Semantic differences in translation

Exploring the field of inchoativity

Lore Vandevoorde

Translation and Multilingual Natural Language Processing

language science press

Translation and Multilingual Natural Language Processing

Editors: Oliver Czulo (Universität Leipzig), Silvia Hansen-Schirra (Johannes Gutenberg-Universität Mainz), Reinhard Rapp (Johannes Gutenberg-Universität Mainz)

In this series:

1. Fantinuoli, Claudio & Federico Zanettin (eds.). New directions in corpus-based translation studies.
2. Hansen-Schirra, Silvia & Sambor Grucza (eds.). Eyetracking and Applied Linguistics.
3. Neumann, Stella, Oliver Čulo & Silvia Hansen-Schirra (eds.). Annotation, exploitation and evaluation of parallel corpora: TC3 I.
4. Czulo, Oliver & Silvia Hansen-Schirra (eds.). Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation: TC3 II.
5. Rehm, Georg, Felix Sasaki, Daniel Stein & Andreas Witt (eds.). Language technologies for a multilingual Europe: TC3 III.
6. Menzel, Katrin, Ekaterina Lapshinova-Koltunski & Kerstin Anna Kunz (eds.). New perspectives on cohesion and coherence: Implications for translation.
7. Hansen-Schirra, Silvia, Oliver Czulo & Sascha Hofmann (eds). Empirical modelling of translation and interpreting.
8. Svoboda, Tomáš, Łucja Biel & Krzysztof Łoboda (eds.). Quality aspects in institutional translation.

ISSN: 2364-8899

Semantic differences in translation

Exploring the field of inchoativity

Lore Vandevoorde

language
science
press

Lore Vandevoorde. 2018. *Semantic differences in translation: Exploring the field of inchoativity* (Translation and Multilingual Natural Language Processing).
Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/000

ISBN: 978-3-96110-072-9 (Digital)
978-3-96110-073-6 (Hardcover)

ISSN: 2364-8899
no DOI
Source code available from www.github.com/langsci/000
Collaborative reading: paperhive.org/documents/remote?type=langsci&id=000

Cover and concept of design: Ulrike Harbort
Fonts: Linux Libertine, Libertinus Math, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
Unter den Linden 6
10099 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

no logo

On ne voit bien qu'avec le cœur

L'essentiel est invisible pour les yeux

(Antoine de Saint-Exupéry)

Contents

Preface

Acknowledgements

This book began as my PhD thesis; I would therefore like to express my sincerest gratitude to my supervisor, Gert De Sutter, my co-supervisor, Koen Plevoets, and the members of my doctoral advisory committee Sandra Halverson, Dagmar Divjak and Els Lefever. I would also like to thank the Hogeschool Onderzoeksfonds, who funded the research to begin with, and Ghent University to provide me with the necessary research facilities to continue my research.

I would like to thank my colleagues Alexandra, Annelore, Bert, Chloé, Christophe, Isabelle, Katrien, Kristien, Lynn, Pauline, Peter, Sevdag, Sofie, Stefaan and a special *merci* to Désirée. I could not have finished this work without the loving support of my friends and extended family Ann, Caroline, Christina, Irene, Elisa, Jasper, Sylvie, Bjoke, Evy, An-Sofie, Camille, Ruben, Chen, Lieve, Jan, Valérie, Anne and Francis. A special thank you to my sweet sister and brother-in-law.

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).

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

1 Introduction

The notion of *meaning* has always been at the core of translation as a task as well as of translation studies (TS) as a discipline. As a task, translation is considered as an act of "communicating the overall meaning of a stretch of language" (Baker 1992: 10). Within the discipline of TS, meaning is an essential concept of the metalanguage of translation and plays, with equivalence – with which it is so closely intertwined – a central role in translation theory (Halverson 1997).

If, as a task, translation is considered as an act of communicating meaning, as the above definition by Baker suggests, this seems to imply that the essence of the task lies within the *transfer* of that overall meaning. The idea that translation is an act of transfer is furthermore suggested by the etymology of the English word *translation*, which means "to carry across". It is at this point that meaning becomes what I would call *the invariant of translation*. Meaning is what is transferred, it is the carefully wrapped content of a box labelled *fragile* that at all times needs to be held securely, carried by a vigilant translator-delivery boy or girl. When the box is opened upon delivery, the deliverer's mission will only be considered successful if the content of the box, once unwrapped, appears to be in the exact same state as when it was wrapped and dispatched by the sender. Any alteration to the box's content is inconceivable, any broken glass or fades colors will necessarily be charged to the deliverer and the box will be returned to sender: invariant content (meaning) is the conditio sine qua non for delivery (translation). Many of the metaphors that are used to talk about translation, such as the one invoked here, adopt the idea of transfer (delivery), of the packing, unpacking and repacking (the box) of a message and its meaning (the content of the box). What is *not* put into question whenever these metaphors are used, is that the content of the box needs to remain unaltered, in other words, that meaning presupposedly remains invariant, if not, delivery (translation) will not take place. But in times of Amazon and DHL, most of us would sense that, when opening the metaphorical box that was just delivered, we at times have this gut feeling that the long travel, the bumpy ride through rain and harsh weather may have somewhat impacted not only the box itself (the form) – as a logical consequence of the dispatch – but also the box's content (the meaning), not so much that it is

immediately apparent, but still. If we take the fragile-box-metaphor to the level of language, could it then be that the meaning of a word, so many times transferred from sender (a source language) to receiver (target language) by so many deliverers (translators), becomes somewhat altered upon reception in the target language, where the entirety of delivered goods constitutes the pile named *translated language*? Is it possible that the meaning of a word, in translated language, is or becomes (slightly) different from its meaning in non-translated language? And how can we investigate something as ephemeral as word meaning in translation? These are exactly the types of questions that the methodology proposed in this book aims to tackle.

Although many scholars explicitly or implicitly accept the idea that meaning is the invariant of translation [1], it is however not a generally accepted given in TS. On the contrary, the idea that meaning is *not* stable has generated a large body of research within socio-cultural studies of translation (Baumgarten 2012). This postmodernist view on meaning dismissed the linguistic view on meaning in TS, and in this way, the debate shifted away from the linguistic, "stable meaning" views in TS to a deconstructed, unstable view on meaning, embedded in cultural studies. In recent years, linguistically-oriented studies in TS have again come to the fore, but the status quo of meaning as the "invariant of translation" seems to be maintained. The aim of this book is therefore to investigate, from a linguistic viewpoint, meaning (un)stability in translation. Admittedly, the empirical investigation of meaning is not a straightforward endeavor, but in neighboring disciplines to TS such as lexical semantics, methodological solutions have been proposed. The development of a methodological solution to compare variance in meaning between translated and non-translated texts is one of the main objectives of this book. It will be illustrated by the investigation of the semantic relations of lexemes in the semantic field of inchoativity in Dutch, leading to a comparison of the semantic field of inchoativity in non-translated Dutch (SourceDutch) to the semantic field of inchoativity in translated Dutch (TransDutch). If meaning is indeed stable, semantic fields of translated and non-translated language should be identical. If however, meaning is not completely stable in translation, differences between the semantic fields are to be expected.

Apart from being the content of the box in a translation task, meaning is also a metalinguistic concept in translation theory, where it is probably as pervasive as

[1] The acceptance of meaning as the invariant of translation by a wide range of TS scholars is apparent from definitions of the concept *tertium comparationis*. This so-called 'third comparator' is based on the idea "that an invariant meaning exists" (Hatim & Munday 2004: 31), independent of both the source and the target text, and that it "can be used to gauge or assist transfer of meaning between ST and TT" (Hatim & Munday 2004: 31).

that of equivalence, although that does not mean that there is a consensus about the meaning of *meaning*. In fact, metalinguistic discussions about the discipline's core elements such as meaning and equivalence are more difficult in Translation Studies than in other disciplines due to the nature of the discipline itself:

> Above and beyond that the very nature of the discipline [Translation Studies] means that the discourse is conducted in and through a number of different languages, and with language being both the object of discussion and the means of communication, the risk of non-communication is only increased (Snell-Hornby 2007: 314).

As if to avoid venturing onto thin ice, recent theoretical paradigms in TS such as the universals[2] paradigm, seem to circumvent the whole idea of meaning, implicitly considering it as an integrative part of what translation is, rather than engaging in what seem to be an endless theoretical discussion. Numerous corpus-based studies within this paradigm (Malmkjaer 1997; Laviosa 1998; 2002; Mauranen 2000; Olohan & Baker 2000; Baker 2004; Bernardini & Ferraresi 2011; Delaere et al. 2012b; De Sutter et al. 2012b; Kruger 2012b) have focused on lexical and grammatical phenomena (the packaging of the box) and have somewhat neglected the semantic level (Laviosa 2002: 28)[3] (the content of the box). To my knowledge, the question whether (universal) tendencies of explicitation, simplification, normalization or levelling out can be found on the semantic level has not yet been raised in TS.

With this work, I want to answer the three questions that arise here about meaning in translation:

- How can we investigate semantic differences in translated vs non-translated language?
- Are there any differences on the semantic level between translated and non-translated language?
- If there are differences on the semantic level, can we ascribe them to any of the (universal) tendencies of translation?

In order to answer these questions, I will first propose a methodological framework which offers a strategy to operationalize the idea of semantic difference

[2] Universals are "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).

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

between translated and non-translated texts. Secondly, the exploration of the semantic field of inchoativity in Dutch will enable me to tackle the second and the third question I aim to answer with this study. All this will finally lead to the formulation of a number of recommendations for future research about (universal) tendencies of translation on the semantic level.

The outline of this book is as follows. Chapter 2 provides the theoretical foundation of this work. Every section of this chapter constitutes a building block necessary to arrive at the methodology presented in Chapter 3. In the first part of the theoretical chapter, I will zoom in on a number of aspects of corpus-based translation studies (CBTS) which form an integral part of this study: the use of corpora in TS, the translation universals and the cognitive turn in TS. I will equally discuss the place of the study of meaning within CBTS as well as the relationship between (the study of) universals, (the study of) meaning and the notion of equivalence. In the second part of this chapter I will look into different sub-disciplines of linguistics such as contrastive corpus linguistics and corpus semantics, which have, compared to CBTS, a much longer tradition of investigating meaning relationships. The theoretical foundations for the development of a bottom-up, statistical visualization method of semantic fields in both translated and non-translated language will be laid here. I 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.

Chapter 3 contains a thorough description of the methodology. The method which is developed is an extension of an existing method, the Semantic Mirrors Method (SMM) (Dyvik 1998; 2004; 2005); it is corpus-based, uses statistical visualization techniques and consists of two parts (two extensions to the SMM). The first extension allows the potential user of the method to select candidate-lexemes for a semantic field. The second extension to the SMM proposes a way to visually inspect the retrieved data set(s). The ultimate goal of these extensions is to enable the user to compare visualizations of semantic fields of translated and non-translated language to each other.

In Chapter 4, I apply the methodology to the semantic field of inchoativity in Dutch. The choice of inchoativity as a 'test case' is certainly not the most obvious choice, but offers a number of advantages: (i) I expect to find 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 1996a). Inchoativity is therefore a compelling test case when one is interested in revealing meaning differences. The results are presented and described on the basis of three main visualizations, one for a semantic field of inchoativity in non-translated Dutch (SourceDutch), one for translated Dutch with English as a source language ($\text{TransDutch}_{\text{ENG}}$) and one for translated Dutch with French as a source language ($\text{TransDutch}_{\text{FR}}$). The goal is to explore the semantic field of inchoativity in Dutch and by doing so, to formulate an answer to the second and the third question of this study: are there any differences between translated and non-translated language on the semantic level, and, if there are, can we ascribe them to any of the (universal) tendencies of translation (we will focus on levelling out, normalization and shining through)?

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 this study: the Gravitational Pull Hypothesis, developed by Sandra Halverson and the Neurolinguistic Theory of Bilingualism, by Michel Paradis.

Chapter 6 concludes this book with an overview of the main findings with regard to the differences and similarities of the semantic relationships in translated Dutch compared to non-translated Dutch for the semantic field of *beginnen*/inchoativity. In the conclusive discussion, I will comment on the methodological contribution this work possibly makes to the empirical study of semantics in translation, especially with regard to the impact of translation on semantic representations. Finally, a number of recommendations for future research about (universal) tendencies of translation on the semantic level will be made. This book will then end where research into semantics in translation could begin, with the possibility of taking the conclusive ideas of this work as a starting point.

2 Theoretical considerations

2.1 Introduction

Modern corpus linguistics (CL) 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 subdisciplines 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 (CBTS) was born. It is within this paradigm that this work is situated.

In the first part of this chapter (§2.2), I will introduce the discipline of CBTS. 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 §2.3, I will investigate a number of contrastive corpus studies. I 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 (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 §2.4.1, I will elaborate on the notion of translational equivalence. Its operationalization within Word Sense Disambiguation (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, I choose to create bottom-up, statistical visualizations of semantic fields in translated and non-translated language. In §2.4.2, it will be shown 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 §2.4.3, I propose to combine the corpus-based quantitative visualizations with a theoretical framework from cognitive linguistics. I will propose to use the prototype model of category structure as a necessary basis for a coherent interpretation of the statistical visualizations.

2.2 Corpus-based translation studies

In the first part of this section (§2.2.1) I will zoom in on the different types of corpora, which constitute the main methodological tool in CBTS. In the second part (§2.2.2), I will focus on how precisely this new sub-discipline arose within translation studies, by further exploring the research program set up by Baker. I will give extensive consideration to the translation universals paradigm and I will show why, in my opinion, research into universals on the semantic level has barely had any uptake within CBTS[1]. In addition, I will determine which universals seem best suited for the investigation of semantic relationships in translation. In §2.2.3, I 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 §2.2.4.

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 to create typologies for the general purpose of corpus linguistics (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.

[1] 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). I choose, however, to focus on the universals research program which has undeniably dominated the field since the 1990s.

Instead of undertaking a (necessarily non-exhaustive) overview of existing corpora, I will 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 selection of a corpus that suits one's research needs.

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 size mattered was quickly assimilated. To overcome the obstacle of corpus size, the logical step was 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. By that time, the world wide web had started to be used as a corpus too. 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 & Steiner 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 TS 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 (comparable corpora have nevertheless been frequently used in CBTS).

2.2.1.2 Content

While for most of the history of CL, definitions of a corpus most often limited its content to files of text, the recent appearance of multimodal corpora 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 interest-

ing one, I will not further explore this type of corpora (since this 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 n^{th} to the n^{th} 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) versus 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 Languages 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, which is presented below. Her corpus typology focuses on the applicability of corpora to the study of translation. Given the focus of this book on translated and non-translated language, I will maintain Laviosa's typology:

- 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) – 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 being 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.

Laviosa established this corpus typology because she considered it to be "an essential step towards developing a coherent methodology in corpus-based translation studies" (Laviosa 2002: 38).

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 & 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 &

Otani (2010: 328), who see no reason to consider CL as a theory, but consider it is a methodological paradigm.

A second issue concerns 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, one 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 one compiles or selects his 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, it may be difficult to know which

forms in the compared languages of the comparable corpus have similar meanings and pragmatic functions, and consequently, which forms can be compared and which ones not (Dyvik 1998: 5). On the other hand, comparable corpora seem to be easier and faster to compile than parallel corpora since it is usually more straightforward to identify texts as original texts or as translations than it is to find source texts with their matching translations. A drawback of parallel corpora, however, is that all texts labeled as *original/non-translated* in a parallel corpus (representing for instance the native, standard variety of a language) have at some point been selected to be translated (since all non-translated texts in a parallel corpus are a 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 should be kept in mind that the presence of texts in a parallel corpus can be based on their 'suitability' to be translated (and hence, their absence can be based on their unsuitability). In order to overcome this problem, it is possible to include a monolingual reference corpus for supplementary comparison, but studies that have done so, faced comparability issues due to corpus size or due to 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), Gellerstam (1986) and Frawley's idea of third code (Frawley 1984) – 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 §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" (Baker 1993: 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).

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 CL. 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 can actually 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 cor-

pora (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 this volume). 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), I 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.

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, 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. Numerous studies were carried out to test the explicitation hypothesis, e.g. Øverås (1998), Olohan & Baker (2000), Olohan (2003), Mutesayire (2004), Puurtinen (2004) and many others (see Kruger (2012a) or Zanettin (2013) for overviews of the translation universals literature).

A study on syntactic explicitation, carried out by Olohan & Baker (2000) 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 2010). If Baker's statements (1993: 244, Baker (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 less 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" (2010: 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).

With regard to the operationalization of simplification 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 (1996) 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 or corpus is (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 2012a). The use of this measure is based on the assumption that "word length can be seen as a measure of morphological complexity. […] [M]ean 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 2012a: 366).

The measures proposed to investigate simplification on the level of a text or (part of) a corpus are mainly quantitative and very little or no doubt can arise as to how a type-token ratio or a mean word length should be operationalized. However, one might wonder to what extent these measures really indicate simplification in Baker's sense of "making things easier for the reader" (Baker 1996: 182). Some of the measures discussed here such as average sentence length do not seem to "make things easier" (Baker 1996) at all. In addition, from the viewpoint of readability research, which is equally concerned with "what makes some texts easier to read than others" (DuBay 2004; cited by Clercq et al. 2014), the 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 […] (Clercq et al. 2014: 294).

As a consequence, a more up-to-date and complete measure of simplification in TS would 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 my 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. Finally, if one is indeed interested in discovering whether translated texts are *easier to understand* than non-translated texts, the question whether they are *simpler* might just be the wrong question. Rather, researchers in TS could ask themselves: 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. This tendency is quite possibly influenced by the status of the source text and the source language, so that the higher the status of the source text and language, the less the tendency to normalise (Baker 1996: 183).

The third universal feature of translation, normalization – also referred to as *conventionalization*, *standardization* or *sanitization* (Zanettin 2013: 23) – is defined as a tendency to conform with typical features of the target language, and will depend on the status of the source language: a higher status of the source language will decrease the tendency to normalize. By virtue of its own definition, normalization then in fact dismisses itself as a universal strictu sensu: if normalization is susceptible to the status of the source language, it cannot be universal anymore, since universals are by definition "not the result of interference from specific linguistic systems" (cfr. the quote by Baker in §2.2.2). Despite Baker's own concession to the strict interpretation of universals, source-language related phenomena such as interference have often been excluded from the universals research paradigm, under the pretext of 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-Schirra 2011) and degrees of formality of pairs of near synonyms (De Sutter et al. 2012a). The results of these studies are however far from univocally 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" and De Sutter

et al. (2012b: 338) conclude (i) that degrees of formality in translated texts may differ depending on the source text, (ii) that translated texts are not always more formal (more conservative) than non-translated texts and (iii) that, when translating into the same target language, translators normalize less when translating from one source language and more when translating from another source.

As Baker intuited about normalization, it seems very hard to pretend that normalizing trends in translated texts are (completely) source language independent. As a consequence, source language influence on translated texts has been taken into account with regards to the normalization hypothesis, and researchers like Teich have 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, my emphasis)

While the second hypothesis corresponds to the classic definition of normalization (adherence to target language norms), the first one hypothesizes that normalization can also take place in the opposite direction (adherence to source language norms). In such cases, the source language is said to literally *shine through* in translated texts. 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" where the specific features observed in translated texts would 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.

2.2.2.4 Levelling out

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

While the above 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. From the perspective of levelling out, the question is raised whether lexical density levels amongst translated texts are more similar than lexical density levels 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 (2012a) 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 2012a: 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, to my knowledge, no studies have so far focused on this universal in particular. This is most probably because levelling out 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 my 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 therefore be needed (a common requirement for both the investigation of meaning as well as for the universal of levelling out).

2.2.2.5 Universals: the more the merrier?

Over the last more two and a half decades, the notion of translation universals has set many tongues wagging. In the mid-2000s, after ten years of universals research, 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 (Mauranen & Kujamäki 2004: 1). Some universals were indeed refuted (see e.g. Becher 2010) and new ones, such as the Unique Items Hypothesis (Tirkkonen-Condit 2004) or the Asymmetry Hypothesis (Klaudy 2009) were added to the list. The Unique Items Hypothesis states that some features that are unique to the target language will appear less or will not appear 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", Chesterman 2004: 39). Chesterman's S-universals are features such as lengthening (translated texts tend to be longer than their originals) – a feature of translated language that had been proposed by Vinay & Darbelnet (1958: 185) – and Toury's (1995) well-known laws of interference (source text features are transferred to the target text) and grow-

ing 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 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). Following Chesterman, any of the generalizations made on the basis of a corpus study should indeed first and foremost apply to the language-pair(s), the period, the text genre(s) etc. which are selected by the researcher (and often determined by the chosen (sub-)corpus 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 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 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 seem to try to do away with this idea of universal applicability. Other scholars have introduced 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 in-

clude 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 "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. (2012a); Diwersy et al. (2014) and the studies presented in the edited volumes by Oakes & Ji (2012) and Sutter et al. (2017) have shown that multivariate statistical methods can be successfully implemented into CBTS. In this work, I will use multivariate statistical techniques 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 §2.4 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 §2.2.2 I stated 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 left 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 opera-

tion), 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 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 CBTS that address the problem of meaning. In §2.2.3, I will show that any CBTS study concerned with meaning will at some point be confronted with the concepts *equivalence* and *correspondence* or will 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: 82). 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, in order to know if it is true that meaning is the invariant of translation, one 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 §2.2.3, some important investigations of meaning in translation carried out over the last two decades will be highlighted.

2.2.3 The cognitive turn in translation studies

2.2.3.1 Translation and Meaning series

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) cover an incredibly wide range of subjects. 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…is included. However, very few studies in the series take a linguistic viewpoint on meaning. Most of the papers fit the series' subject because of the general acceptance that translation *is* meaning, and that meaning is the invariant of translation, leading to the possibility of including virtually any branch of TS into the series. Laviosa-Braithwaite's "Comparable corpora: Towards a corpus linguistic methodology for the empirical study of translation" in the third part of the series (1996) 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. Unfortunately, they do not consider the scholarly contributions lexical semantics could make to the discipline of TS but (only) explain how lexical semantics could be of help for the professional translator (Klaudy 2010: 100). The same applies for corpora: a number of studies in the Translation and Meaning series use corpora to investigate (linguistic) meaning. They often focus on the utility of such tools for translation teaching and translation quality assessment but do not challenge the idea of meaning itself (e.g. Bednarczyk 1997; Lan & Bilbow 2007; Oster & van Lawick 2008). One of the few studies that actually does engage in 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 meaning variation in translation.

2.2.3.2 The re-introduction of meaning in translation studies

The corpus turn from the 1990s – which ensued from descriptive translation studies (hence: DTS) – was the necessary prerequisite for TS to become a discipline in its own right. According to (Lewandowska-Tomasczyk (2002: 41), there were two "leading recurrent themes" in translation theory in the 1990s: corpus research and

cognitive approaches. Within the first current – corpus research – the so-called universals paradigm was quickly adopted by a large group of scholars. As for the second current, one would have expected an emphasis on linguistic meaning and the status of meaning in TS, parallel to cognitive linguistics (Lewandowska-Tomasczyk 2002: 41). However, the earliest attestations of a cognitive TS did not immediately show interest in a re-introduction of linguistic meaning (and equivalence) but rather focused on corpus-based empirical and experimental (process) research and introduced, for instance, think-aloud protocols (see e.g. the volume edited by Tirkkonen-Condit & Jääskeläinen 2000) which equally benefited from the cognitive-translational setting. Rojo & 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).

Although the above mentioned "notion of referential truth" was abandoned within most cognitively oriented studies of translation, a number of scholars did propose a "linguistic-cognitive orientation" (House 2013). House (2013) makes a strong plea in favor of a "linguistic-cognitive orientation" in TS, since "translation is above all an activity involving language and its cognitive basis" (House 2013: 47). In her opinion, TS scholars have been so pre-occupied uncovering and investigating all the "external social, cultural, personal, historical etc. factors impinging on translation 'from the outside' " that they have somewhat forgotten about what translation is made of 'from the inside' (House 2013: 47). She argues in addition that a cognitive-linguistic view on translation could counterbalance the wide focus on these external factors:

> For balance, I think it is also necessary and insightful to describe and explain how strategies of comprehending, decision making and re-verbalization come about in a translator's bilingual mind" (House 2013: 46).

In the following paragraphs, I will highlight a number of studies which have specifically engaged with this linguistic-cognitive orientation on TS. This non-exhaustive overview focuses on those studies which in my opinion have provided important insights for the study of translation within a linguistic-cognitive framework, especially with regards to the status of meaning in translation or the

impact of translation on meaning. Early attestations of this linguistic-cognitive orientation include Tabakowska (1993), who introduced 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.

2008 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" (2008: 249). She suggests that (translation) scholars "should aim at a description of prototypical features, inherent or contextual" rather than "attempting an exhaustive analysis of a lexeme" (2008: 251). One of her key focus areas is semantic prosody, i.e. "the spreading of connotational colouring beyond single word boundaries" (Partington 1998: 68 in: Korning Zethsen 2008: 256:), which has not often been investigated by contrastive comparison, and even less so in translational settings. 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' 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 semantic 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 of such subtle issues as semantic prosody in translation.

Klaudy (2010) takes the point of view of translation universals to investigate lexical specification and generalization. Her contrastive view on translation (translation as transfer) is in apparent opposition with the mainstream research into universals, which generally 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. She introduces the concept *lexical transfer operations*:

> In our interpretation, "lexical transfer operations" is a collective term for 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).

The identification of these "operative moves" can provide additional information on language differences not yet distinguished by contrastive lexicographers and bilingual dictionary builders. Klaudy furthermore expresses a special interest in the impact of the process of translation on systemic language differences (2010: 82) and questions the "firmly embedded idea that meaning remains unchanged in translation" (2010: 82). Klaudy makes a distinction between *meaning* and *sense* to explain the processing of meaning in translation. *Meaning* covers "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" (Klaudy 2010: 83). The process of translation involves the recreation of *sense* in the TL, "instead of retaining the SL meaning" (2010: 83). 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), and this is where according to Klaudy, meaning (in)variance ties up with translation universals.

In the volume edited by Rojo and Ibarretxe-Antuñano, Martín de León (2013) explores how cognitive models of meaning can be of use for translation. "Different cognitive approaches provide different visions of meaning, and they also lead to different theoretical frameworks for empirical translation research" (Martín de León 2013: 99). She explains how the *classical paradigm* (which is coherent with truth-conditional semantics and generative grammar, and has influenced translational models such as the ones proposed by Nida (1964) and Kade (1968)) is unsuitable to explain translation processes:

> If meaning could emerge intrinsically from a system of abstract symbols, the degree of equivalence between two texts could be determined just comparing their respective linguistic systems (Martín de León 2013: 103).

Meaning cannot be seen as invariant ("transferable, invariable information units"), since that would resume the translator's task to a mere transferring of

information encoded in the source language into the target language (Martín de León 2013: 99). According to Martín de León, the fact that no consensus has yet been reached about equivalence in TS (precisely because translation involves more than a simple transfer of encoded information), shows that the classical paradigm "does not provide a model for meaning construction [...][and] cannot explain the processes involved in human translation" (Martín de León 2013: 103). By contrast, if 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). If one adheres to the idea that meaning is dynamic, a *connectionist* view on meaning can be taken as a starting point. Connectionists see meaning construction as "a dynamic process of pattern recognition and construction" (Martín de León 2013: 100). Unlike the classical approach, connectionism allows for a view on translation as a process of meaning construction *beyond* the decoding-recoding metaphor, (Martín de León 2013: 105). An approach that is coherent with connectionism is prototype theory:

> Prototype theory is a model of human categorization processes in which the internal structure of a category is defined by a series of family resemblances, with some features applying to a subgroup of the category, some others applying to another subgroup, and a most representative element or prototype (Martín de León 2013: 106).

Martín de Leóns proposal for a connectionist view, and more specifically a prototype-based view on meaning in translation is then indeed coherent with the idea that meaning construction is variant.

One of the few scholars who has been consistently occupied with the study of meaning in translation is Halverson (2003; 2010; 2017; 2013). Since the beginning of the 2000s, she has been developing a hypothesis that accounts 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. 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 or structure more or less prominently than another one. In Chapter 5, I 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.

2.2.3.3 Conclusion

Translation universals research generally takes meaning invariance as its baseline and does not problematize the impact of translation on the (in)variance of meaning. The studies that were presented in this section all agree – some do so implicitly, others explicitly – on the important point that the re-introduction of meaning research should not be considered as an obstacle to the study of universals. On the contrary, cognitive-linguistic views on meaning and translation are put forward as possible explanations for (some) translation universals.

The authors cited above have contributed in various ways to the study of meaning in translation and their contributions are of major importance for this study. Firstly, Korning Zethsen explicitly linked corpus-based cognitive (lexical) semantics to the study of translation. In that regard, her methodological intentions are closely linked to the ones put forward in this study. Furthermore, Halverson clearly explained 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, I will apply the Gravitational Pull Hypothesis so as to explain some of the phenomena that emerged from the comparison of semantic fields of translated and non-translated language. Secondly, both Martín de León and Korning Zethsen opt for a prototype-based view on meaning in TS. Martín de León convincingly showed that a connectionist view on meaning construction in translation is necessary to be able to explain the human translation processes from a cognitive point of view. A prototype-based view on meaning in translation is also pro-

posed by Korning Zethsen. In this study, such a prototype-based view will also be adopted (see §2.4.3). Finally, the importance of the notion of equivalence was indicated by both Klaudy and Korning Zethsen. 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-introduction of the source texts into the comparisons. Korning Zethsen's research into semantic prosody further put into question the notion of the invariance of equivalence. The contrastive methods I will rely on to carry out an intralingual comparison between different varieties of Dutch equally necessitate such an equivalence concept. In the next section, I will therefore define the notion of equivalence in view of the current study.

2.2.4 On a tightrope with equivalence

The notion of equivalence is one of the most heavily loaded concepts in TS. A number of developments within the discipline – ranging from Nida's sociolinguistic translation analysis (Nida 1964; Nida & Taber 1969) to skopos theory (Nord 1997) and including cultural, power and other turns – favored 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 itself. 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[2].

A linguistic-oriented study of translation such as this one 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 while relying on and investigating contrastive relations between source and target language. Because of this linguistic-cognitive view on translation, and in view of formulating my own definition, I am particularly interested in how the 'early' linguistics-oriented scholars defined equivalence.

[2]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).

In his work 'A Linguistic Theory of Translation' Ian Catford (1965) discriminates 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 is an important distinction 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-oriented idea of equivalence seemed thus to reside in its source-orientedness as well as in its prescriptive nature.

In the early 1990s, Reiss' and Vermeer's Skopos theory (1991) emphasizes 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" (Toury 1980: 11). 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 explicates 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.

My proposal for a definition of equivalence is now 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. My definition does not impose any prescriptive rule on what is acceptable or not as equivalence, it 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 greatly indebted to Catford's and Toury's ideas about equivalence as an empirical relation. Rather than imposing on the equivalence relation a need to be "the closest natural equivalent", in my view, equivalence can be thought to represent the relation between the source and the target text, that what connects source and target, irrespective of the nature of what is represented in this connection. This definition forms the baseline of my idea of equivalence. The operationalization of the equivalence concept for the purpose of this study will require an extremely pragmatic application of this definition so that it can be applied to a manual word-level annotation procedure of a sentence-aligned corpus (see § 3.3).

2.2.5 Conclusion

The study of meaning variance in translation is still largely unexplored in CBTS. Within the universals paradigm – most probably the most pervasive one in current CBTS – (in)variant meaning and equivalence are usually not problematized. An empirical corpus study such the present one which questions meaning invariance, however needs a workable definition of equivalence such as the one formulated in this section. Although such a definition is a necessary starting point, it still does not provide a methodological procedure to investigate meaning relationships in translation. In fact, very few studies have suggested and even less so have actually developed methodological procedures for this kind of research in CBTS. In the next section, I will 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, I will focus on corpus approaches that have manifested an explicit interest in the contrastive study of meaning via corpora. The goal of this section is to find a way in which an equivalence-like concept coming from contrastive linguistics 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 that should be kept in mind throughout this section is as follows: how can a linguistically inspired notion of translation equivalence be used in such a way that it meets the following requirements? Firstly, the adopted notion of translation equivalence needs to allow for a comparison of translated and non-translated language. From the point of view of TS, translated and non-translated language are considered as different varieties; and it is therefore necessary to distinguish between translated and non-translated language. Secondly, whenever a relation of translation equivalence is established, the existence of that relation should not imply that the conveyed meaning is invariant. In this section, I will first focus more generally on the use of translations in contrastive studies (§2.3.1), before I focus on the procedure of back-translation, which is considered as one (of the most) fruitful applications of translations in a contrastive context (§2.3.2). I will also explore two successful applications of back-translation in contrastive analysis: Mutual Correspondence (§2.3.3) and Semantic Mirroring (§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-fertilizations that exist between the two disciplines (Vandepitte & De Sutter 2013 for a survey) are all linked to this one element that 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 theories 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 & 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 center 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). 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 TS the so-called corpus turn coincided with the emergence of parallel corpora. Although McEnery & Hardie (2012: 20) claim that parallel corpora are typically used for translation research and comparable corpora for contrastive studies, this is only partially true. Comparable corpora have equally been used for translation research (think of earlier mentioned research by Baker and Laviosa-Braithwaite) and parallel corpora have been both extensively and fruitfully used in contrastive studies. Within contrastive studies, translation equivalence – necessarily established on the basis of parallel corpora – was considered as "the best available tertium comparationis […] [to] 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:

> [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).

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 were worried 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 by a number of other scholars such as Krzeszowski (1971; 1972) from a contrastive linguistic point of view (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). According to Ivir, information from translations can be of valuable use for contrastive linguistic analysis. His main concern is therefore to show "how translation equivalence enables the analyst to isolate formal correspondents" (Ivir 1981: 58). Ivir states that formal correspondents in the way defined

by Catford "can hardly be said to exist" (1981: 54) and he therefore proposes to adapt Catford's definition of formal correspondence so that it becomes defined "with reference to translationally equivalent texts" (Ivir 1981: 55) rather than to linguistic systems. In this way, formal correspondence becomes a text-based and equivalence-based concept, in which the relationship between the correspondents is a one-to-many relationship (1981: 55): 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" (1981: 55). He further on repeats that "such multiple correspondents are important analytical pointers to distinctions of meaning in the source language" (1981: 56). It is exactly this idea that will be exploited for the development of the Semantic Mirrors Method (see §2.3.4) which uses translations to lay bare different meanings. At all times, Ivir remains conscious about the difference in nature between translation and contrastive linguistic analysis (Ivir 1969: 15, Ivir 1970: 17, Ivir 1983: 173): translation aims at semantic equivalences 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 the distinction between *parole* and *langue* has become somewhat obsolete in contemporary corpus-based cognitive linguistics – which is considered 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 a contrastive linguist such as Ivir. His concern with the *langue* vs. *parole* dichotomy ultimately led to the formulation of a practical solution allowing many corpus and contrastive linguists to fruitfully use translational data: back-translation.

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):

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).

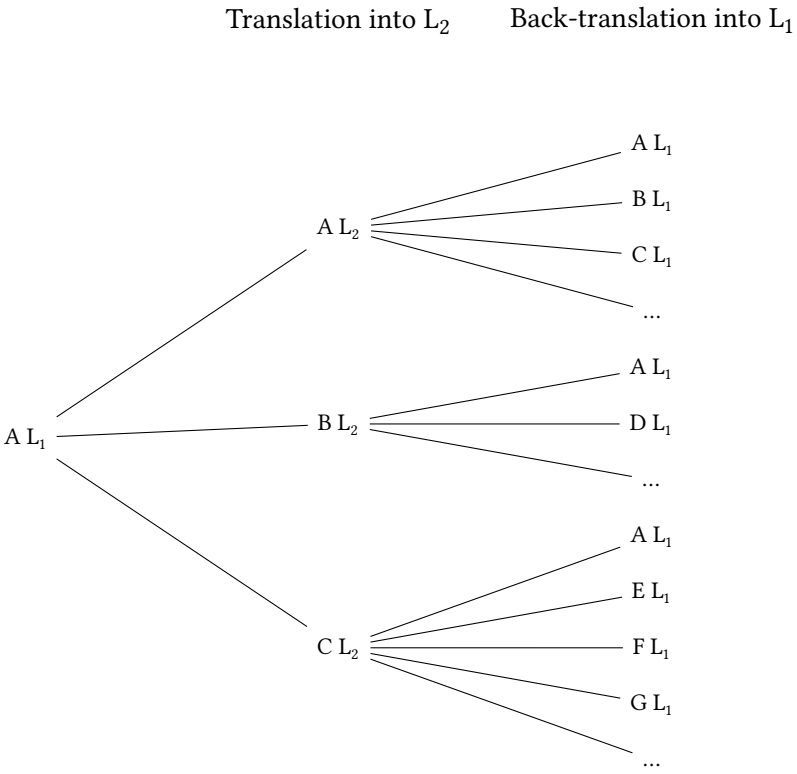


Figure 2.1: Back-translation procedure for contrastive analysis (Ivir 1987: 478)

Second, Ivir's concern with the distinction of contrastive correspondents equally allows the (translation studies) researcher to separate the "irrelevant differences that are due to the translator's idiosyncrasies or motivated by particular communicative or textual strategies" (Altenberg & Granger 2002: 7:17) from what Dyvik will call the Linguistically Predictable Translations. The back-translation procedure seems thus to be suitable for both (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

Within contrastive linguistics a consensus was reached about the fact that information from translation corpora could be used as "an empirical basis for se-

mantic claims" (Noël 2003: 758). According to Ebeling & Ebeling (2013: 24-28), back-translation formed the basis for using translational data and parallel corpora in contrastive analyses. Ivir's work on back-translation has indeed been further used and developed within contrastive linguistics by, amongst other researchers, Altenberg (1999) and Altenberg & Granger (2002).

Altenberg and Granger's application of back-translation is called Mutual Correspondence (henceforth MC; Altenberg 1999: 254 ff. Altenberg 2007: 9, Altenberg & Granger 2002: 7:18) and combines the idea of back-translation with a quantitative equivalence concept (such as Krzeszowski's notion of statistical equivalence (Krzeszowski 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:
>
> $$\frac{A_t + B_t}{A_s + B_s} \times 100$$
>
> where A_t and B_t are the frequencies of the compared items in the translations, and A_s and B_s 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 offered by back-translation "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 (Gilquin 2008). 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 agree 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 in-

dividually; the overall 'network' of relationships between the source language lexeme(s) and all attested translations stays somewhat out of the picture.

Although MC is an interesting application of back-translation for semantic research, it is, due to its clear contrastive nature, incompatible with the research objective of this book, which is to compare semantic field representations in translated and non-translated language, since such a comparison 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 (hence: SMM) 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 use in machine translation and other types of multilingual processing" (Dyvik 1998: 51) and later also the derivation of wordnet relations within the previously mentioned project (Dyvik 2004: 311). The idea of the SMM – which will be at the heart of the methodological tool I will propose – in fact finds itself at this crossroads of linguistic software development and lexical-semantic investigations. In this section, I 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.

According to Dyvik, the semanticist first needs to get persuaded of the usefulness of translation for linguistic semantics. Apart from his argument that translation is a large scale activity, bringing about a multi-lingual perspective on lexical semantics, Dyvik additionally and convincingly argues that translation is a normal kind of linguistic activity within wich meaning is evaluated and that, as a consequence, this evaluation takes place without any kind of metalinguistic,

philosophical or theoretical reflection (Dyvik 1998: 51). Hence, the relations between the texts, which are the observable results of the translator's evaluation of the meaning under scrutiny, can be considered as empirical evidence about semantic relatedness (Dyvik 1998: 51).

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 to produce 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 (Dyvik 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: "bad translations" need to be filtered out of the data and so-called Linguistically Predictable Translations[3] need to be iso-

[3] 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 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.

lated from the totality of the data (Dyvik 1998); the latter ones are 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. The difference between the two proposals 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[4], 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 subsenses of the meaning of *heks*: each of the English words indicates one of the many possible ways of dividing the semantic potentiality of *heks* (Dyvik 2005: 31).

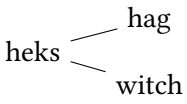


Figure 2.2: Translational correspondence

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:

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).

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

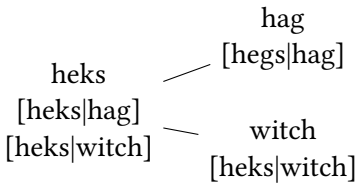


Figure 2.3: Translationally derived features

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" (Dyvik 2005: 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 will adopted for the extension of the SMM (see §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 (Dyvik 2005: 31). In this

way, "sets of translationally corresponding items across languages [can be seen] as *the primitives of semantic descriptions*" (Dyvik 2005: 31), and the contrastive pairs can be considered as semantic features, assignable to lexical items (Dyvik 2005: 31).

2.3.4.3 Ivir and Dyvik

Although Dyvik's mirroring method shows quite some resemblances to Ivir's ideas about contrastive correspondents and back-translation, Dyvik seems to develop his method 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). Since the current study focuses on translation itself, and not merely on its exploitation as a (logical) tool, the asymmetry of the translational relation will necessarily have to be taken into account.

One could wonder why such an effort is made here to present Dyvik's technique, if in fact Ivir's previously formulated ideas were so similar. There are several important reasons to prefer the SMM to Ivir's 'pure' back-translation as a basis for this methodological tool. First, Dyvik makes an important link between a technique, back-translation, and a specific research objective: lexical semantic research, an objective which I 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, I 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. Overlap is part of the procedure of back-and-forth translation, and forms an additional dimension which will be used in the extended version of the SMM (see §3.4).

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" in semantic field research. They mainly focused on discourse particles (Simon-Vandenbergen 2013), pragmatic markers (Simon-Vandenbergen & Aijmer 2002; Aijmer & Simon-Vandenbergen 2004; Aijmer et al. 2006) and adverbs (Simon-Vandenbergen & Aijmer 2007; Simon-Vandenbergen 2013)[5] . 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 the supplementary information which translations are able to provide about semantic similarity into their analysis. They show an interest in using the back-and-forth translations as a *tertium comparationis* (Simon-Vandenbergen & Aijmer 2002: 16, Aijmer & Simon-Vandenbergen 2004: 1795), but their main interest in Dyvik's proposal stems from its aptitude to construct and compare semantic fields (Aijmer & Simon-Vandenbergen 2003: 1131, Aijmer & Simon-Vandenbergen 2004: 1782, Simon-Vandenbergen & Aijmer 2002: 13). A number of adaptations and specifications are made by Aijmer and Simon-Vandenbergen to Dyvik's orig-

[5] 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 from 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.

inal method:

Firstly, Aijmer and Simon-Vandenbergen always use at least three languages; i.e. the language under study (English) and two mirror languages: either Dutch and 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[6] 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 §??). Combining different mirror images into one result also implies that data are obtained from different corpora and need to be combined while staying comparable.

Secondly, whereas Dyvik's "ranking of signs in a semantic field" is done "quite independently of frequency of occurrence"[7] and based on the "overlap relations among *t*-images" (Dyvik 1998: 73)[8], Aijmer & 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

[6] Note that this interpretation of overlap differs from the interpretation of the notion in this study.

[7] "(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).

[8] 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).

> 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 (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" (Simon-Vandenbergen & Aijmer 2007: 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[9], 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" (Simon-Vandenbergen & Aijmer 2007: 248).

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 their use of multiple languages to arrive at a final semantic map as well as the integration of frequency information, although without statistically incorporating this information into the analysis.

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 to visualize semantic relatedness instead of the Venn diagrams proposed by Dyvik. Eldén et al. (2013) propose to visualize the

[9] 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.

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

Back-translation is a technique that can be used as a contrastive linguistic tool. It enables the researcher to isolate formal correspondents (renamed and re-defined by Ivir as contrastive correspondents) and to detect 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. Although the SMM has indeed the potential to lay bare meaning relationships, a number of issues remain unsolved. First, a operationalizable notion of translation equivalence allowing for valid comparisons between translated to non-translated language is still to be defined. Both Dyvik and Ivir established equivalence on the basis of a symmetric notion of the translation relation, but the idea that equivalence is symmetric is incompatible with the viewpoint of CBTS which is taken in this book. 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 the first two issues can be overcome, a theoretical framework within which those comparisons can be interpreted, is still missing. Solutions to each of these problems can be found within corpus-based semantics.

2.4 Corpus semantics

Various theoretical insights from different areas of corpus-based semantics are brought together in this section. These insights are needed to underpin the methodology which will be presented in Chapter 3. 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 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. I 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), I deal with the notion of translational equivalence as it was developed in Word Sense Disambiguation. 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 this study).

In §2.4.2, I will show 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, Speelman & Geeraerts 2012: 17) to study semantic relationships in fields of translated and non-translated language.

In §2.4.3, I 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 §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[10] are either based on the distri-

[10] 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).

butional 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)[11].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 et al. (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[12], methods that rely on translational equivalence as a meaning-structuring device have not yet had 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

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

[12]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; Gries & Divjak 2009.

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:

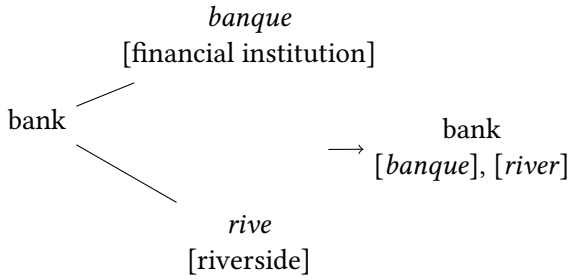


Figure 2.4: Different senses of the English lexeme *bank* are assigned based on its French translations

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,* hence, ambiguity remains unresolved (Figure 2.5). 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 2.6).

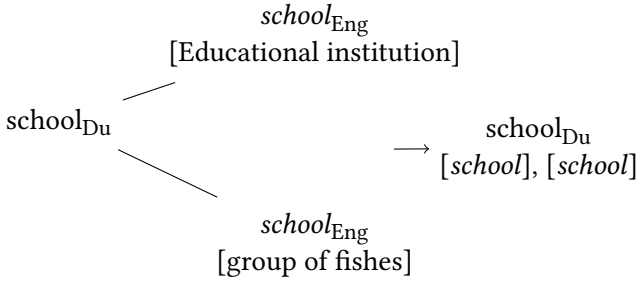


Figure 2.5: Unresolved disambiguation via one language

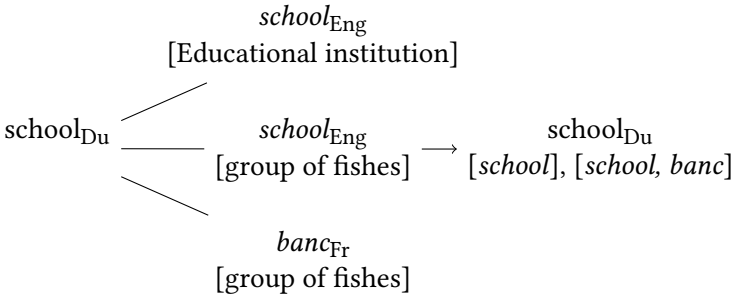


Figure 2.6: Resolved disambiguation via two languages

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 my

point of view that translation is necessarily asymmetric (see §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). By considering the relation of translational equivalence in the SMM as identical to the one in a WSD disambiguation task with several languages, i.e. not necessarily symmetric and lexicalized cross-linguistically – the SMM can be used for the disambiguation task carried out in this study.

2.4.2 Vector Space Models

The SMM can be used to reveal semantic relationships, but 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 it does not offer solutions as to how the disambiguated senses can be objectively compared to each other. Objective comparisons would indeed require objective visualization methods, which neither the SMM nor WSD straightforwardly offer. In this section, I will turn to linguistic semantics and corpus-based cognitive semantics, which are mainly occupied with the empirical study of lexical meaning. 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" (Heylen et al. 2012).

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 2006c: 6). Geeraerts (2016: 242) and Stefanowitsch (2010) discern three major perspectives: experimental research, the referential method and the distributional, corpus-based approach. My 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 2016: 242). 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 (2006a) and Divjak (2010a) and correspondence analysis by Stefanowitsch (2010) (Geeraerts 2016: 243). The third approach discerned by Geeraerts, 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: in order to know how semantically similar two words are in translated and non-translated language, equally the spatial proximity between those two words can be measured in both varieties. For instance, the semantic similarity between stoel [chair] and bank [bench] can be measured in translated Dutch and compared 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 this one, 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" (Heylen et al. 2012: 17) will be needed. Such an interface (a visualization) can then be obtained via statistical analysis of those data matrices. In this study, Correspondence Analysis and Hierarchical Cluster Analysis will be applied to yield such visualizations (see §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, mostly in an attempt to arrive at a more empirical account of lexical meaning. Heylen et al. compared 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)[13]. Gries (2006a) explains that, by bridging the gap between cognitive studies 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 2006a: 57). The integration of a cognitive linguistic framework within a corpus linguistic study is moreover believed to lead to more "theoretical sophistication" (Gilquin 2010: 16). In this section, a "prototype model of category structure" will be proposed as the theoretical basis 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), I will zoom in on the notion of prototypicality so that it can be used in an unproblematic way to further describe and interpret the results presented in the subsequent chapters of this study. In the second part (§2.4.3.2), I will show how Divjak's proposal to opt for a prototype-based categorization for

[13] 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).

low-contrastive verbs expressing abstract concepts also seems to be the better choice for this study. In addition, I will comment on Divjak's two proposals of internal category organization (schematic or radial structure). Just as Divjak, I 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; 1999). One of Rosch's most important findings was that "[m]ost, if not all, categories do not have clear cut boundaries" (Rosch 1999: 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: 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: 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" (Rosch 1999). 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: 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" (Rosch 1999). 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: 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 onwards also (general) linguistics (Geeraerts 2013: 578). As far as cognitive linguistics is concerned, prototype theory is even seen as "one of its cornerstones" (Geeraerts 2006a: 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).

According to Gilquin, 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 et al. 2006: 160-161). Prototypicality was furthermore extended beyond concrete objects to more abstract categories such as past tense and syntactic constructions (Gilquin et al. 2006, referring to 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 2006a). Prototypicality, according to Gilquin, needs to be considered as follows:

> 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 et al. 2006: 180).

By defining prototypicality along these different lines, Gilquin tries to incorporate the four hypotheses uttered by Geeraerts (2006b) 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 (2006b: 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 (2006b: 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, can allow the researcher to compute differences in salience (Geeraerts 2006b: 29). Thirdly, according to the statistical hypothesis, the prototype is that member of a category which is most frequently experienced. Geeraerts (2006b: 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 2006b: 28).

I follow Gilquin's "multi-faceted" view on prototypicality, which incorporates Geeraerts four hypotheses. However, the following question arises: if a prototype-based view on language is taken and claims are made about the semantic relationships within the presumably prototype-based semantic fields, how can one be sure that the chosen method will actually render a prototype-based structure? Given the corpus-oriented scope of this work, the most straightforward way of 'ensuring' that the yielded semantic fields will be prototype-based is to integrate both Geeraerts' second (family resemblance / salience) and the third (statistics) hypothesis. In this way, a cognitivist view on prototypicality – "cognitivists tend to consider the prototype as the cognitively most salient exemplar" (Gilquin et al. 2006: 159) – is united with a corpus-linguistic view which usually considers the prototype as the most frequently corpus-attested item (Gilquin et al. 2006). As Gilquin points out, most of the time, both cognitivists and corpus-linguists assume that salience and frequency coincide with one another (Gilquin et al. 2006). 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 this study, I will not only take frequency as a measure of prototypicality, I will also propose a way to operationalize salience, and I will do so by taking into account the number of overlapping translations. By doing so, I 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, I can apply Geeraerts idea (2006b: 29) that the number of shared attributes (overlapping translations) can be used to compute salience. The principle of overlap will be further developed in §??. In short, I combine the use of frequency – the statistical hypothesis – and overlap

– my operationalization of salience – to determine the status (more prototypical or more peripheral) of the member(s) of the semantic field I 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 the methodology (see Chapter 3). Under the assumption that not every member is equally representative for a category, the method will need to be able to inform about member representativity (this will be done by calculating the distance from each lexeme to its cluster's centroid, see §3.6.3). As far as the family resemblance structure is concerned, it will be integrated by means of the so-called overlap principle. The fuzziness of category boundaries will be dealt with by imposing a minimum threshold for the overlap criterion (see §3.4.3) and the remaining fuzziness will be evaluated by assessing the distance of each lexeme to its cluster's centroid as well as to the centroids of other clusters (see §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 2010a: 150). Given a number of differences between nouns and verbs –verbs are not stable/ time independent, verbs name intangible events, verbs render relational concepts (Divjak 2010a) – it is indeed plausible that "conceptual categories associated with verbs and adjectives function differently from those associated with nouns" (Divjak 2010a). 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. As far as the intangibility of the verb concepts is concerned, Divjak (2010a: 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 2010a: 151). There seems to be no doubt however that "categories for intangible relational concepts also display prototype effects" (Divjak 2010a: 153), as is shown by Schmid (1993); Taylor (1995; 2003); Geeraerts (1985; 1988; 1990) (Divjak 2010a: 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 [...]" (Divjak 2010a: 150).

Since the semantic domain covered in this study also expresses a rather abstract concept (inchoativity), I believe that the above line of reasoning in favor of prototype-based categorization also holds for this study. Divjak herself uses ID tags to set up behavioral profiles for each of the verbs in her study for prototype identification (Divjak 2010a: 158). My 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).

A prototype-based organization for the internal structure of a category seems like a defendable choice; the next question that comes to mind is: what does it look like? (Divjak 2010a: 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 2010a: 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 2010a: 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 2010a: 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 2010a: 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

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. This lack of empirical studies of meaning can be attributed to the typical status of meaning in translation, i.e. meaning as the invariant of translation. However, this alleged invariance of meaning cannot be accepted as a given since investigating meaning in translation could potentially answer the perennial question of the difference between translated and non-translated language. Universal tendencies such as levelling-out and normalization-shining through are well suited to investigate meaning relationships in translation and such studies could indeed even inform the universals research on an explanatory level.

In this Chapter, I 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.

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.

The 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 rely on statistical visualization techniques

that are usually employed in distributional semantics and it will take a prototype-based view on meaning to interpret the statistical visualizations.

In order to apply the SMM for this study, however, two practical issues still need to be solved. First, a way needs to be found 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.

3 Methodology

3.1 Introduction

In this methodological chapter, a technique to visualize semantic fields in translated and non-translated language will be developed. In the previous chapter, I 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. I concluded that this technique could potentially offer a methodological solution for meaning investigation in translation. In this chapter, I will further explore the SMM and see how the technique can now be employed to compare semantic relationships in translated and non-translated language. I will therefore 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 §3.2, the distinction between onomasiology and semasiology will be presented. This distinction is important because it partially determines the interpretation of the visualizations. In §3.3, the corpus that will be used in this study, the Dutch Parallel Corpus, is described. In §3.4, I will give a detailed account of the SMM as it has been developed by Dyvik. In the next §3.5, we will explain my own extensions of the technique. The first extension is an integration of translation direction and the asymmetry of translation into the retrieval task; the second extension focuses on how the output of the retrieval task can be used as an input for a statistical visualization of a semantic field. In §3.6, the first extension of the SMM will be applied to retrieve data sets for the semantic field of *beginnen*/inchoativity in Dutch. In §3.7, the second extension of the SMM is applied via an exploration of a number of statistical methods that will allow for the visualization of semantic fields. In §3.7.1, a first visual exploration of the data on the basis of correspondence analysis is presented before, in §3.7.2, a hierarchical agglomerative clustering is carried out upon the output of the CA. This section also covers

the choice of the distance measure (§3.7.2.1), clustering algorithm (§3.7.2.2) and number of clusters (§3.7.2.3) for the HAC. In the final part of this section (§??), I 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 §3.8 I present a methodological solution to investigate whether the presumed differences between translated and non-translated Dutch on the semantic level might be ascribed to levelling, shining through or normalization on the semantic level. The (changing) prototype-based organization of meaning distinctions within semantic fields of translated and non-translated Dutch and of lexemes within the meaning distinctions revealed by the clusters is based on the calculation of the distances of the clusters to the centroid of the semantic space and of medoids…

3.2 Semasiological and onomasiological perspective

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 my 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, the visualizations would represent the different ways of expressing one and the same concept under study (in our case, the field of inchoativity).

If one wants to discover the different words that can be used to express the concept of inchoativity (onomasiological viewpoint), the best option, in a corpus study such as this one 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 this study starts off

with a single word, i.e. *beginnen*, simultaneously favors a semasiological outlook on meaning. If one wants to explore the different concepts expressed by *beginnen*, the most logical choice would be to start this study with this lexeme itself. Hence, the choice of the initial lexeme *beginnen* allows to take both a semasiological and an onomasiological outlook. I 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).

The visualizations of semantic fields in this work will correspond to the visual output of a statistical analysis via Hierarchical Agglomerative Clustering (see §3.7.2). The different groupings (clusters) in a visual representation (dendrogram) will be considered 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, the lexical items which make up each cluster will be considered 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, represented by its (most prototypical) means of expression. 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 meaningfulness – in terms of meaning distinctions of a central word – of the 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 a double semasiological-onomasiological outlook here: clusters are considered as meaning distinctions of the central word (semasiological outlook) and the lexical items in each cluster are considered as the expressions of the meaning distinction of the cluster (onomasiological outlook). By taking such a double view, the question can be raised 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?).

3.3 The Dutch Parallel Corpus

All data for this study are drawn from the Dutch Parallel Corpus (DPC), which was developed as part of the STEVIN 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 §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.

The DPC offers a number of indisputable advantages. With respect to corpus size the DPC is, to my 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 so-called 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" (Delaere 2015: 59). 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 (De-

laere 2015: 59), 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 the DPC is 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 §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). Delaere rightfully points out that it is important to keep in mind that "these results may have influenced the output results of our corpus queries" (Delaere 2015: 50), since the 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 problem of imperfect lemmatization.

The official web-interface of the DPC[1] displays the results of a search query as concordanced observations. For this study, I used the very user friendly "graphical search engine" developed as part of the COMURE project to access the DPC[2]. 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 informa-

[1]Access to the demo version via http://dpc.inl.nl/indexd.php
[2]Access to the full version (password required) via http://dpcserv.ugent.be/comure/

tion 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 so-called 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 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 proposed method. Given the extreme difficulty of retrieving the source language of a given text post hoc, the observations for which the DPC does not indicate the source language were discarded.

3.4 The Semantic Mirrors Method

In the previous chapter, I concluded that the SMM has the potential to lay bare meaning relationships in translated and non-translated language. The technique was explained on a theoretical level and its usefulness was illustrated 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, I will first present the work flow of the SMM[3] as it was proposed by Dyvik (§3.4.1). After this description of the different stages of the SMM, I will take a step back and explore the prerequisites and assumptions one needs to take into consideration before an SMM can be carried out (§3.4.2). I will further explicitate the rationale behind the overlap threshold (§3.4.3) as a crucial

[3]The SMM as well as the SMM++ in §3.4 were first introduced and described in a less elaborate way in Vandevoorde et al. (2017), an article which is under copyright. Its publisher should be contacted for permission to re-use or reprint the material in any form.

element of the technique which ensures that semantically related lexemes can be separated from semantically unrelated ones.

3.4.1 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[4] . 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:

3.4.2 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" (Dyvik 2005: 31). Based on Dyvik's ideas, and based on the following assumptions (verbatim from Dyvik 2005: 31-32) "each language [will be used] as the 'semantic mirror' of the other". The assumptions Dyvik puts forward are as follows:

- Semantically closely related words tend to have strongly overlapping sets of translations.
- Words with wide meanings tend to have a higher number of translations than words with narrow meanings.

[4]For the sake of clarity, I 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.

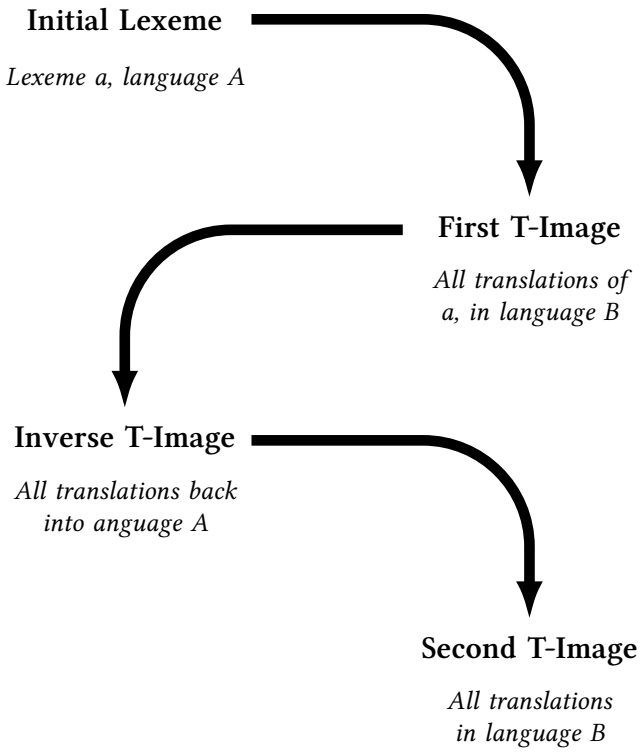


Figure 3.1: Work flow of the SMM

- 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*.
- 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.)
- 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).

3.4.3 Overlap

When the SMM is applied to an initial lexeme, three types of word sense relationships can arise: "related word senses", "unrelated word senses" and "mutually unrelated word senses" (Dyvik 2005: 32). The first step that needs to be taken, is to isolate the mutually unrelated senses of each word (Dyvik 2005: 32) for which the resulting lexemes of the first T-image are used. I 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 3.2), 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 presented in the following sub-sections is based on Dyvik's procedure for word sense isolation.

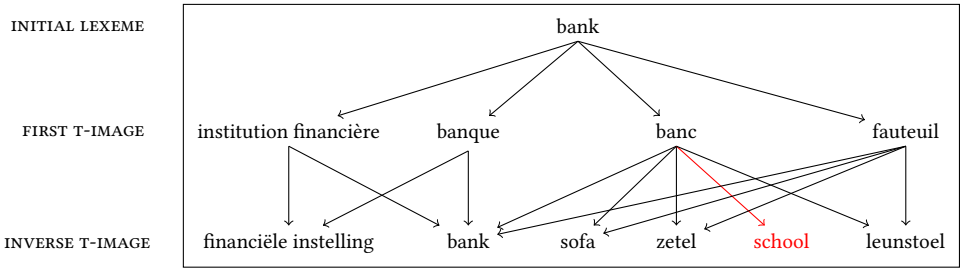


Figure 3.2: Example of the (ficticious) SMM of *bank*

3.4.3.1 Unrelated word senses

The set of translations back into Dutch (the inverse T-image) of *banque* and *banc* only share the initial lexeme *bank* itself in the inverse T-image. *Banque* (Figure 3.3) is connected in the inverse T-image (i.e. 'can be translated back into Dutch as') to *bank* and *financiële instelling*. As a consequence, it could be stated that the inverse T-images *bank* and *financiële instelling* are semantically related to each other (via *banque*):

Banc (Figure 3.4) 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*):

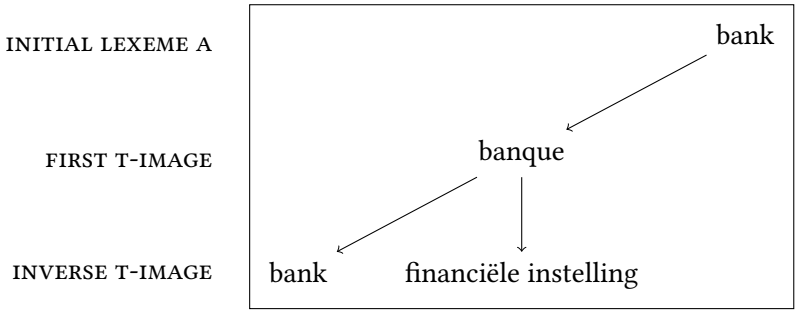


Figure 3.3: Inverse T-image of banque

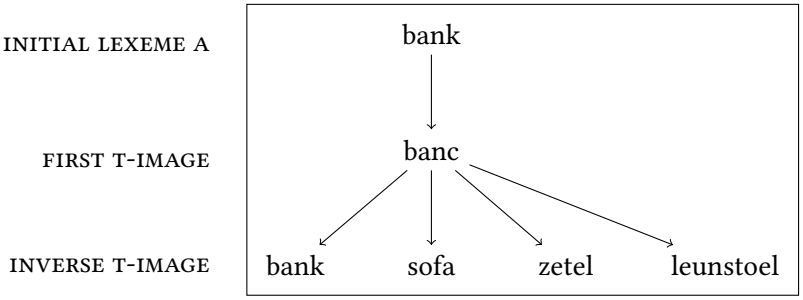


Figure 3.4: Inverse T-image of *banc*

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 Dutch *bank* is contrastively ambiguous between French *banque* [financial institution] and *banc* [seat]). This observation corresponds with Dyvik's assumption (4): the Dutch lexeme *bank* is indeed *homonymous* between *bank* [financial institution] and *bank* [seat]. There is also 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*. Hence, 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*).

3.4.3.2 Related word senses

Looking 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 ??), and that *fauteuil* is connected to *bank, sofa, zetel* and *leunstoel* in the inverse T-image (Figure 3.5). 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.

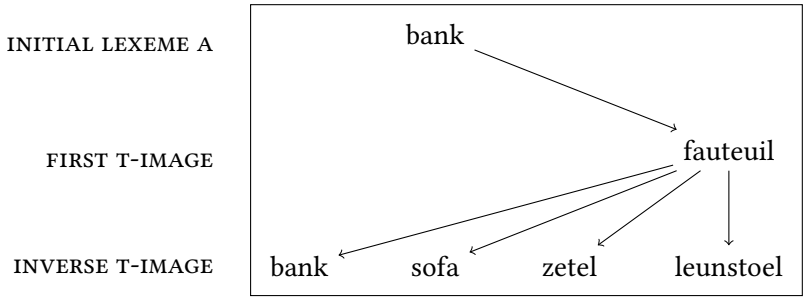


Figure 3.5: Inverse T-image of *fauteuil*

3.4.3.3 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 3.2, the example of the (fictitious) SMM of *bank).* Dutch *school* [school] is a possible translation back into Dutch of the French first T-image word *banc,* in its meaning [school of fishes]. But this latter meaning [schoo] is not a meaning of Dutch *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 3.6). Whereas unrelated senses shared only their initial lexeme in the inverse T-image – enabling a distinction between 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.

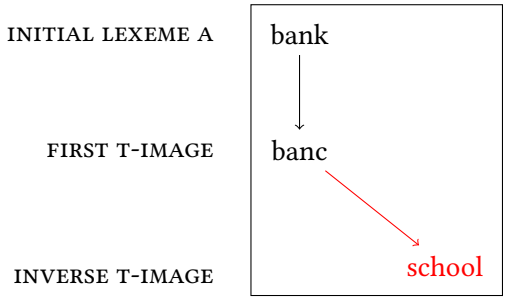


Figure 3.6: Mutually unrelated sense *school*

3.4.3.4 Word sense individuation

The individuation of word senses can now take place: one of the meanings of *bank* [financial institution] can be expressed by *bank* and *financiële instelling*, another meaning of *bank* [seat] can be expressed by *bank, sofa, zetel, leunstoel*. *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, my emphasis, my 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, 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.

3.4.3.5 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 (Dyvik 2005: 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" Dyvik 2005: 29 – 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 it is applied 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 §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 my 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.

3.5 Extended Semantic Mirrors Method: SMM++

The goal of this methodological chapter is to find an adequate way to retrieve lexemes as candidate-members of a semantic field under scrutiny for both non-

translated (original/source) language and translated (target) language and 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 the 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. The SMM is 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 to retrieve candidate-lexemes for a semantic field, but in order to implement 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, I will propose two extensions of the SMM[5]. The first extension is concerned with the integration of translation direction and the asymmetry of translation into the retrieval task (§3.5.1); 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 (§3.5.2).

3.5.1 Extension 1: Translation direction and asymmetry of translation

In the SMM, the translational relation is considered as symmetric, i.e. a relation which exists irrespective of the translation direction. The second T-image results in a set of Language B lexemes which are translations into Language B of the Language A lexemes from the inverse T-image. The second T-image provides the necessary information to establish a semantic field in Language B, 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, and "paired semantic fields in the two languages involved" (Dyvik 2005: 33) 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. When integrating the SMM for research in TS, one inevitably

[5]The two extensions to the SMM in §3.5.1 and §3.5.2 were first introduced and described in a less elaborate way in Vandevoorde et al. (2017), an article which is under copyright. Its publisher should be contacted for permission to re-use or reprint the material in any form.

has to take into account the asymmetric nature of the translational relation as well as the reality of translation direction. This implies that, in my 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.

Two sets of data are therefore created, which can form the basis for a comparison of a semantic field of a lexeme under scrutiny: one data set representing non-translated (original/source) language (in this case non-translated Dutch), and a second data set representing translated (target) language (in this case translated Dutch with English or French as a source language).

Non-translated Dutch and translated Dutch need to be represented by separate sets of data, which furthermore need to be (easily) comparable. In addition to that, the semantic fields created on the basis of these data sets need to consist of lexemes in the same language as the initial lexeme (Dutch). Table 3.1 below shows the original structure of the SMM as it was conceived by Dyvik. In the fourth column, translation direction is added. 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 3.1: 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 3.1, 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.

The first extension thus 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 & Simon-Vandenbergen 2004), I 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. Further references in this book to translated language, will be written as $\text{TransLanguage}_{\text{A}}$ (in this study $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$); referring to the sets 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. References to non-translated (original/source) language, will be written as $\text{SourceLanguage}_{\text{A}}$ (in this study SourceDutch); the underlying data set will be the one obtained in the second T-image with any language A (in this study Dutch) as a source language and any Language B (here: English or French) as a target language.

3.5.2 Extension 2: Statistical implementability of the data sets

In the previous section, I dealt with the asymmetric nature of translation and determined a way to compile sets of translated and non-translated language by extending the existing SMM. The next step is to arrive at comparable visualizations of those sets of lexemes. The information which has so far been obtained only gives the researcher sets of lexemes, but does not propose any kind of organization of those lexemes which could give further information 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 3.2 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*.

A visualization based solely on overlapping t-images is usually realized via Venn diagrams (Dyvik 2011), which tend to get rather complex to interpret. 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 scope of this book to computationally implement the SMM. However, the objective to create visualizations

which can provide more insights into the alleged semantic differences between translated and non-translated language on the basis of the SMM, implies the use of methods 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 is the second major extension to the SMM. If frequency information is now integrated into the previously given fictitious example of bank, the result looks as follows for non-translated (original/source) language bank (Table 3.3) and translated (target) language bank (Table 3.4).

Table 3.2: Overlapping translations of bank (ficticious) in the inverse T-image

is translated as	bank$_{[nl]}$	financiële instelling$_{[nl]}$	sofa$_{[nl]}$	leunstoel$_{[nl]}$
banque$_{[fr]}$	✓	✓	✗	✗
banc$_{[fr]}$	✓	✗	✓	✓
fauteuil$_{[fr]}$	✓	✗	✓	✓

Table 3.3: 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

The occurrence counts in the frequency tables implicitly also contain the number of overlapping translations (or source language lexemes). Hence, the fre-

Table 3.4: 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

quency 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). Advanced statistical techniques can now be applied upon the data sets, opening 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 allow for a visual representation of 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 (Gries & Divjak 2009; Gries 2012; Divjak & Fieller 2014).

3.5.3 Technical fine-tuning

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 this 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 other (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 these instances, but it would not answer one of the main research questions of this book: how to compare semantic relationships in translated language and non-translated language. For this reason, I agree with Dyvik to exclude completely unpredictable translations – translators' idiosyncracies – from our analysis. More specifically, I will apply a frequency threshold of three attestations for every translation, allowing me to work with a manageable number of possible translational pairs. This choice is motivated by pragmatic considerations. Firstly, a frequency threshold below three attestations generates additional manual annotation work, endangering the feasibility of the task. Solutions such as automatic word alignment with GIZA++ did not yield satisfying results due to unsuficient corpus size, and word aligned parallel corpora are not available for the language pairs in this book. Secondly, with respect to the statistical processing of the data, Evert argues against the inclusion of hapax and dis legomena, providing an additional argument in favor of a frequency threshold of three observations:

> [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 cut-off thresholds. Data with cooccurrence frequency f < 3, i.e. the hapax and dis legomena, should always be excluded from the statistical analysis (Evert 2004: 133).

A second restriction of the data is necessary to make sure that the data sets are also acceptably comparable. I 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 fre-

quency 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 ($\text{TransDutch}_{\text{ENG}}$ or $\text{TransDutch}_{\text{FR}}$), 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 3.3 and Table 3.4 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 lead to (i) a selection of a manageable amount of manually controlled data on which a quantitative analysis can be carried out and (ii) the comparability of the two data sets.

3.5.4 Conceptual issue

The application of the SMM++ leads to the creation of comparable data sets of translated and non-translated language. The frequency tables are obtained on the basis of translational data and following a translation-based method. For the data set of non-translated (source) language, both the nature of the data as well as the nature of the method could well be held against it. In this section, I will 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. to 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. Under the assumption that translations can indeed constitute a kind of attribute, 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.

As explained in §3.5.1, the corpus observations which are selected to investigate non-translated (source) language are source language data. As a consequence, translation cannot have affected the use of a specific source language lexeme in its non-translated environment simply because it is not translated. The use of source language data to represent non-translated data is, in my opinion, conceptually acceptable, but one should keep in mind that the mere selection of a text as a source text, i.e. a text selected to be translated, does have a certain impact: 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, sometimes referred to as preliminary norms (Toury 1995). In addition, the lexeme selection method is and remains of course based on a translation-based technique, viz. the SMM++. While its translational nature assures the semantic relatedness, the ungraspable trace of the translational basis of the method on the selection of the lexemes needs to be accepted. 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 visualizations on a different hypothesis which does not rely on translations as semantic features. If, for instance, 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[6].

Some additional steps to keep the possible source language influence to a minimum are taken, ensuring a 'fair' comparison between original language and translated language using the same technique, the same hypothesis and the same data. As a first precautionary measure, I will refer to these data sets and their sub-

[6] 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.

sequent visualizations as SourceLanguage$_A$ instead of OriginalLanguage$_A$. Secondly, I 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, I maximize the neutralization of any possible specific influence of the semantic features (translations) on the visualization of SourceLanguage$_A$.

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

In this section, the SMM++ retrieval task is applied 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 (see §3.3). I will describe how the three resultant data sets were obtained 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. *Beginnen* was chosen as the initial lexeme, because it can be considered 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 (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 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$_{FR}$.

3.6.1 First T-images of beginnen$_{ENG}$ and beginnen$_{FR}$

7

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

The SMM++ was first carried out with English as a pivot language. 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 the criterion of translation direction (Dutch as a source language, English as a target language). Each of the 382 sentences was 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*:

- 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. Three reasons for elimination were distinghuised. 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.

- 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, I chose to disregard the erroneously aligned sentences out of practical considerations.

- 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:

- 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 not addressed in the current study.

- 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:

- 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 reamins a sluggish affair.]

TARGET: While the European Union is gradually preparing 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*. Observations where an inchoative verb is translated by the syntactic structure such to be+ing-form were excluded for further. Although annotation was perfectly possible on the technical side, the inchoative aspect of the structure 'to be+ing-form' is often very subtle (Smith 1997) and open for debate, as the following example (4) clarifies:

- 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 (*een rol spelen* [to play a role]). 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 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. The reason is that one should also envisage and annotate the translation of those patterns into still other patterns in the target language. 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 observations where a verb pattern is proposed as a translation for the lexeme under study, could be seen as a shortcoming of this study; a solution for complex annotations is definitely needed However, in this first application of the SMM++, I reasonably limited the factors of complexity and disregarded 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 3.5) 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 3.5, the lexemes in bold meet the frequency threshold of 3 observations and are selected for further analysis. Table 3.6 gives a summary the first step of the SMM++ retrieval task for beginnen$_{ENG}$.

The retrieval task of the SMM++ was also carried out with French as a pivot language. Table 3.7 summarizes the information of the first T-image of beginnen$_{FR}$:

3.6.2 Inverse T-images of beginnen$_{ENG}$ and beginnen$_{FR}$

The next step of the SMM++ consists in querying the lexemes from the first T-image as source language lexemes in the DPC. For beginnen$_{ENG}$, all English sentences containing each of the 9 lexemes from the first T-image are queried, only those sentences where English is the source language and Dutch the target lan-

Table 3.5: First T-image of beginnen$_{ENG}$ (raw frequencies)

beginnen			
already	1	to embark	2
as from	1	to emerge	1
aspiring	1	to enter	2
beginning (adj)	2	to gain	1
beginning (n)	3	to go ahead	1
first of all	3	to go into	1
fundamental	1	to kick off	1
initial	1	to launch	2
introduction	1	to let	1
nascent	2	to open	5
new	1	to result	1
original	1	to see	1
start (n)	7	to set up	3
start-up (n)	1	to start	171
to adopt	1	to start off	2
to assume	1	to start out	6
to be rooted	1	to start up	5
to bear	1	to take up	2
to begin	89	to talk	1
to come	1	to try	1
to commence	2	to undertake	2
to develop	1	young	1
TOTAL: 336			

Table 3.6: First T-image of $\text{beginnen}_{\text{ENG}}$

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 ranslations 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

guage are selected. For each observation, the translation back into Dutch of the lexeme is annotated, which leads to the summary in Table 3.8:

Table 3.9 summarizes the results of the inverse T-image of $\text{beginnen}_{\text{FR}}$:

With regard to the inverse T-image of $\text{beginnen}_{\text{FR}}$, there are two points which require 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.

The lexeme *prendre son départ* was initially selected as one of the source language lexemes of the inverse T-image of $\text{beginnen}_{\text{FR}}$ (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 de-

Table 3.7: 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			
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	11.	entrer
	2.	commencer	12.	lancer
	3.	d'abord	13.	lancer, se
	4.	début	14.	mettre, se
	5.	débutant (adj)	15.	ouvrir
	6.	débutant (n)	16.	partir
	7.	débuter	17.	prendre cours
	8.	démarrer	18.	(prendre son départ)
	9.	entamer	19.	recommencer
	10.	entreprendre		

scribe 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[8] corpus confirm:

- 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)

- Quant au chemin de fer touristique du Tarn , il prend son départ à l' an-

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

Table 3.8: Inverse T-image $\text{beginnen}_{\text{ENG}}$

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.	beginning (n)	5.	to open
	2.	first of all	6.	to set up
	3.	start (n)	7.	to start
	4.	to begin	8.	to start out
	9.	to start up		
Selected target language lexemes	1.	aanvang	13.	opening
	2.	(allereerst)	14.	oprichten
	3.	begin	15.	opstarten
	4.	*beginnen*	16.	opzetten
	5.	eerst	17.	sinds
	6.	gaan	18.	start
	7.	inzetten	19.	start-
	8.	komen	20.	starten
	9.	krijgen	21.	steeds meer
	10.	maken	22.	van start gaan
	11.	ontstaan	23.	vanaf
	12.	openen	24.	worden

Table 3.9: Inverse T-image of beginnen$_{FR}$

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

> cienne 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:

- Le propos de Laplanche prend son départ , en effet , de l' idée qu' érosliaison 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 *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. As a consequence, further mirroring cannot be carried out for this verbal expression.

A second observation which can be made here is that the final selection of data for $\text{beginnen}_{\text{FR}}$ is proportionally smaller than the selection for $\text{beginnen}_{\text{ENG}}$ – a little over 70%, compared to more than 80% for $\text{beginnen}_{\text{ENG}}$. 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 (Baker 1992: 40). It is an interesting phenomenon which should not be neglected and from which interesting findings can ensue. In this study, for example, no translation into Dutch could be formally indicated in 59 out of 226 observations for the French adverb *d'abor*d (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, observations of translation by omission have to be discarded from this study as zero translations cannot be selected and retrieved as a source language lexeme in the next step of the SMM++.

3.6.3 Second T-images of beginnen$_{ENG}$ and beginnen$_{FR}$

The following step of the SMM++ consists in querying the lexemes from the inverse T-image as Dutch source language lexemes in the DPC. For beginnen$_{ENG}$, the translation back into English of each selected observation of one of the 24 source lexemes is annotated. Recall that the data in the second T-image are selected according to an additional restriction, i.e. translations have to be identical to one of the source language lexemes of the inverse T-image. In practice, there are two implications of this additional restriction for the data set beginnen$_{ENG}$. First, the total number of selected observations is 17 times smaller than the (enormous) total number of queried observations[9], 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 3.10:

Table 3.11 recapitulates the results of the second T-image of beginnen$_{FR}$:

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.

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 *effectuer 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

[9] In order to cope with the large amount of observations, a preliminary statistical word alignment using GIZA++. Every statically word-aligned observation was subsequently manually verified. The author wants to thank Els Lefever for her precious help with the statistical word alignment.

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

[11] 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.

Table 3.10: Second T-image of $\text{beginnen}_{\text{ENG}}$

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[10]			
Source language lexeme(s)	1.	aanvang	13.	opening
	2.	(allereerst)	14.	oprichten
	3.	begin	15.	opstarten
	4.	beginnen	16.	opzetten
	5.	eerst	17.	sinds
	6.	gaan	18.	start
	7.	inzetten	19.	start-
	8.	komen	20.	starten
	9.	krijgen	21.	steeds meer
	10.	maken	22.	van start gaan
	11.	ontstaan	23.	vanaf
	12.	openen	24.	worden
Target language lexemes	1.	beginning (n)	5.	to open
	2.	first of all	6.	to set up
	3.	start (n)	7.	to start
	4.	to begin	8.	to start out
	9.	to start up		

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:

- 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

Table 3.11: 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[11]			
Source language lexeme(s)	1.	aanvang	19.	ondernemen
	2.	aanvangen	20.	ontstaan
	3.	aanvankelijk	21.	op basis van
	4.	aanvatten	22.	openen
	5.	begin	23.	oprichten
	6.	begin-	24.	opstarten
	7.	*beginnen*	25.	opzetten
	8.	doen	26.	sinds
	9.	een aanvang nemen	27.	sluiten
	10.	eerst	28.	start
	11.	gaan	29.	starten
	12.	in werking treden	30.	ten eerste
	13.	ingaan	31.	uitgaan van
	14.	komen	32.	van start gaan
	15.	krijgen	33.	vanaf
	16.	lanceren	34.	vanuit
	17.	maken	35.	vertrekken
	18.	nemen	36.	worden
Target language lexemes	1.	à partir de	10.	entreprendre
	2.	commencer	11.	entrer
	3.	d'abord	12.	lancer
	4.	début	13.	lancer, se
	5.	débutant (adj)	14.	mettre, se
	6.	débutant (n)	15.	ouvrir
	7.	débuter	16.	partir
	8.	démarrer	17.	prendre cours
	9.	entamer	18.	recommencer

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)

- 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[12]. However, because of the restrictions on the second T-image, *ontwikkelen* has been excluded from the analysis.

Zich storten [throw oneself, plunge]

[12]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 $\text{beginnen}_{\text{FR}}$. This is considered as a reliable agreement (Carletta 1996).

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.

3.6.4 Final selection of candidate lexemes

Tables 8 and 9 (summarizing the second T-images of $\text{beginnen}_{\text{ENG}}$ and $\text{beginnen}_{\text{FR}}$) 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 $\text{beginnen}_{\text{ENG}}$ and $\text{beginnen}_{\text{FR}}$ with the inverse T-images, the *common* lexemes of the second T-images of $\text{beginnen}_{\text{ENG}}$ and $\text{beginnen}_{\text{FR}}$ need to be selected. As the summaries of $\text{beginnen}_{\text{ENG}}$ and $\text{beginnen}_{\text{FR}}$ 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 $\text{beginnen}_{\text{ENG}}$ and $\text{beginnen}_{\text{FR}}$. In total, 17 lexemes have been independently selected by both the mirroring of $\text{beginnen}_{\text{ENG}}$ and $\text{beginnen}_{\text{FR}}$. 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][13].

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 9, but renouncing this final restriction of the data set would have two implications.

[13] 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-*.

Firstly, the data of the second T-images of $\text{beginnen}_{\text{ENG}}$ and $\text{beginnen}_{\text{FR}}$ could not be merged, meaning that the data set of SourceDutch would be based on either $\text{beginnen}_{\text{ENG}}$ or $\text{beginnen}_{\text{FR}}$ – which would consequently take away the previously established 'safety mechanism' of merging the two sets in order to eliminate possible target language effects. Secondly, the sets of lexemes whose visualizations will be compared 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 the data sets with respect to their informativity, I opt for the security of comparing likes with likes in the final visualization step by selecting only those lexemes which the SMM++ of $\text{beginnen}_{\text{FR}}$ and $\text{beginnen}_{\text{ENG}}$ have in common.

3.7 Statistical visualization

After the application of the newly developed SMM++ for the retrieval of candidate-lexemes, the final methodological step of statistically analyzing the data is presented in this section. 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. Parallel to the 'natural' step in distributionalist semantics towards statistical methods, an appropriate statistical visualization method will be selected, which takes into account this newly obtained frequency information.

In order to select such an appropriate technique, a careful analysis of the type of data is needed. The data resulting from the SMM++ are resumed in frequency tables, also called matrices[14]. The matrices list observations in the rows; the columns are considered as the attributes or properties of those rows (Baayen 2008: 118). By grouping the observations according to their properties, (hidden) patterns or structure in the data sets can be laid bare. 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). A first visual exploration of the data on the basis of correspondence analysis will be presented in §3.7.1. A visualization of CA represents the first two latent dimensions of the CA. However, for the data in this study, the first two latent dimensions represent less than the

[14] The contingency tables for all data sets can be found in appendices A to F.

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 described semantic field – 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, a combination of Correspondence Analysis with Hierarchical Cluster Analysis is proposed in §3.7.2. 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, a Hierarchical Agglomerative Clustering will be carried on the output of the CA. This means that the obtained coordinates of the CA will be used as an input for the HAC, a procedure which allows to filter out noisy data. Each of the remaining sub-sections of §3.7.2 is concerned with a particular choice which needs to be made before the HAC can be carried out. In §3.7.2.1 and §3.7.2.2, I will put forward the choice of a particular (dis)similarity measure (Euclidean) and clustering algorithm (Ward's). In §3.7.2.3, I will explain the procedure to determine the number of clusters and I will propose a validation procedure for the number of clusters. Finally, section 3.7.2.4 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 3.7.3 I will propose the use of a number of statistical tools to reveal the prototype-based organization of the clusters in a dendrogram and of the lexemes within each cluster. In section 3.7.4, I also put forward two additional analyses which can be of help to interpret the influence of a specific source language on the translated semantic fields: the visualization of the SourceField of the language B and Multiple Correspondence Analysis on the Burt tables of the TransDutch fields. All the analyses were carried out with the open source statistical software R (R Core 2014). While most analyses can be carried out using existing packages in R, I used the svs-package (Plevoets 2015) which contains "various tools for semantic vector spaces" for a number of analyses. I 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 this case, I will use the output of a CA as the input for a HAC).

3.7.1 Correspondence Analysis

Correspondence Analysis, "a special case of multidimensional scaling" (Baayen 2008: 136), seems a good candidate technique to map 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 3.12, 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 3.12: 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 3.7).

Now, in order to be able to visually present the specific geographic position of each of the Dutch lexemes in the rows, their position in the n-dimensional space is reduced to a two-dimensional space. All five Dutch lexemes can be then represented as points in this space (Figure 3.8):

Next, the best fitting two-dimensional space is computed (Figure 3.9). 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

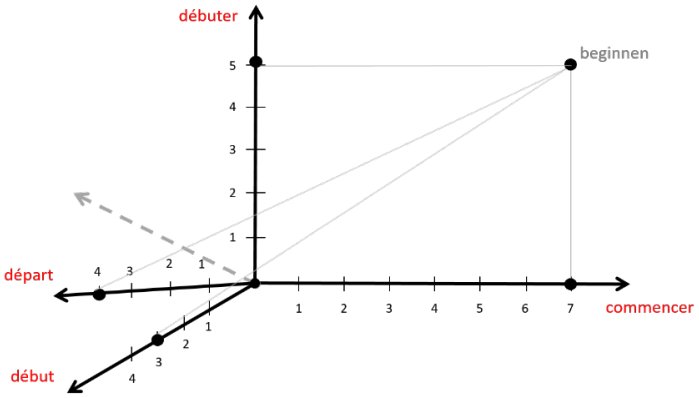


Figure 3.7: Spatial map with n dimensions for beginnen

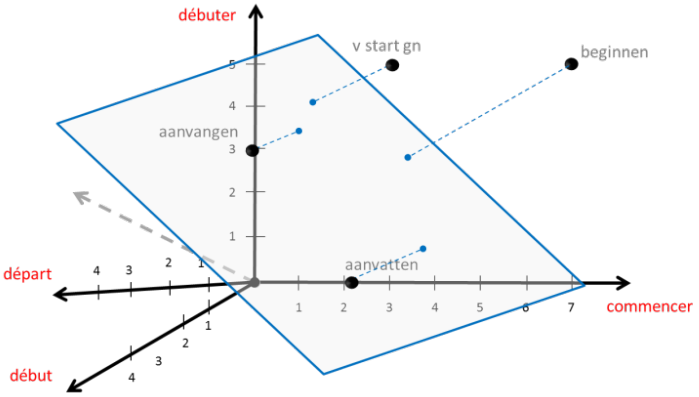


Figure 3.8: Reduction to a two dimensional space for all rows

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 3.9):

When CA is applied to the data sets gathered for this study, a first visualization via CA of the SourceDutch field of *beginnen* looks is obtained (Figure 3.10):

What is immediately striking is the outlying position of *vanaf*. Although the selection of lexemes has been done through a carefully developed technique, described in the previous sections, it is decided to exclude *vanaf* from all data sets. Looking back at the frequency tables (the second T-images of $\text{beginnen}_{\text{ENG}}$ and

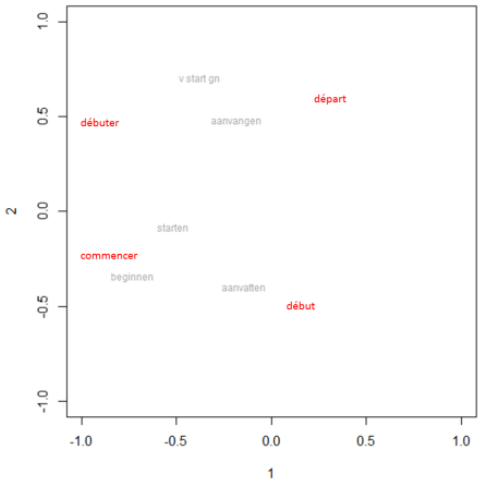


Figure 3.9: Bi-plot for fictitious data matrix in Table 3.12

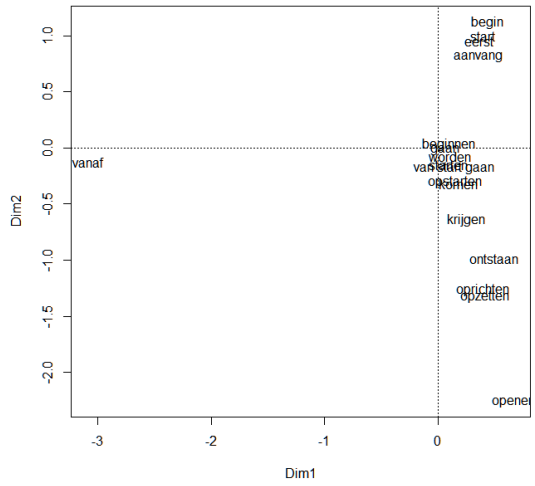


Figure 3.10: First Correspondence Analysis of SourceDutch field for *beginnen*

beginnen$_{FR}$, see appendices E and F) for SourceDutch, it is indeed striking 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$_{FR}$, we also see that the relative weight of *vanaf* is rather high (0.1505792; representing 15 % of the total number of observations) and contributing to a 0.1953608 – over 19% – rise of the total inertia[15] 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 3.17) are indeed more spread out in the two-dimensional space, which will facilitate the interpretation. Based on the above, *vanaf* is removed from all data sets.

Before I further analyze a visualization via CA, the degree of representativeness of the plots with respect to the total variation in each of the data sets needs to be assessed. 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%. The scree plots for SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$, show 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) in order to represent 80% of the total variation visually. This presents a practical problem, however, as 4 or 5 dimensional plots are not easily visualized. Although a visualization via CA for SourceDutch only represents around 40% of the inertia, the visualization in Figure 3.17 is presented as a first, exploratory analysis of the field of SourceDutch.

In Figure 3.17, one large central cluster is observed, situated around the origin (the 'zero-point') of the plot which contains, amongst other lexemes, the initial lexeme *beginnen*. This central cluster is considered 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

[15] "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).

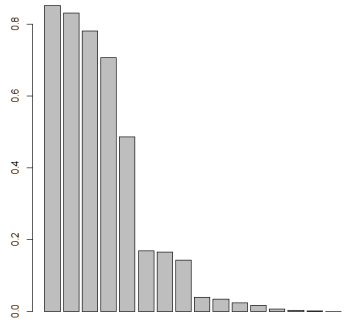


Figure 3.11: Scree plot for SourceDutch

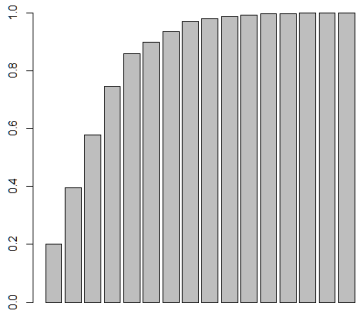


Figure 3.12: Cumulative scree plot for SourceDutch

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[16] counterpart of *begin* and *start*. In the lower right corner, *eerst* [firstly] 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*. *Oprichten* [to establish] and *opzetten* [to set up] are furthermore

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

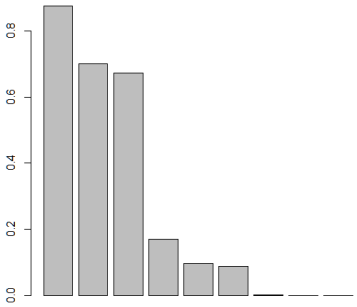


Figure 3.13: Scree plot for TransDutch$_{ENG}$

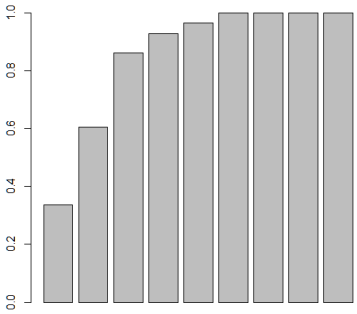


Figure 3.14: Cumulative scree plot TransDutch$_{ENG}$

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, the 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 periphery of the graph. On the basis of the CA, three different clusters can be discerned: 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

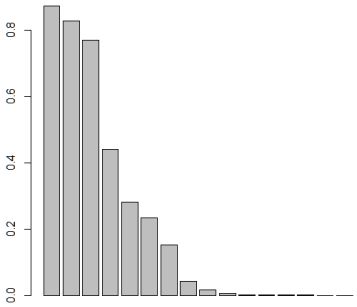


Figure 3.15: Scree plot for $\text{TransDutch}_{\text{FR}}$

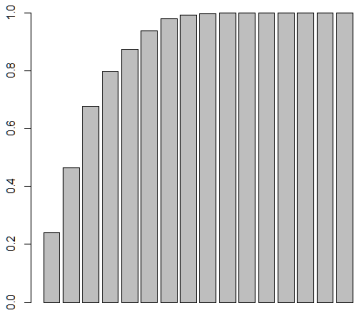


Figure 3.16: Cumulative scree plot for $\text{TransDutch}_{\text{FR}}$

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, the delimitation of clearly distinct clusters can appear difficult. A drawback of CA moreover is that it does not allow 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, the following observations can be made on the basis of this preliminary CA. Firstly, an outlying data point which was distorting the overall interpretation of the data (*vanaf*) was detected and removed. Secondly, the scree

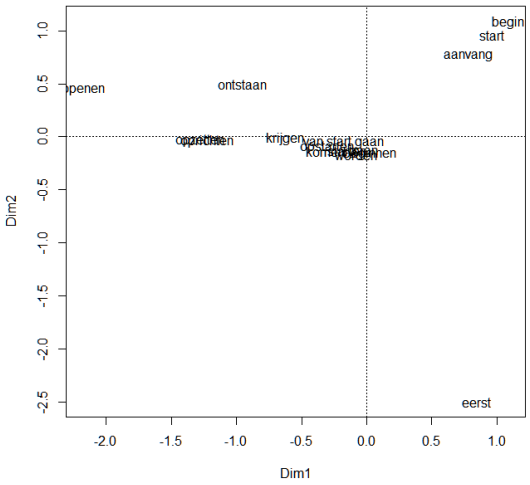


Figure 3.17: Correspondence Analysis of SourceDutch field for *beginnen* without *vanaf*

plots for SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$ showed that for these data sets, 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 visualization. Thirdly, a first exploration of the SourceDutch field on the basis of the CA allowed to formulate some preliminary insights into the semantic field. However, the delimitation of clearly distinct clusters appeared difficult and the exact relations between the lexemes in the central cluster could not further be examined. It was therefore decided to use Hierarchical cluster Analysis for the visualization of the semantic fields of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$. The HCA will be carried out on the output of a CA, a procedure which will be further explained in the next section.

3.7.2 Hierarchical Cluster Analysis

Hierarchical Cluster Analysis (HCA) can be defined as "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). In this section, I will first describe which type of cluster analysis seems the best choice for this study. In addition, as every cluster analysis is crucially dependent on both a particular similarity measure and clustering

algorithm, I will elaborate on these measures in §3.7.2.1 and §3.7.2.2 respectively. Next, I will explain the procedure for determining the number of clusters for which I will rely on the R package pvclust (Suzuki & Shimodaira 2006) (§3.7.2.3). Finally, I propose a validation of the combined choice of a particular similarity measure and clustering algorithm and of the number of clusters in the cluster solution (§??)[17].

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 & Rice 1995; Rice 1996)" (Gries 2006a: 81). The method has also been extensively used by Gries and Divjak (see for example Divjak & Gries 2006; Divjak 2010a; Divjak & Fieller 2014; Divjak & Gries 2006; Gries & Divjak 2009; Gries 2006b; Gries & Divjak 2009; Gries & Otani 2010; Deshors & 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 2010b: 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 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:

> 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 1999: 495).

[17] Exhaustive overviews of the existing clustering techniques can be found in Manning & Schütze (1999: 495-523), Baayen (2008: 138-148), Everitt et al. (2011: 71-110), Gries (2013: 336-349) and Divjak & Fieller (2014).

In order to maintain terminological clarity, I propose to use the following terminology (visualized in Figure 3.18), 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, I 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 in this book) instead of the term *label* proposed by Everitt et al. (2011).

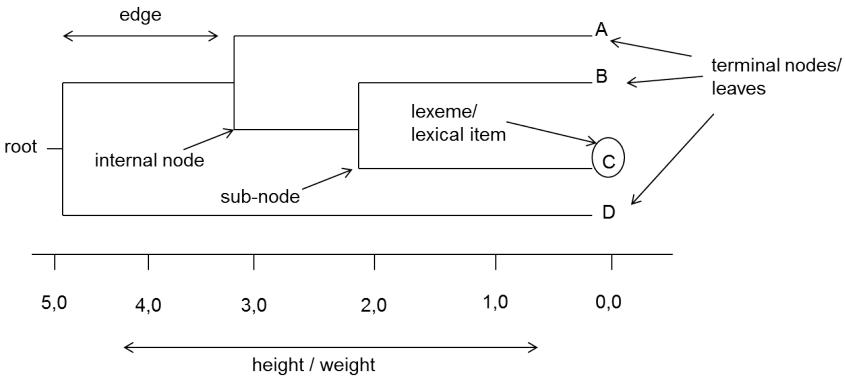


Figure 3.18: Terminological description of a dendrogram (adapted from Everitt et al. 2011: 89)

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 (Manning & Schütze 1999: 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 the visualized result in Figure 3.19, which shows a so-called *chaining effect* when applying divisive clustering for $\text{TransDutch}_{\text{ENG}}$. 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" (Manning & Schütze 1999: 506). In particular, this corresponds to "a model like the Gaussian distribution" (Manning & Schütze 1999: 506). 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 (Manning & Schütze 1999: 506). Because the dendrograms will be interpreted as semantic fields of *beginnen*, organized in a prototype-based manner – with the different clusters representing the meaning differentiations of the lexeme under study – I will prefer a clustering which indeed reflects my intuition that the clusters are centered around a central point and avoids large, elongated clusters caused by a chaining effect.

In summary, I follow Everitt et al. (2011: 92) who state 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 and colleagues 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" Everitt et al. (2011: 92). Figure 3.20 illustrates such a space-dilation effect, of which I will also be wary.

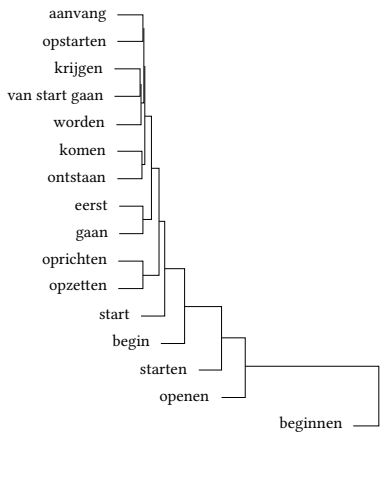


Figure 3.19: Divise clustering of the field of $\text{TransDutch}_{\text{ENG}}$, displaying chaining effect

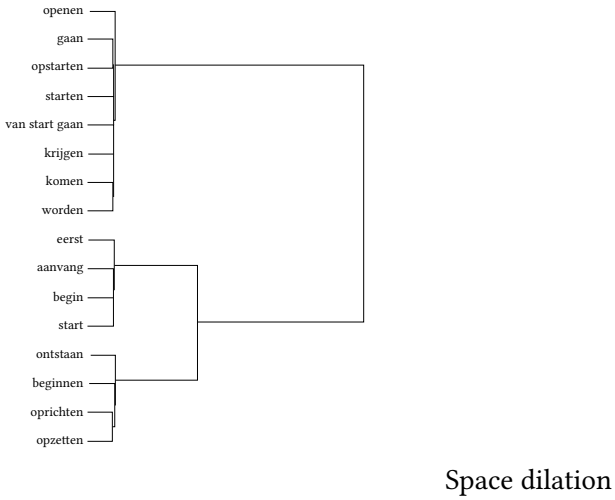


Figure 3.20: Agglomerative clustering of the field of SourceDutch, displaying space-dilation

As a consequence, the data will be further explored with hierarchical agglomerative clustering (HAC). In addition, the HAC will be carried out on the resultant coordinates of the CA. I 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, such an 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 to detect informative dimensions of variation to the detriment of noise, and with HAC meaningful structure(s) in the data cloud can be discerned.

Since I will use the output of the CA as input for a HAC, I 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.

3.7.2.1 (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, but I will limit the 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 2010a: 132). Based on the outcome of the comparison (which will be presented in §??), Euclidean will be chosen as the distance measure for the analyses in this book.

3.7.2.2 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 (Manning & Schütze 1999: 503); the clusters moreover tend to show a chaining effect ((Manning & Schütze 1999: 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 (Manning & Schütze 1999: 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[18] of the 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 2010a: 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, 493–95) 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 this purpose, is not a trivial question, as different algorithms yield different dendrograms. Divjak (2010a: 132), following Speece (1994: 35) emphasizes to choose the algorithms whose " "side-effects" of the mathematical properties […] fit the phenomenon under investigation, and, consequently, yield easily interpretable results".

The single-linkage method is discarded 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 my 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

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

with the complete and average linkage methods in §??. HAC is carried out on the output of the CA with the function pvclust() from the package pvclust which relies on the function hclust() (the choice of pvclust will be substantiated in the next section).

3.7.2.3 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. There is, however, no 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: the computation of the average silhouette width and the use of bootstrap validation.

The optimal cluster solution will be determined by means of a bootstrap validation technique (I 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 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, I will always cut the tree at the highest significant node attaining a confidence level of 95%. 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 be less interpretable than. As a consequence, I 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 the validated cluster solution meets the first two criteria mentioned by Divjak & Fieller for good cut height (acceptable cluster size and meaningful clusters).

3.7.2.3.1 Validation of the number of clusters

In the first part of this section bootstrap p-values were proposed to determine the number of clusters. That procedure is now complemented with two 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 & 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: 138). 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. 2011: 128); values close to -1 imply misclassification and values around 0 suggest that the classification is unclear (Everitt et al. 2011: 128). 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. 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, I will use it as a validation of the HAC. More specifically, I 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 this can be considered 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, 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 (Manning & Schütze 1999: 515). The algorithm then consists in iteratively assigning each data point to the cluster to the center of which it is closest (Manning & Schütze 1999: 515) 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.

In §3.7.2, I substantiated my choice to carry out a HAC on the output of a CA. In addition 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 (Turney & Pantel 2010: 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 this case, the algorithm of LSA (which is usually applied to a document-term matrix) is applied to a source–target language matrix.

In the subsequent comparison, I will include 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 3.13. Because of the high number of combinatorial possibilities – 18 in total – I only compare the combinations for the data set SourceDutch. I selected three validation criteria which have in common their ability to assess the overall strength of the clustering structure: agglomerative coefficient, chaining effect and p-values.

Firstly, the agglomerative coefficient for each combination is calculated. 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 I am using the same data set for each dendrogram in this comparison, the agglomerative coefficients are comparable. An agglomerative coefficient higher than 0.80 is considered as satisfactory.

Secondly, the output for each combination is visually inspected for presence of chaining effect. For this 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 3.13, *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.

Finally, the p-values (which were introduced in §3.7.2.3 to determine the cluster solution) will be used as a third element to assess the overall strength of the clustering structure. To do so, the number of significant nodes (i.e. with a p¯value of 0.95 or higher) in the dendrogram will be counted. Each of the dendrograms presented in the comparison counts 14 nodes. If ≥ 7 nodes in the dendrogram are significant, this will be considered as an indication of a strong overall clustering structure. The number in the third column of Table 3.13 thus indicates the number of significant p-values (≥0.95) on a total of 14 nodes.

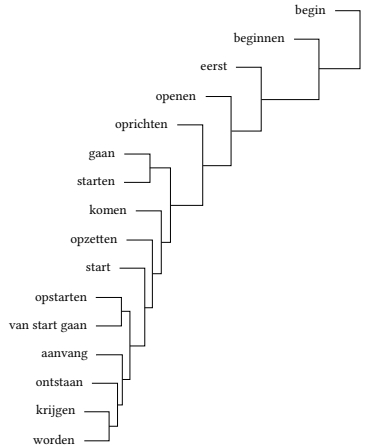


Figure 3.21: Euclidean, Average (1)

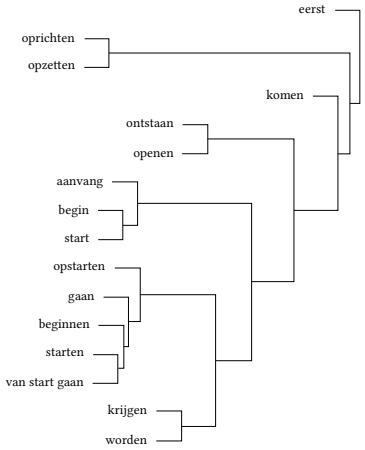


Figure 3.22: Euclidean, Average, on CA (2)

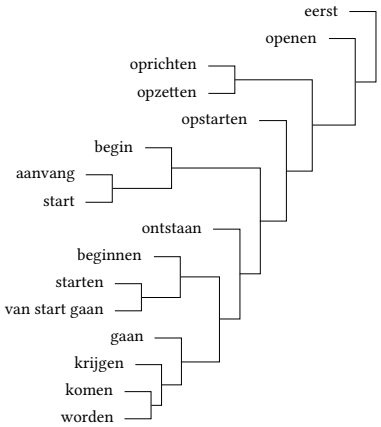


Figure 3.23: Euclidean, Average, on LSA (3)

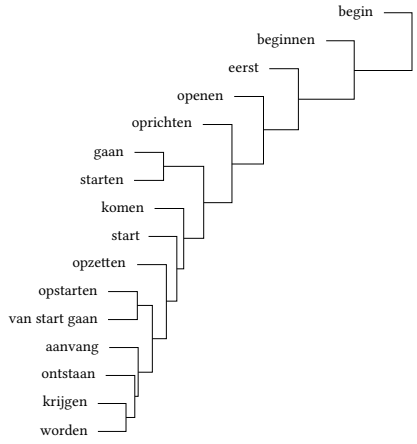


Figure 3.24: Euclidean, Complete (4)

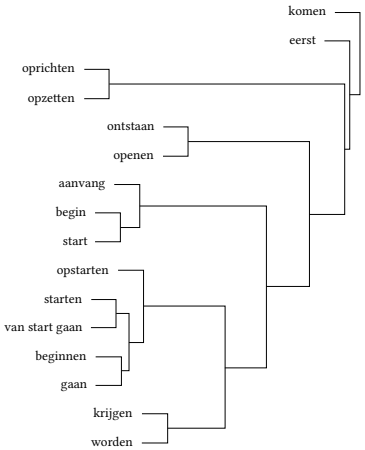


Figure 3.25: Euclidean, Complete, on CA (5)

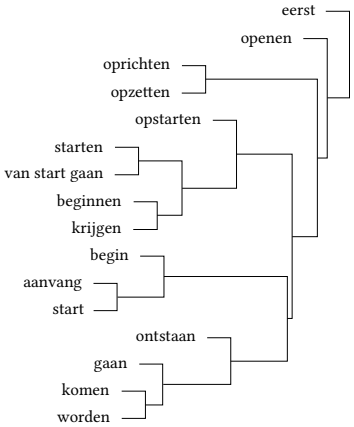


Figure 3.26: Euclidean, Complete, on LSA (6)

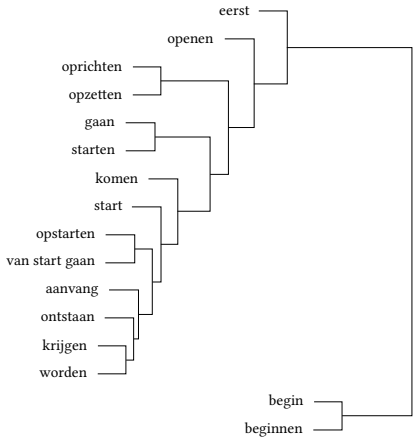


Figure 3.27: Euclidean, Ward's (7)

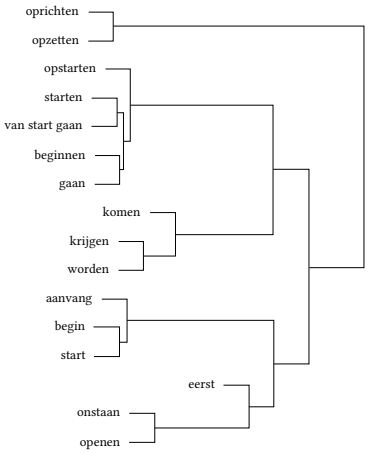


Figure 3.28: Euclidean, Ward's, on CA (8)

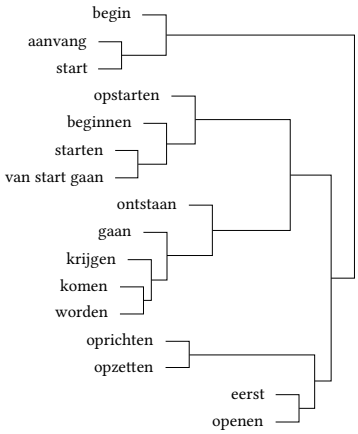


Figure 3.29: Euclidean, Ward's, on LSA (9)

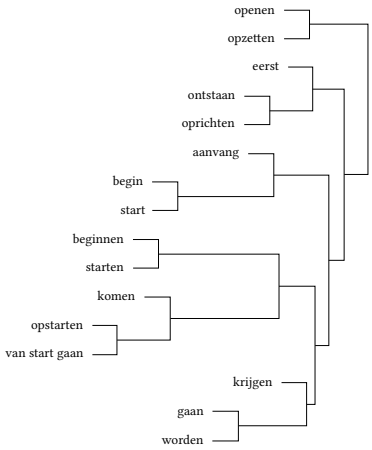


Figure 3.30: Canberra, Average (10)

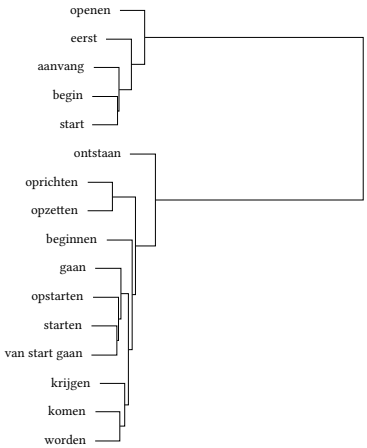


Figure 3.31: Canberra, Average, on CA (11)

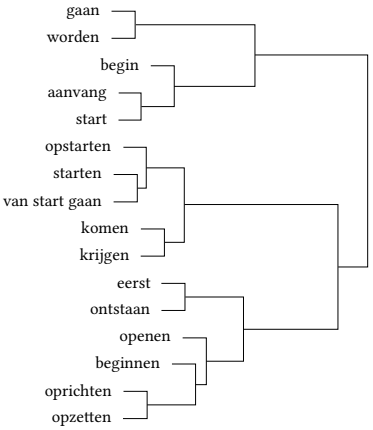


Figure 3.32: Canberra, Average, on LSA (12)

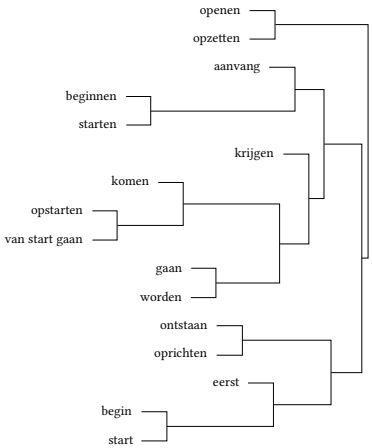


Figure 3.33: Canberra, Complete (13)

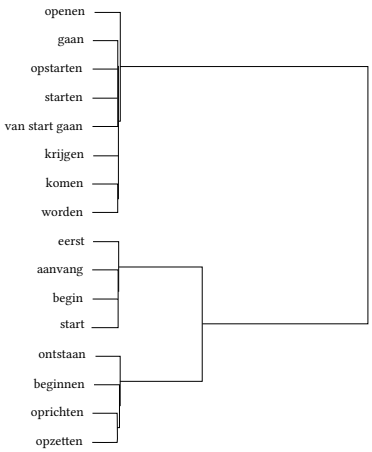


Figure 3.34: Canberra, Complete, on CA (14)

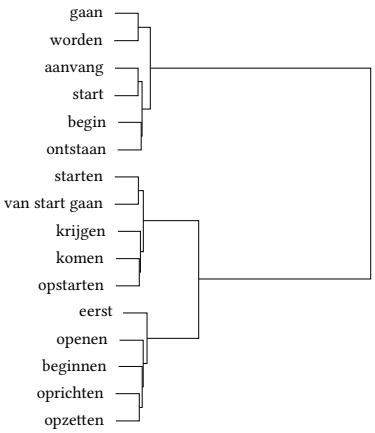


Figure 3.35: Canberra, Complete, on LSA (15)

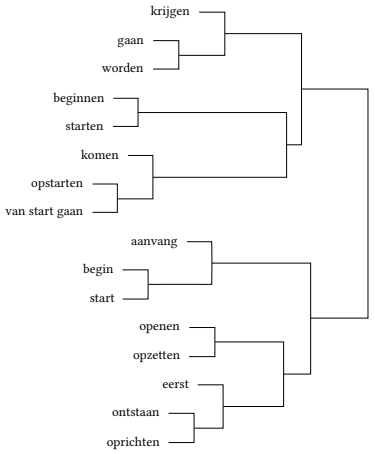


Figure 3.36: Canberra, Ward's (16)

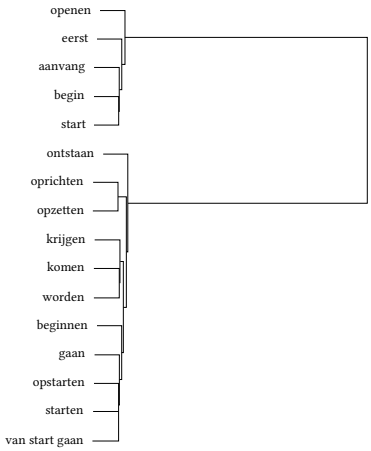


Figure 3.37: Canberra, Ward's, on CA (17)

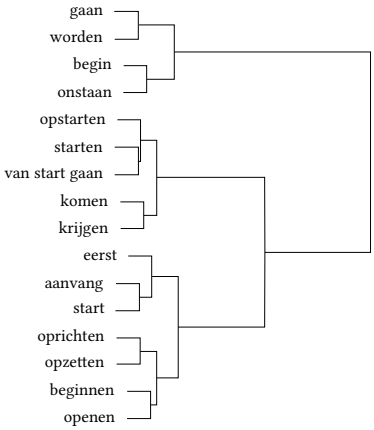


Figure 3.38: Canberra, Ward's, on LSA (18)

Table 3.13: Table 13 Combinatory possibilities of the selected distance measures, clustering algorithms and "spatial maps". †: (+ high space dilation)

	Procedural combination	Agglomerative coefficient	Chaining effect	p-values
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†	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†	7
15	Canberra, Complete, on LSA	0.99	low†	9
16	Canberra, Ward's	0.43	NO	2
17	Canberra, Ward's, on CA	0.99	low†	5
18	Canberra, Ward's, on LSA	0.96	NO	3

Table 3.13 (and the accompanying Figures 30 to 47[19]), shows 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, for the combinations with Canberra distance, none of the analyses carried out on the raw data display a satisfactory agglomerative coefficient.

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 Min-

[19]For each Figure, the number between brackets refers to the number of the combination in Table 3.11 it represents. I will use these numbers to refer to the different combinations (not the Figure numbers).

imum Variance Method) do not suffer from chaining. As for the combinations with Canberra distance, 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. Chaining and space-dilation effects seem not to be limited to the complete linkage algorithm but appear irrespective of the clustering algorithm.

For the combinations with Euclidean distance, all but 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). For the combinations with Canberra distance,we 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, it can be concluded 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 these data. Preference goes to combination 8, because no chaining was observed at all (in combinations 14 and 15 space-dilation in the high nodes and chaining in the low nodes was observed). 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 was explained in §3.7.2.2). On the basis of this comparison, it is decided to apply combination 8 (Euclidean, Wards, on CA) to all data sets of the case study of *beginnen*.

The previous comparison also leads to some more general observations: when Euclidean distance is used, chaining effect, relatively high agglomerative coefficients (although lower than for Canberra) and a high number of significant p-values are more likely to appear. 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 the different spatial maps are taken as point of departure, it appears that analyses on the raw data render low agglomerative coefficients and that analyses on the CA are prone to chaining.

3.8 Statistical approach of universals on the semantic level

In the previous section §3.7, I explained the different decisions that led me to choose HAC carried out on CA to visualize the semantic fields of translated and non-translated inchoativity in Dutch. This methodological development is my answer to the first research question "how to investigate semantic differences?". In Chapter 4, the technique developed in the current chapter will be applied to the field of inchoativity in Dutch. In this way, the second research question: "are there any differences on the semantic level between translated and non-translated language?" will be answered, although of course limited to the differences between translated and non-translated Dutch within the field of inchoativity. However, before the results of the case study can be presented in Chapter 4, more theoretical reflection is needed with respect to the third research question: "if there are any differences between the fields of translated and non-translated Dutch inchoativity, can we ascribe them to any of the universal tendencies of translation?". In the sub-sections §3.8.1 and §3.8.2 below, I will make a number of methodological and conceptual propositions that will enable me to investigate whether the presumed differences between translated and non-translated Dutch on the semantic level might be ascribed the following universal tendencies: levelling out – which has received very few attention from CBTS researchers (see §2.2.2.4) – or normalization and shining through – which take into account specific source and specific target language influence on translated language (see §2.2.2.3). For each of these universal tendencies, a difference will be furthermore made between the semasiological and the onomasiological perspective (see §3.2), leading to different operationalizations for comparison on the semantic level.

3.8.1 Measuring prototypicality effects as a proxy for levelling out

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 §2.2.2.4). On the semantic level, levelling out can be examined on the semasiological level by comparing the variation of the feature *meaning distinctions* in translated language to its variation in non-translated language. For this particular case study, appearance of semantic levelling out on the semasiological level would imply that in translated language, *beginnen* displays fewer meaning distinctions compared to non-translated language. Semantic levelling out could also be investigated on the onomasiological level by comparing the variation of the feature *number of lexical expressions per meaning distinction* in translated language to its variation in non-translated language. For *beginnen*, semantic lev-

elling out on the onomasiological level would manifest itself through the use of fewer lexical expressions to express the different meanings of *beginnen*.

Under the assumption that the meaning distinctions for *beginnen* will be very subtle, I expect that the semantic variation between the fields of translated and non-translated Dutch inchoativity will be small and hence difficult to observe by mere inspection of the clusters in the dendrograms since these clusters are all on an equal par, i.e. they simply represent a partitioning of the lexemes. As a solution to this, I will measure the centrality of each of the meanings (represented as clusters) and focus on possible changes within their prototype-based organization. By determining which clusters are more central in the semantic space and which ones are more peripheral, changes in the prototype-based organization of the meanings within the semantic fields are assessed. Semasiological levelling out will consequently be investigated by looking at the prototype-based organization of the clusters within each dendrogram (SourceDutch, TransDutchENG and TransDutchFR). This will be done by comparing the distances-to-centroids of the clusters within each dendrogram (§3.8.1.1). Onomasiological levelling out will be investigated by comparing the prototype-based organization of the lexemes in each cluster and for each field (SourceDutch, TransDutchENG and TransDutchFR) to each other. This will 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 (§3.8.1.2).

In §3.8.1.3 I will further explore how centroids and medoids may represent different views on prototypes. In addition, each cluster in a dendrogram will also receive a meta-label as a solution to capture the specific meaning distinction of each cluster (§3.8.1.4).

3.8.1.1 Organization of clusters within each dendrogram

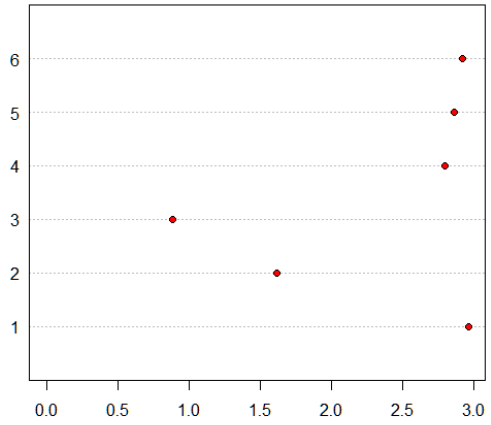
The prototype-based organization of the clusters within each dendrogram will be explored by assessing the distance of each cluster's center (its centroid) to the zero-point of the semantic space. Centroids correspond to the average of all points in the cluster (Tan et al. 2006: 494). They can be calculated 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 Lakoffs

idea (1987, cited by Tyler & Evans 2003) that lexical categories and polysemy networks are structured with respect to their prototypical meanings is accepted, and if Dyvik's basic idea that "semantically closely related words ought to have strongly overlapping sets of translations" is equally accepted, from which it follows that strongly overlapping sets of translation ought to reveal semantic relatedness; then this leads to the assumption 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 be informative about the more prototypical or more peripheral position of each cluster (meaning distinction) in the semantic space (the semantic field it belongs to).

The coordinates of the cluster center (the centroid) are calculated on the output of the CA (i.e. the coordinates of the CA) with the built-in function centers_ca() from the svs-package. Next, the Euclidean distance from each centroid to the zero-point of the semantic space is calculated with the helper function dist_wrt() from svs. Finally, the distances of the centroids to the origin of the semantic space are visualized with a dot chart. The example in Figure 3.39 shows the distance of each of the clusters in the HAC visualization for SourceDutch to the origin of the semantic space. Since the zero-point of the semantic space is held to be the prototypical center, clusters that are closer to the zero-point of the semantic space are considerd as more prototypical and clusters further away from the zero-point as more peripheral.

3.8.1.2 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 3.40). 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 3.40, 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



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 3.39: Dot chart presenting the distance of the cluster centroids to the zero-point of the semantic space of SourceDutch

the abstract prototype contained in the centroid (see §??).

The stability of the cluster membership of each lexeme can also be determined on the basis of this analysis. 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, the hard clustering is somewhat nuanced. 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[20] since it is the average of all points.

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

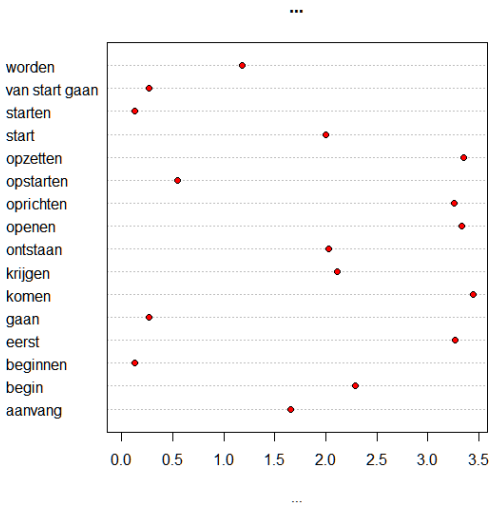


Figure 3.40: Dot chart representing the distance of the lexemes to the centroid of cluster n°4 for SourceDutch

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 (Divjak 2010a: 164). Everitt et al. note that the term medoid was coined by Kaufman & 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, I 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 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.

3.8.1.3 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 prototype is regarded 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 2004: 42). 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). However, it is also possible to 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 the medoid of a cluster is regarded as the prototype of the cluster it belongs to (the particular meaning distinction), this would correspond to Murphy's "best example idea" (Murphy 2004: 42), where "a single prototype could represent a whole category" (Murphy 2004: 42). The medoid then indicates the best example as the prototype of the cluster it belongs to.

3.8.1.4 Manual assignment of meta-labels

A meta-label will be assigned to each cluster in the dendrogram in an effort to name the specific meaning distinction indicated by the cluster. There are several options to arrive at such a label. Firstly, either the lexeme closest to the centroid or the medoid of each cluster can be selected as its meta-label. However, since only 16 lexemes will be making up the dendrograms, several small clusters are to be expected (with 3 or fewer members). 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, other quantitative techniques can be applied in an attempt to provide

supplementary information about each cluster. This would, however, require an expansion of the amount and nature of annotated data in the 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 would consist in coding each item occurring in each of the sentences for a number of variables, known as ID tags)[21]. 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, I 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 §3.3). 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). 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 I propose[22]. 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

[21] While such an analysis would have certainly yielded new insights, it could not be carried out within the scope of this study.

[22] 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.

consequently allow for more fine-grained descriptions of the clusters – while simultaneously adding interpretative power – I did not further investigate this option within the purview of this study, mainly because the main focus of this book is to explore as many potentialities as possible of translational data 'alone' for semantic description, without using any additional annotative information in the analysis.

3.8.2 Semantic fields of commencer and to begin

The investigation of semantic normalization and shining through on both the semasiological and the onomasiological level requires a number of additional visualizations.

On the semasiological level, target language influence on the meaning distinctions in translated Dutch inchoativity (**semasiological normalization**) will be investigated by comparing the meaning distinctions in translated Dutch to those present in non-translated Dutch, for which the visualizations are available the basis of the methodology clarified above. Source language influence on the meaning distinctions in translated Dutch inchoativity (**semasiological shining through**) will be investigated by comparing the meaning distinctions in translated Dutch to those present in the source languages. This will be done by visualizing the semantic fields of the closest equivalents of *beginnen* in the source languages of $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$, viz. SourceEnglish *to begin* and SourceFrench *commence*r. The meaning distinctions in the fields of *to begin* and *commencer* are compared to those present in $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$ respectively, to see whether the specific meaning distinctions within the semantic fields of SourceEnglish and SourceFrench might have influenced the organization of the meaning distinctions in $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$. The resulting semantic spaces of inchoativity in French and English are independent of TransDutch and correspond to the second T-images[23] of *commencer* and of *to begin.*

With **onomasiological normalization**, I refer to the possible influence of non-translated Dutch on the prototype-based organization of the lexemes within each meaning distinction of *beginnen* in translated Dutch. This can be assessed by comparing the prototype-based organization of the lexemes in each meaning distinction in SourceDutch to the organization of the lexemes in each mean-

[23]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. 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 $\text{beginnen}_{\text{FR}}$ and $\text{beginnen}_{\text{ENG}}$).

ing 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. **Onomasiological shining through** would manifest itself as an influence of the source language on the organization of the lexemes within each meaning distinction of *beginnen* in translated Dutch. In order to asses such an influence, 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 – will be visualized together with the Dutch target language lexemes. In this way, it can be observed how the specific organization of the lexical items within 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, I 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 HAC, enabling us to visually inspect which Dutch target lexemes are associated with which French or English source lexemes.

3.9 Conclusion

In this chapter, I have proposed a methodology to investigate semantic differences between translated and non-translated language. The method is an extension of an existing method (the SMM); it is corpus-based, uses statistical visualization techniques and consists of two parts (two extensions to the SMM). The first extension allows the potential user of the method to select candidate-lexemes for a semantic field. This selection mechanism (retrieval method) is translation-driven and uses the different translational statuses (either source or target language) of parallel corpus data. The second extension to the SMM proposes a way to visually inspect the retrieved data sets via a combination of CA and HCA. First, CA is applied in order to construct a low-dimensional semantic space of the data.

Second, HAC is applied in order to distinguish clusters of lexemes within the semantic spaces. The technique is calibrated by the Euclidean distance measure and Ward's Minimum Variance Method as the amalgamation rule. In this methodological chapter, the way was furthermore paved to investigate levelling out, shining through and normalization on both the semasiological and the onomasiological level.

In the next chapter, I will apply the two extensions of the SMM to the semantic field of inchoativity in Dutch. The comparison of different visualizations representing the semantic fields of SourceDutch, $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$ will enable me to tackle the second question I aim to answer with this study: "Are there any of the (universal) tendencies of translation that also apply to the semantic level?" as well as the third one: "If there are differences on the semantic level, can we ascribe them to any of the (universal) tendencies of translation?".

4 Results

4.1 Introduction

The outline of this chapter is as follows. In §4.2, §4.3 and §4.4, I 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, yielded on the basis of the methodological procedure developed in the previous chapter. 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. 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. Finally, (v) an in-depth 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. The meta-labels that I will assign should be understood as a post-hoc, interpretative tool, applied to enhance my understanding of the rendered dendrograms. It should be clear that my attempt to present such 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. In §4.5, §4.6 and §4.7 I will present my

insights with respect to tendencies of levelling out, shining through and normalization each time on both the semasiological and the onomasiological level. The interpretations of the fields of SourceDutch, $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$ described in the previous sections will be used as a basis here. 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.

4.2 SourceDutch

4.2.1 Results of the Hierarchical Agglomerative Cluster analysis

Following the procedure described in Chapter 3, I carried out a HAC on the output of a CA. I first applied 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 4.2) shows that at least 5 dimensions are needed to represent more than 80% of the variation:

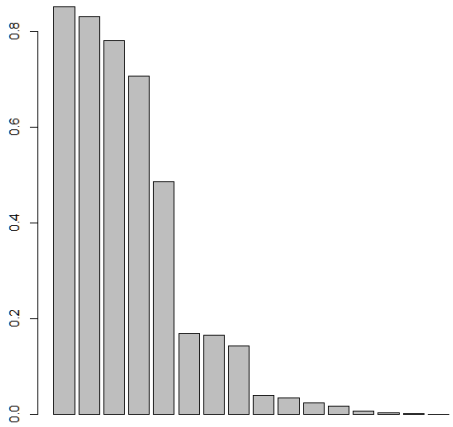


Figure 4.1: Scree plot for SourceDutch

On the basis of this scree plot, I reduced the number of dimensions of the CA to 5. This step is important to avoid noisy (less informative) data patterns. A

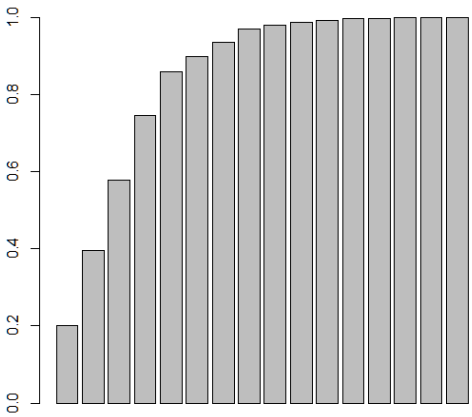


Figure 4.2: Cumulative scree plot for SourceDutch

HAC was then carried out on the output of the CA. The cut-off point was set at a height of 4 (following the rationale described in Chapter 3)[1], 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]. I consider the result presented in Figure 4.3 as a possible visualization of a semantic field of *beginnen*/inchoativity in SourceDutch.

In order to validate the chosen cluster solution with 6 clusters, I calculated the average silhouette width. I obtained 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 §3.7.2.3).

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

[1]Note that with pvrect(), which cuts off each cluster at the highest possible node with a significant p˙value – the same cluster solution would have been obtained.

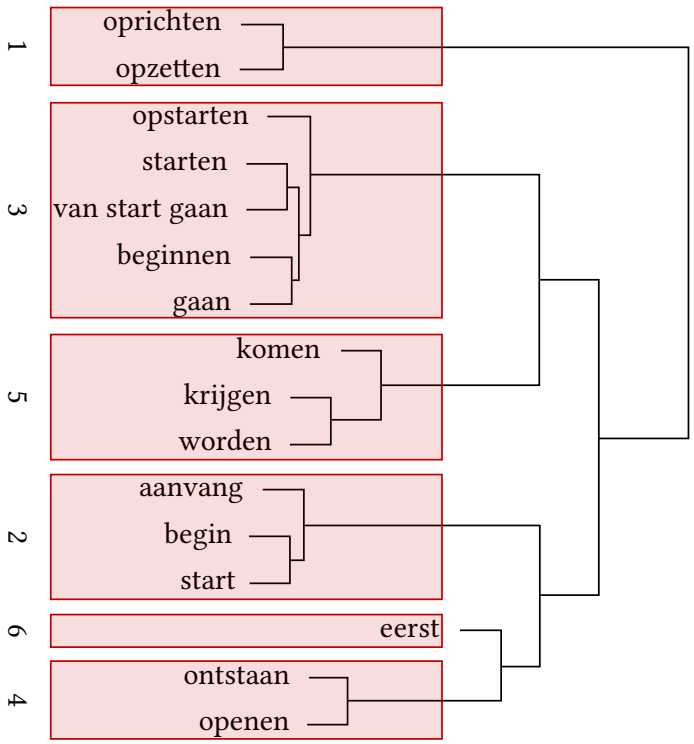


Figure 4.3: Dendrogram representing a semantic field of *beginnen*/inchoativity for SourceDutch

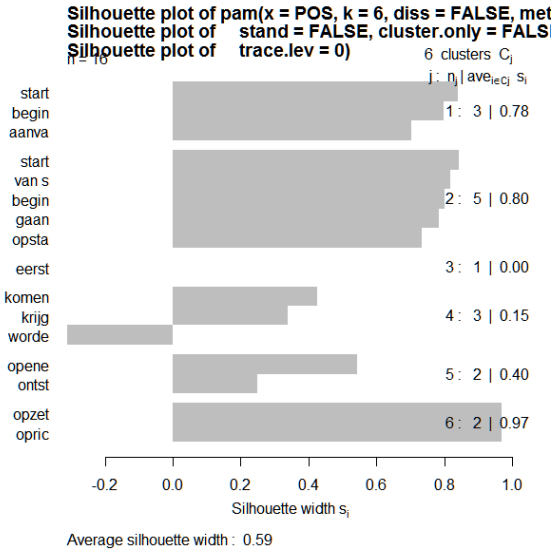


Figure 4.4: Average silhouette width for cluster solution with 6 clusters for SourceDutch

```
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
```

Figure 4.5: K-means clustering with 6 clusters for SourceDutch

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, I consider my cluster solution for SourceDutch as a good classification. In addition, as a result of the K-means clustering it can be conclused that the clustering of the polyfunctional verb *worden* seems to be uncertain.

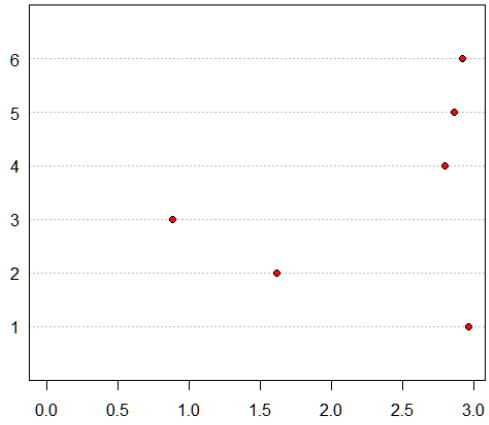
4.2.2 Prototype-based organization of the clusters in the dendrogram (semasiological level)

In order to obtain more information about the prototype-based organization of the clusters (meaning distinctions) within each dendrogram, I determined the distance of the centroids of each cluster to the origin or zero-point of the semantic space (the prototypical center). The centroids were subsequently mapped onto a dot chart (Figure 4.6). The cluster closest to the zero-point is considered as the most central one in the semantic space.

Note that the numerals on the y-axis of the dot chart in Figure 4.6 were 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 4.6. 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)

Next, the prototype-based organization of the lexemes within each cluster was inspected by measuring the distance of the lexemes of each cluster to the centroid of the cluster they belong to. In addition, I calculated the medoid of each cluster. 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 I 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.



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 4.6: Dot chart presenting the distance of the cluster centroids to the zero-point of the semantic space of SourceDutch

4.2.3.1 Centroids

Each of the six dot charts (Figures 56 to 61) 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 allows me 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*).

Since the difference in distance of the members of a same cluster to their cluster's centroid is often minimal, I used the calculated distances (which are represented by the dots in the dot charts) to evaluate the distances to the centroids (see appendix H).

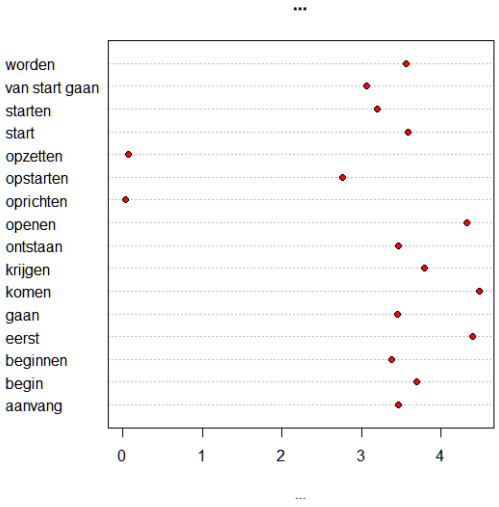


Figure 4.7: Cluster n°1 for SourceDutch

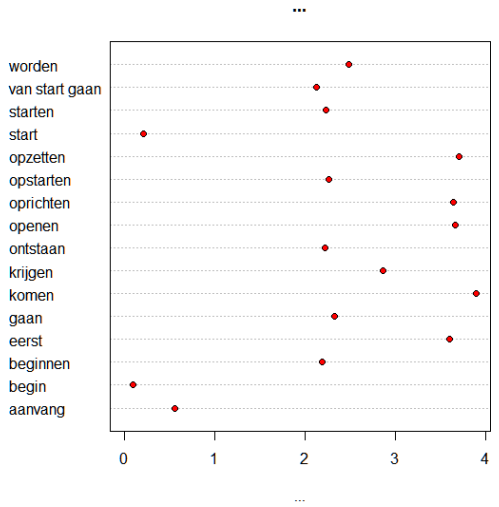


Figure 4.8: Cluster n°2 for SourceDutch

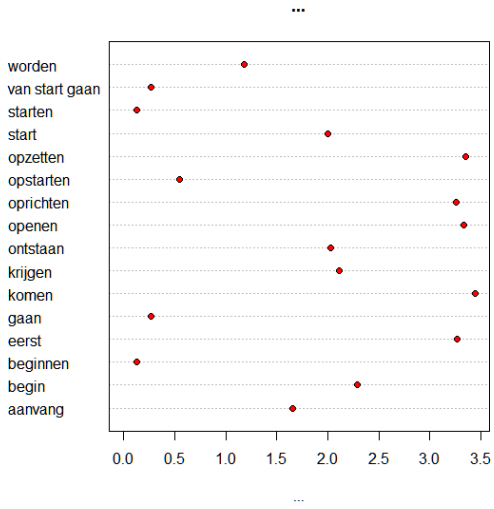


Figure 4.9: Cluster n°3 for SourceDutch

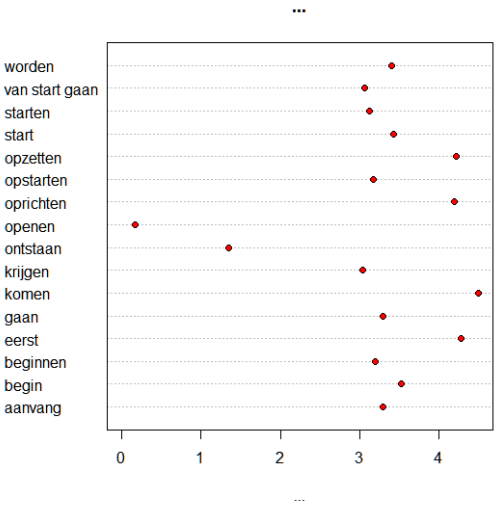


Figure 4.10: Cluster n°4 for SourceDutch

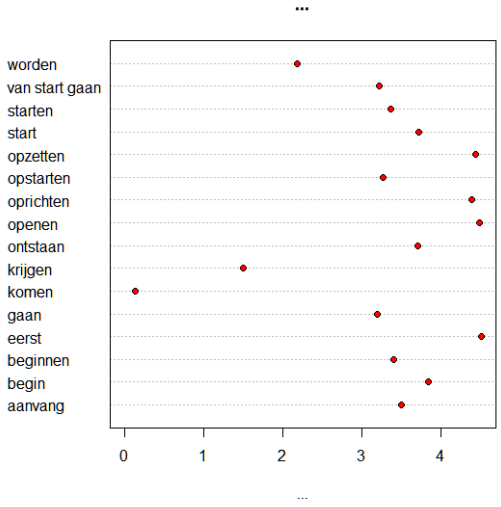


Figure 4.11: Cluster n°5 for SourceDutch

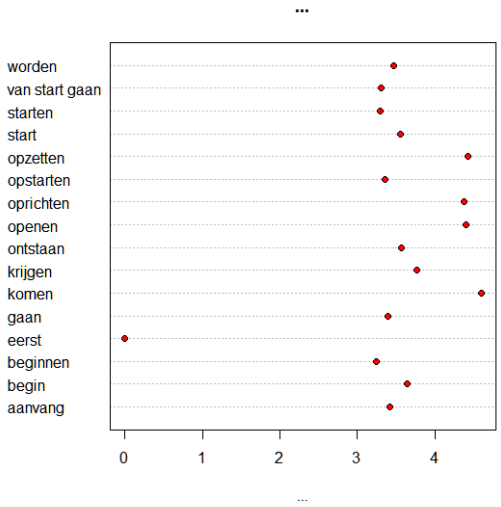


Figure 4.12: Cluster n°6 for SourceDutch

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, four lexemes are very close to the zero point. *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 4.10 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 2010a: 164). The medoids for the clusters are summarized in Table 4.1 and compared to the lexeme closest to the centroid as determined above. The table shows that the medoid and the lexeme closest to the centroid never converge (clusters n°1, 4 and 6 are disregarded since they have only one or two members):

Table 4.1: Comparison of medoids and lexemes closest to the centroids for SourceDutch

Cluster	Medoid	Lexeme closest to centroids
Cluster n°2	Start	Begin
Cluster n°3	Starten	Beginnen
Cluster n°5	Krijgen	Komen

The divergence between the closest lexeme to the centroid and the medoid of a cluster for all clusters increases the uncertainty about which lexeme can be considered as the most central one. In addition, the difference in distance to the centroid is minimal for some clusters, especially for cluster n°3 (*beginnen* vs *starten*) and 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.

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 they have been linked to different views on prototype. As a consequence, neither the lexeme closest to the centroid nor the medoid will be used 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

I will now provide an interpretation of the visualization representing a semantic field of *beginnen*/inchoativity for SourceDutch [2]. 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 I will assign should be understood as a post-hoc, interpretative tool, applied to enhance my understanding of the rendered dendrograms. Note that I do not consider the meta-labels as a validation of the discerned cluster organization – if this had been my intention, I should have determined the labels beforehand. As determined in §3.8.1.4, 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).

I 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.

[2] Substantial parts of the interpretations of the visualizations of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$ in §4.2.4 §4.3.4 and §4.4.4 of this book have been first presented in (Vandevoorde et al. 2017), an article which is under copyright. Its publisher should be contacted for permission to re-use or reprint the material in any form.

First, on the semasiological level his 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, Figures 56 to 61 – which depict the distances of the lexemes to each of the centroids of the other clusters – show 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 the SMM++ retrieval task since it is considered as the most prototypical expression of inchoativity (based on corpus frequency and etymological age). The cluster containing *beginnen* is believed to hold the most prototypical expressions of inchoativity.

Cluster n°3 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 my 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 1996b: 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, my emphasis).

The subdivision observed in the (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, 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. Recall that the two lexemes closest to the centroid of this cluster are indeed *beginnen* (0.1254173) and *starten* (0.1264550); the minimal difference in distance to the centroid between these two lexemes fur-

ther 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, which classifies all the lexemes of cluster n°3 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"). According to the lexical-semantic database Cornetto (Vossen et al. 2008), *gaan* [to go][3], which somewhat oddly seems to be clustered with *beginnen*, is 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 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. The label of REFERENCE CLUSTER/GENERAL ONSET is assigned 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. An additional meaning distinction is furthermore discerned within this cluster between ACTION verbs (to which I 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, 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 Dale et al. 2015). Although the majority

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

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 this 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 this 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[4] (e.g. a noun is more likely to be translated by a noun, and a verb by a verb), which could explain why the translational method favors a word-class dependent clustering of lexemes. Based on the previous reflection, the meta-label GENERAL ONSET (NOUN) is chosen for cluster n°2. GENERAL ONSET indicates 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*. I 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:

- 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).

- 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, all emphases are mine).

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

[4]but not word choice

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. I will therefore add the label SPECIFIC ACTION to cluster n°1.

The lexemes *komen* [to come], *krijgen* [to get], *worden* [to become] in cluster n°5 share the semantic characteristic that their inchoative aspect is non-lexicalized. By this I 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):

- '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, my 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:

- 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, my 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:

- 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, my 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 in the set of candidate lexemes does not cluster with any other lexemes, further substantiating the previously made observation that the method favors word-class dependent clustering.

In sum, I labeled the different meaning distinctions (clusters) within the semantic field of *beginnen*/inchoativity as follows (see Figure 4.13): 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, I 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 my 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.

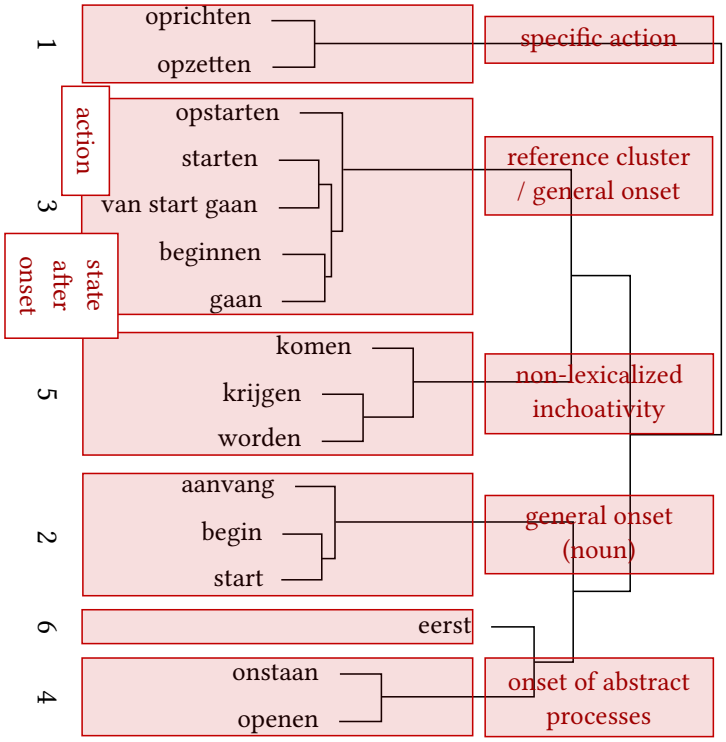


Figure 4.13: Dendrogram representing a semantic field of *beginnen*/inchoativity for SourceDutch with meta-labels

4.3 TransDutch$_{ENG}$

For the description and interpretation of TransDutch$_{ENG}$, I repeated the steps carried out 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 4.14 and Figure 4.15. The number of dimensions of the CA is reduced to 4[5].

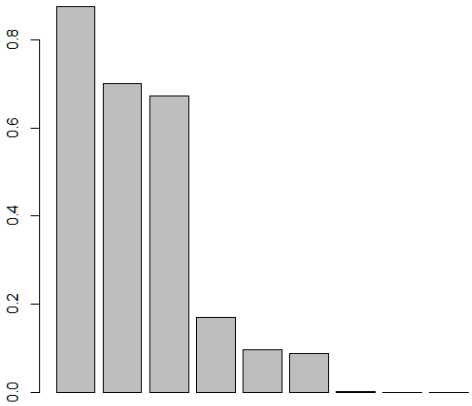


Figure 4.14: Scree plot for TransDutch$_{ENG}$

A HAC was carried out on the output of the CA. The cut-off point was set at a height of 2, offering a cluster solution with 6 clusters[6]. 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 holds *komen* [to come] and *opstarten*

[5]Although 3 dimensions seemed to suffice here to represent more than 80% of the variation, I opted for 4 dimensions, which is the minimum number of dimensions required to carry out pvclust() in the next step of this analysis.

[6]Note that –pvrect() would have yielded same cluster solution.

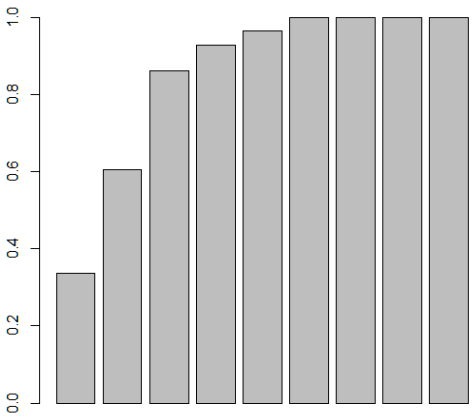


Figure 4.15: Cumulative scree plot for TransDutch$_{ENG}$

[to start up], cluster n°5 consists of *ontstaan* [to come into being] and *openen* [to open] and cluster n°6 contains *begin* [beginning]. I consider the result presented in Figure 4.16 as a possible visualization of a semantic field representing *beginnen*/inchoativity in TransDutch$_{ENG}$.

The chosen cluster solution was validated on the basis of the average silhouette width. For a solution with 6 clusters for TransDutch$_{ENG}$ I obtained an average silhouette width of 0.57, which I consider to indicate a good classification.

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

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, it was concluded 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)

The centroid of each cluster was calculated and its distance to the zero-point of the semantic space was assessed by mapping the centroids onto a dot chart

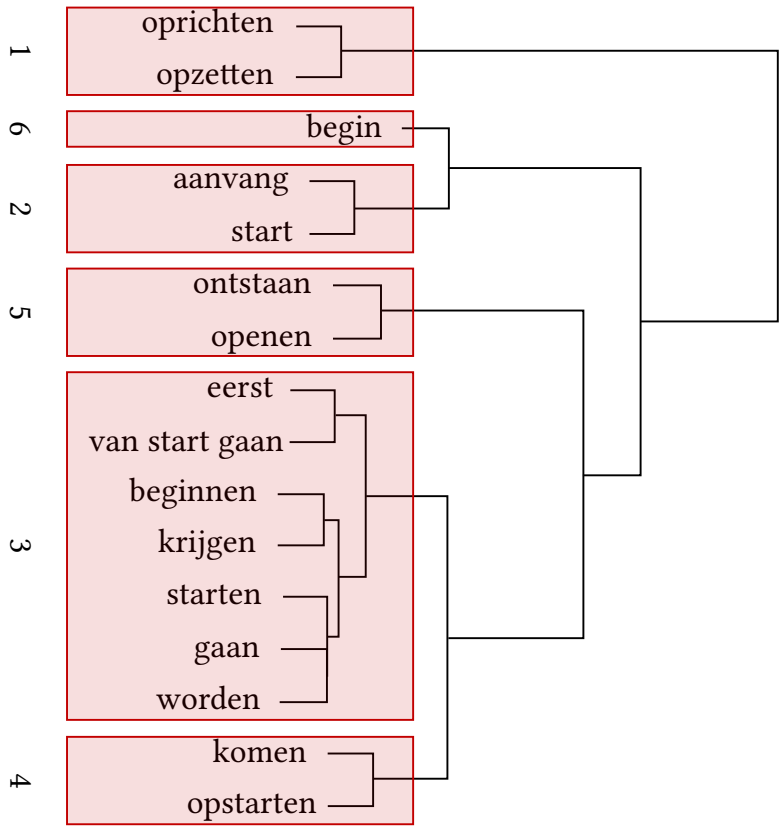


Figure 4.16: Dendrogram representing a semantic field of *beginnen*/inchoativity for TransDutch$_{ENG}$

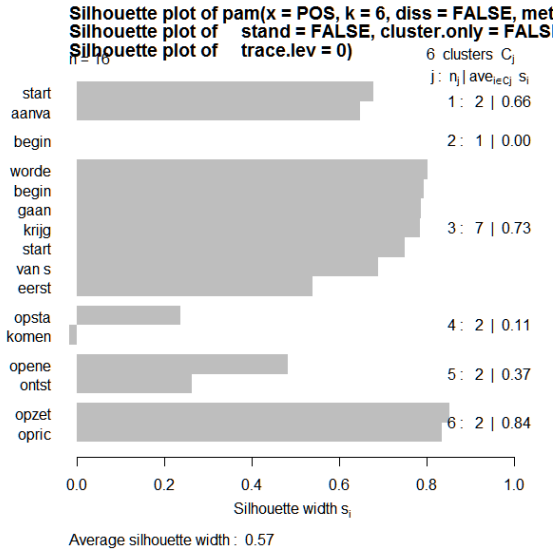
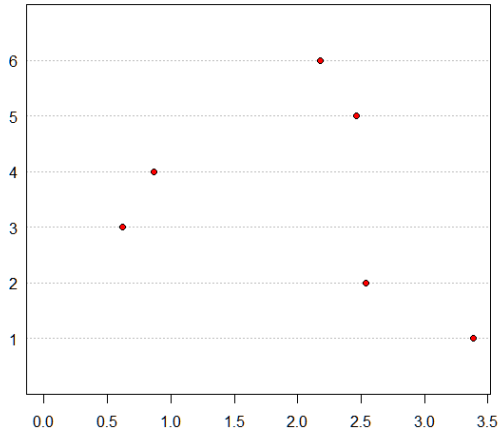


Figure 4.17: Average silhouette width for cluster solution with 6 clusters for $\text{TransDutch}_{\text{ENG}}$

```
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
```

Figure 4.18: K-means clustering with 6 clusters for $\text{TransDutch}_{\text{ENG}}$

(Figure 4.18). The content of each cluster number in the dot chart was summarized in the table accompanying Figure 4.18[7]:



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

Figure 4.19: Dot chart presenting the distance of the cluster centroids to the zero-point of the semantic space of TransDutch$_{ENG}$

The dot chart shows 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.

[7]Parallel to SourceDutch, the numerals on the y-axis of the dot chart in Figure 4.18 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 Prototype-based organization of the lexemes within each cluster (onomasiological level)

The prototype-based organization of the lexemes within each cluster was examined by measuring the distance of the lexemes within each cluster to the centroid of the cluster they belong to, as well as by calculating the medoid of each cluster. Both measures were used to determine which lexical item can be considered as the most prototypical expression of the cluster it belongs to.

4.3.3.1 Centroids

The dot charts in Figures 4.20 to 4.25 represent the distance of all the lexemes to the centroid (the abstraction of the prototype) of a particular cluster.

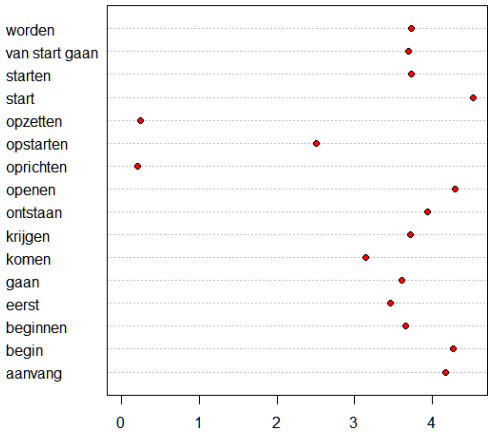


Figure 4.20: Cluster n°1 for TransDutch$_{ENG}$

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

The calculated distances for cluster n°1, show that *oprichten* is slightly closer to the cluster's centroid (0.2004520) than *opzetten* (0.2476172). As for cluster n°2, *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)

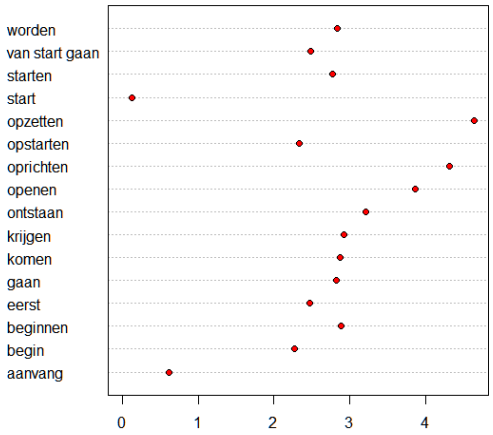


Figure 4.21: Cluster n°2 for TransDutch$_{ENG}$

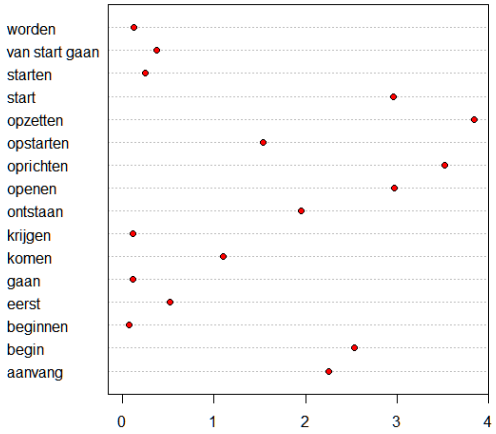


Figure 4.22: Cluster n°3 for TransDutch$_{ENG}$

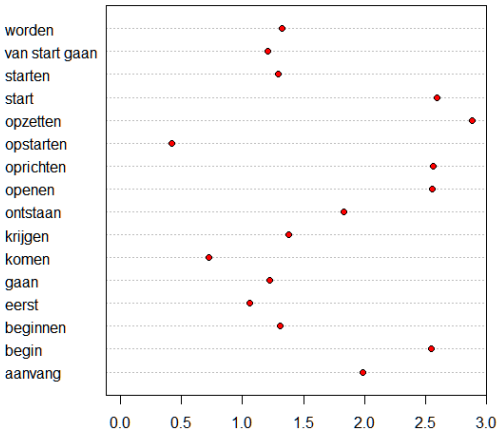


Figure 4.23: Cluster n°4 for TransDutch$_{ENG}$

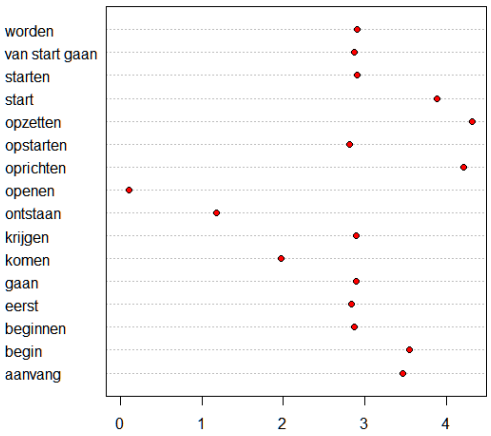


Figure 4.24: Cluster n°5 for TransDutch$_{ENG}$

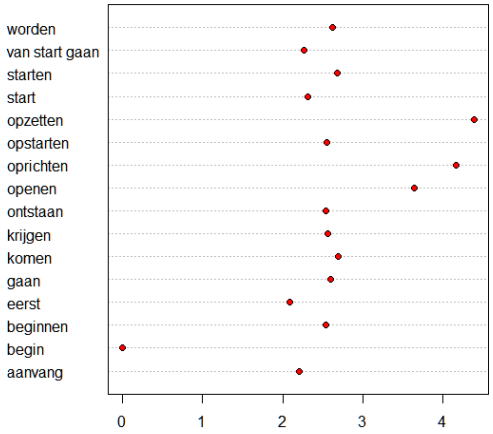


Figure 4.25: Cluster n°6 for $TransDutch_{ENG}$

and *worden* (0.12738579). For cluster n°4, *opstarten* is undoubtedly the closest lexeme to the centroids of the cluster. 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, *komen* is indeed 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 it is a singleton cluster with *begin*.

4.3.3.2 Medoids

Table 4.2 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).

The medoid and the closest lexeme to the centroid of cluster n°3 do not coincide. In addition, 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: for cluster n°3, *beginnen* is now closely followed by *gaan*, *krijgen* and wor-

Table 4.2: Comparison of medoids and lexemes closest to the centroids for $\text{TransDutch}_{\text{ENG}}$

	Medoid	Lexeme closest to centroids
Cluster n°3	worden	beginnen

den. Starten – for which a more central position in the cluster was expected – is situated slightly further away.

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

The following interpretation of a semantic field of *beginnen*/inchoativity for $\text{TransDutch}_{\text{ENG}}$ includes the assignment of a meta-label for each meaning distinction. The specific meaning distinctions determined for SourceDutch will be used as a point of reference to interpret the field of $\text{TransDutch}_{\text{ENG}}$. I will consequently attempt to assign the meta-labels that were chosen on the basis of the SourceDutch field to the field of $\text{TransDutch}_{\text{ENG}}$.

Cluster n°3 can be considered 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, 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, 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). In addition, the initial lexeme *beginnen* (considered as the most prototypical expression of inchoativity) can be found within the REFERENCE CLUSTER, strengthening my assumption that this cluster is holding the most prototypical expressions of inchoativity. The REFERENCE CLUSTER has furthermore 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 in-

choativity in TransDutch$_{ENG}$, compared to SourceDutch.

Just as for SourceDutch, I 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$_{ENG}$. I 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, 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, at the onomasiological level, *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. 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. The dot chart in §4.3.2 revealed 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) with respect to the distance of the lexemes *komen* and *opstarten* to the lexemes of the REFERENCE CLUSTER (Figure 4.23), it can be observed 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, the distance of the centroid of cluster n°2 and cluster n°6 to the zero-point of the semantic space shows 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 4.21) compared to SourceDutch (Figure 4.7).

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, 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$_{ENG}$ compared to SourceDutch. Figure 4.26 below now shows the semantic field of *beginnen*/inchoativity for TransDutch$_{ENG}$ with the meta-labels.

4.4 TransDutch$_{FR}$

The interpretation of the visualization of TransDutch$_{FR}$ follows 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 was decided to reduce the number of dimensions of the CA to 4.

A HAC was carried out and a cut-off point at a height of 5 was chosen, 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].

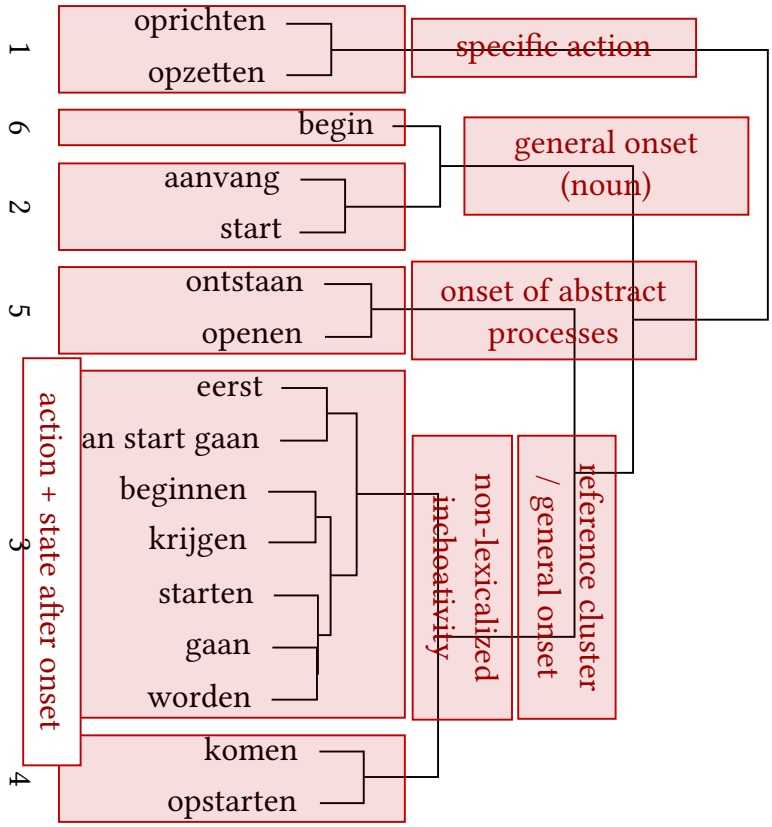


Figure 4.26: Dendrogram representing a semantic field of *beginnen*/inchoativity for TransDutch$_{ENG}$ with meta-labels

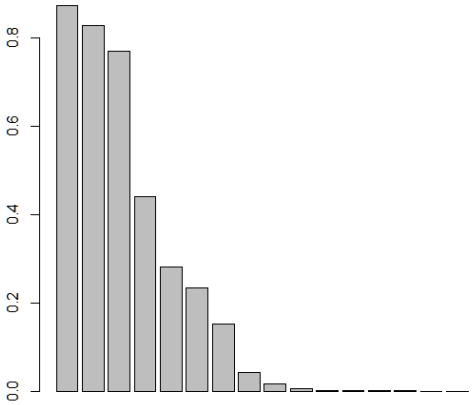


Figure 4.27: Scree plot for $\text{TransDutch}_{\text{FR}}$

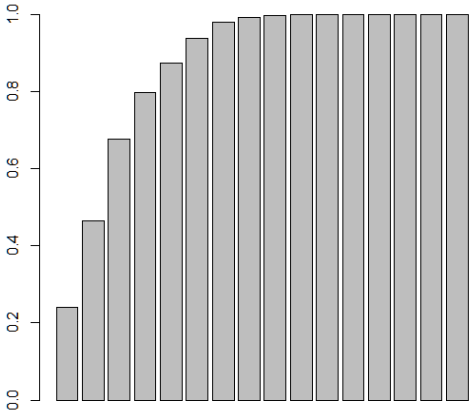


Figure 4.28: Cumulative scree plot for $\text{TransDutch}_{\text{FR}}$

The result presented in Figure 4.29 is considered as a possible visualization of a semantic field representing *beginnen*/inchoativity in TransDutch$_{FR}$.

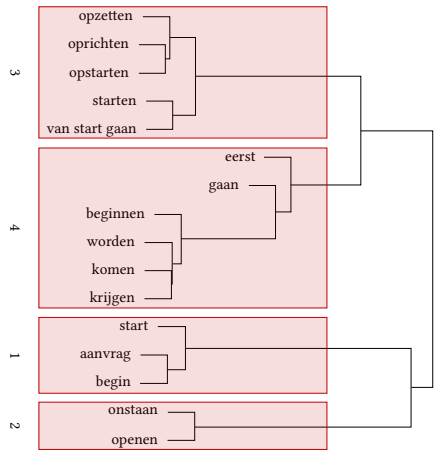


Figure 4.29: Dendrogram representing a semantic field of beginnen/inchoativity for TransDutch$_{FR}$

The cluster solution is validated by the average silhouette width for a solution with 4 clusters (average silhouette width = 0.53) (Figure 4.31) 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, I conclude that the chosen cluster solution for TransDutch$_{FR}$ can be considered a good classification.

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 4.32). The content of each cluster number in the dot chart is summarized in the table accompanying Figure 4.32:

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*

```
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
          3
```

Figure 4.30: K-means clustering with 4 clusters for TransDutch$_{FR}$

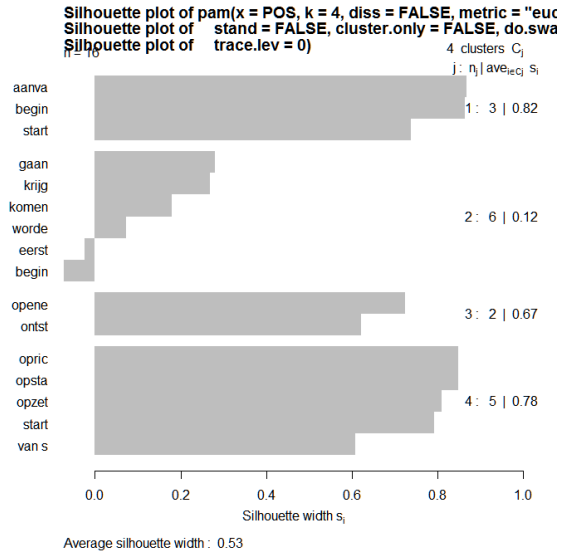
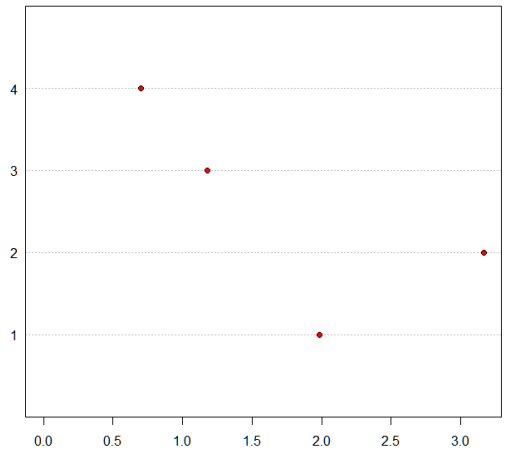


Figure 4.31: Average silhouette width for cluster solution with 4 clusters for TransDutch$_{FR}$



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

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: the distance of the lexemes within each cluster to the centroid of the cluster they belong to and the medoid of each cluster.

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. I again used the calculated

distances (which are represented by the dots in the dot charts) to evaluate the distances to the centroids (see appendix J).

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, I conclude that *starten* is the closest one to the centroid of the cluster. Finally, for cluster n°4, *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). The distance of *beginnen* to the centroid of the REFERENCE CLUSTER it belongs to is the smallest for $\text{TransDutch}_{\text{ENG}}$ (0.06521312) and the largest for $\text{TransDutch}_{\text{FR}}$ (0.6576414); for SourceDutch it is 0.1254173.

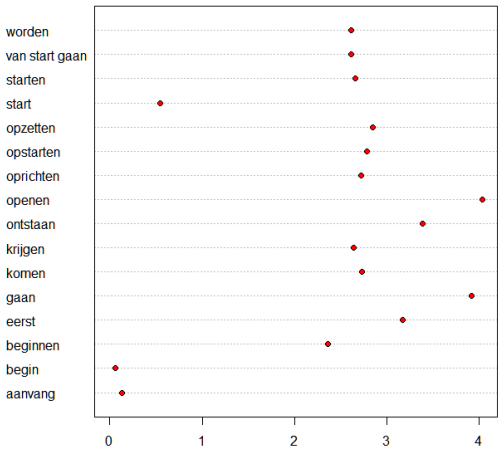


Figure 4.33: Cluster n°1 for $\text{TransDutch}_{\text{FR}}$

4.4.3.2 Medoids

In Table 4.3 below, the lexemes closest to the centroid of clusters n°1, 3 and 4 are compared to their respective medoid.

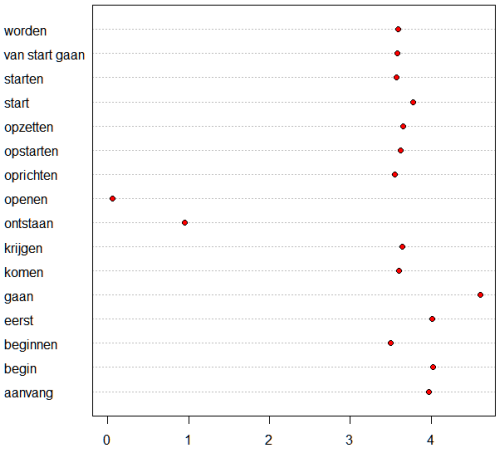


Figure 4.34: Cluster n°2 for TransDutch$_{FR}$

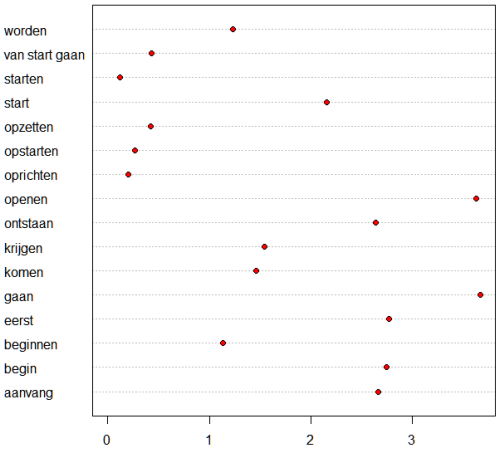


Figure 4.35: Cluster n°3 for TransDutch$_{FR}$

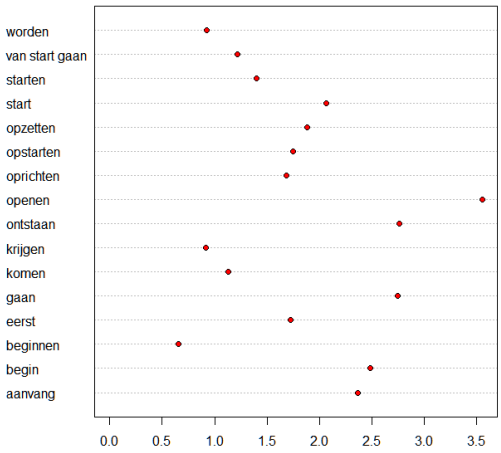


Figure 4.36: Cluster n°4 for TransDutch$_{F}$

Table 4.3: . 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, the medoids of clusters n°3 and n°4 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$_{FR}$

In the following 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. Just as for TransDutch$_{ENG}$, I will attempt to assign these meta-labels to the field of TransDutch$_{FR}$.

Cluster n°4 is considered as the most central cluster in the dendrogram, representing the idea of GENERAL ONSET. As I showed in §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. 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, two significant terminal nodes (*eerst* and *gaan*) can be discerned, 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 §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 two meaning distinctions can be discerned: SPECIFIC ACTION (*oprichten* and *opzetten*) as well all ACTION (*starten* and *van start gaan*). In TransDutch$_{FR}$, the distinc-

tion 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. In addition, *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:

- 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, my emphasis)
- 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, my emphasis)

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. The distance of the lexemes to the centroid of either cluster (Figures 81 and 82), shows a quite different image from 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 4.32), 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, the distance from the lexemes to their cluster's centroid shows 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 as small differences on the onomasiological level are observed: 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 4.37 shows the semantic field of *beginnen* for TransDutch$_{FR}$ with integration of the 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 *beginnen* is considered as the most prototypical expression of inchoativity, I believe 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

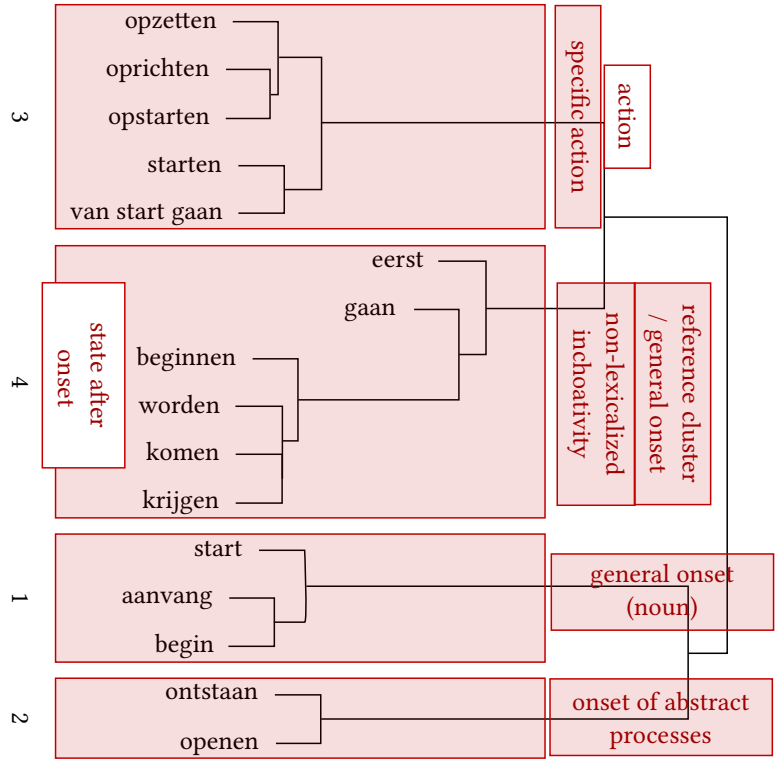


Figure 4.37: Dendrogram representing a semantic field of *beginnen*/inchoativity for $\text{TransDutch}_{\text{FR}}$ with meta-labels.

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, it cannot be maintained that clustering is totally independent of word class, since in the TransDutch fields, *eerst* becomes part of the REFERENCE CLUSTER and clusters with lexemes of a distinct word class.

4.5 Levelling out

In §4.2, §4.3 and §4.4 I formulated 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, I 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, I will call the phenomenon semasiological levelling out.

On the onomasiological level, I 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. I 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. My method does however not allow me 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 §3.5.3)[8]. Observations on the onomasiological level will inform me 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.

I first give a schematic overview of the 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.

4.5.1 Semasiological levelling out

On the semasiological level, I observe the following changes for the REFERENCE CLUSTER/GENERAL ONSET (Figure 4.38):

- 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.

For GENERAL ONSET (NOUN) (Figure 4.39), the following observations can be made:

- 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$_{ENG}$.

[8]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.

Table 4.4: SourceDutch

Cl.	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 INCHOACTIVTIY	*komen* *krijgen* *worden*		
4	ONSET OF ABSTRACT PROCESSES	*ontstaan* *openen*		
6		*eerst*		

Table 4.5: TransDutch$_{ENG}$

Cl.	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 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)

Table 4.6: $\text{TransDutch}_{\text{FR}}$

Cl.	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 > $\text{TransDutch}_{\text{ENG}}$); 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*

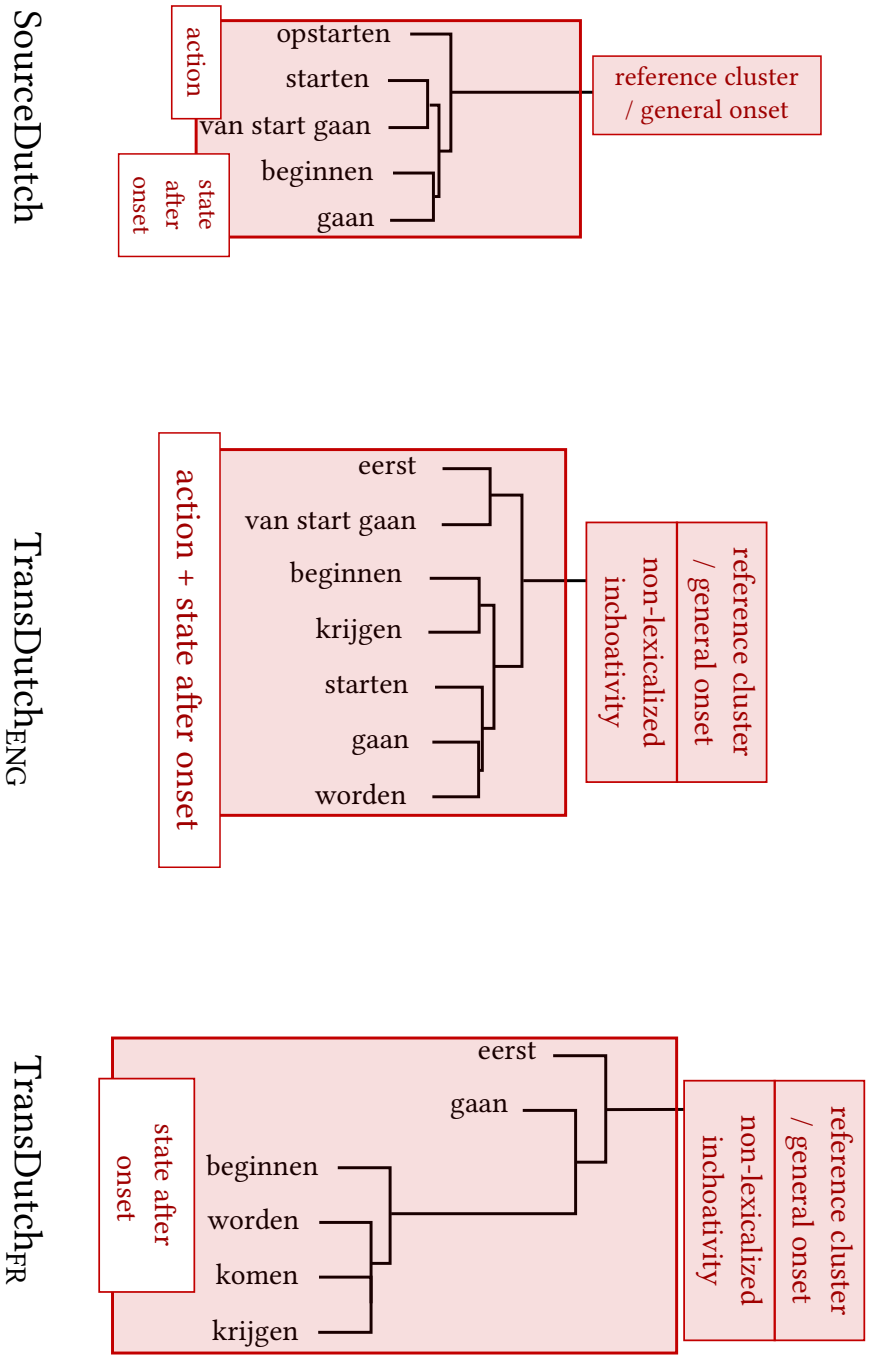


Figure 4.38: REFERENCE CLUSTER/GENERAL ONSET of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$

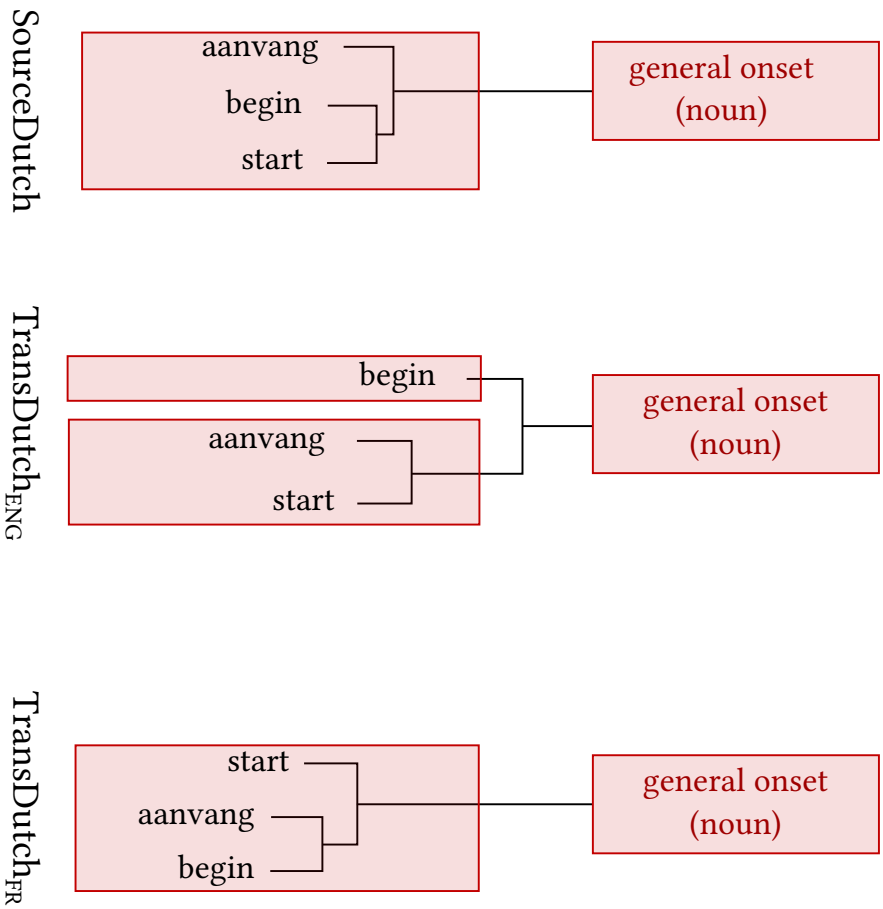


Figure 4.39: GENERAL ONSET (noun) of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$

For ACTION (Figure 4.40), my observations are as follows:

- 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.

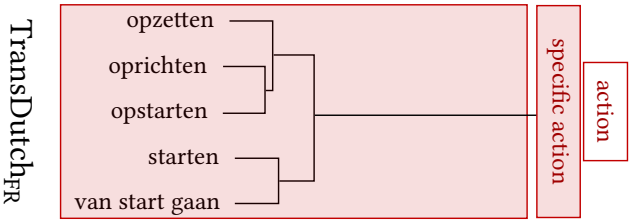


Figure 4.40: ACTION/SPECIFIC ACTION for TransDutch$_{FR}$

From a semasiological point of view, it can be concluded 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. Fewer of the meanings that were distinguished in SourceDutch are distinguished in the TransDutch fields. As a consequence, a presence of semantic levelling out on the semasiological level can be claimed here. Two observations seem however to go against this statement. First, for TransDutch$_{FR}$, 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), while 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, the following changes can be observed for the 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). For both TransDutch semantic representations, *beginnen* and *starten* are less near-synonymous than in SourceDutch.

In GENERAL ONSET (NOUN), *start* and *begin* compete for the position closest to the abstract prototype 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$_{FR}$, *begin* is closest to the abstract prototype, but *start* is situated much further away from the abstract prototype. *Begin* and *start* are thus less near-synonymous in both TransDutch fields compared to SourceDutch.

In 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, a number of small differences in the prototype-based organization of the lexemes are observed in TransDutch compared to SourceDutch. S*tarten* 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, it could be concluded 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 this trend has only been observed 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

Semasiological shining through (source language influence on the meaning distinctions in translated language) is investigated by comparing the meaning distinctions in translated language to those present in the source language of the translation. To do so, the semantic fields of the closest equivalents of *beginnen* in the source languages of $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$ are visualized: SourceEnglish *to begin* and SourceFrench *commencer*.

Ideally, I 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 comparing the different meaning distinctions (clusters) in the fields of *to begin* and *commencer* to the meaning distinctions in $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$. I 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 $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$ could be determined. Obviously, such a description would enhance the insights into the influence on the target language of attested differences between the source language and the target language semantic fields. I 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 $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$ 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*[9] 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$_{ENG}$, this could possibly be interpreted this as source language influence.

The semantic field of SourceEnglish was visualized on the basis of data retrieved via the SMM++ with *to begin* as initial lexeme and Dutch as a language B. The exact same procedure as for SourceDutch was followed. 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 was 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)[10]. The outcome of the SMM++ retrieval task rendered a set of 30 English lexemes (911 observations). I carried out a HAC on the output of the CA and choose a cluster solution with 5 clusters (average silhouette width 0.7).

The dendrogram (Figure 4.41) for SourceEnglish shows that *beginning* is part of

[9] Since the clustering of *komen* is unstable (§4.3.1), the analysis will mainly focus on *opstarten*.

[10] 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, I would have needed a tri-directional corpus (where all three languages can be used as languages B to carry out a SMM++) which I do not have at my disposal. Another solution would have been 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.

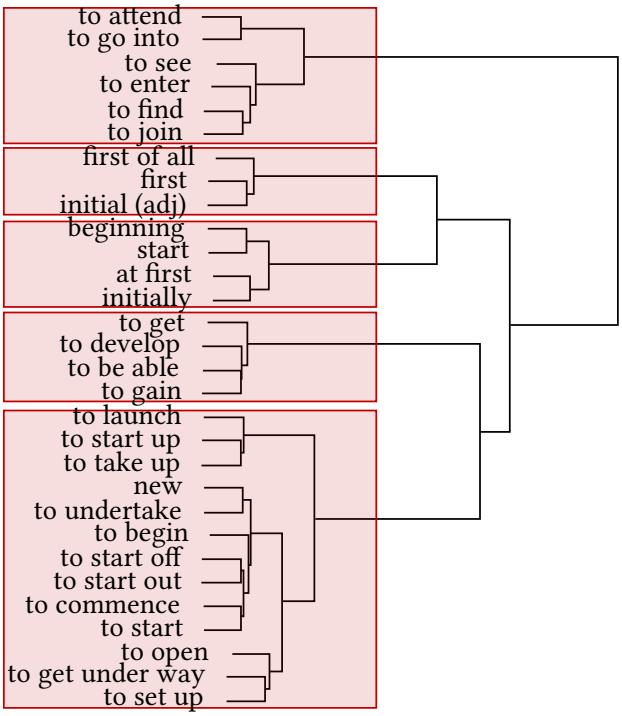


Figure 4.41: Dendrogram representing a semantic field of *to begin* for SourceEnglish

a cluster with *start, at first* and *initially* so that no separate meaning distinction of *beginning* is implied by SourceEnglish. Consequently, the separate meaning distinction of *begin* in TransDutch$_{ENG}$ could not have been triggered by an existing meaning distinction in SourceEnglish.

The verb *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[11]. On the basis of this information, I conclude that the separate meaning distinction of *komen* and *opstarten* in TransDutch$_{ENG}$ is not caused by an existing meaning distinction in SourceEnglish.

As for the unclear distinction between ACTION and STATE AFTER ONSET in TransDutch$_{ENG}$, no clear division between ACTION and STATE AFTER ONSET is marked in SourceEnglish either. The prototypical ACTION verb *to start* and the prototypical STATE AFTER ONSET verb *to begin* are both part of the same, most central 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 §4.7.1).

4.6.1.2 Semasiological shining through of SourceFrench

For TransDutch$_{FR,}$ 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, I will 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

[11] *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.

of SourceEnglish, the field of SourceFrench (Figure 4.42) 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). I carried out a HAC on the output of the CA. The chosen cluster solution with 4 clusters obtained an average silhouette width of 0.54.

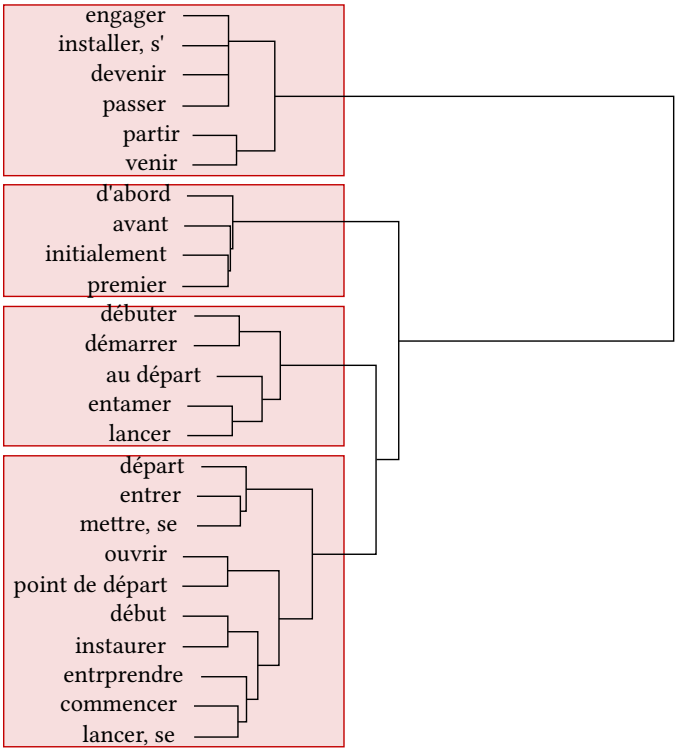


Figure 4.42: Dendrogram representing a semantic field of commencer for SourceFrench

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, there is inideed a cluster containing these ACTION verbs *entamer*, *débuter*, *démarrer*, *au départ* [initially] and *lancer* [to launch]. Within the cluster containing *commencer*, some of the lexemes 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*, there are also a number of lexemes present 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.

The above interpretation is of course preliminary, and can only hint 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

Two additional visualizations for TransDutch$_{ENG}$ and TransDutch$_{FR}$ are presented in this section, containing the English and French source language lexemes together with the Dutch target language lexemes. In this way, onomasiological shining through can be investigated. In other words, it can be determined 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, I 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 §4.6.1, I concluded that semasiological shining through could not account for the separate clustering of *begin*, nor for the separate clustering of *opstarten* in TransDutch$_{ENG}$. I will now explore whether this separate clustering could be the result of an instance of onomasiological shining through (the influence of a specific 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 §3.8). I 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 4.43[12] below is identical to that in Figure 4.16 (all clusters correspond to either a cluster or a sub-node)[13].

On a general level, it is striking 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). I discern the following pairs: *beginning-begin*; *start-*

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

[13] 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.

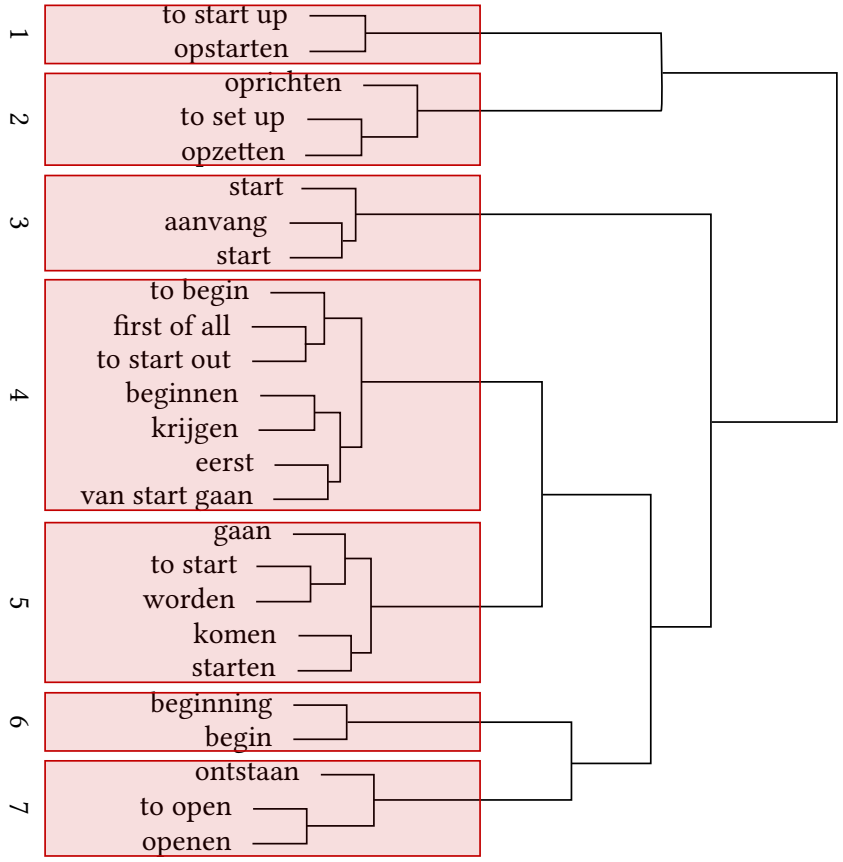


Figure 4.43: Representation of HAC on the MCA for $\text{TransDutch}_{\text{ENG}}$

start; *to open-openen*; *to begin-beginnen*; *to start-starten*; *to start up-opstarten*; *to set up-opzetten*.

Dutch *begin* is clustered with its English close cognate *beginning* (cluster n°6), revealing the preference of *begin* to be used 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 $\text{TransDutch}_{\text{ENG}}$ 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 §4.6.1.2, I tentatively accounted for the clear (over-emphasized with respect to SourceDutch) meaning distinction between ACTION and STATE AFTER ONSET in $\text{TransDutch}_{\text{FR}}$ via semasiological shining through. The joint clustering of ACTION and SPECIFIC ACTION could however not be explained on the semasiological level. In this section, I 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 4.44[14] shows the same semantic field of *beginnen* for $\text{TransDutch}_{\text{FR}}$ as the dendrogram of the HAC for $\text{TransDutch}_{\text{FR}}$ in Figure 4.29 (all clusters correspond to either a cluster or a sub-node).

The cluster reuniting SPECIFIC ACTION and ACTION in the HAC visualization in Figure 4.29 corresponds to clusters n°5 and 6 in Figure 4.44. 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 $\text{TransDutch}_{\text{FR}}$ 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 $\text{TransDutch}_{\text{FR}}$) argues against onomasiological shining through.

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

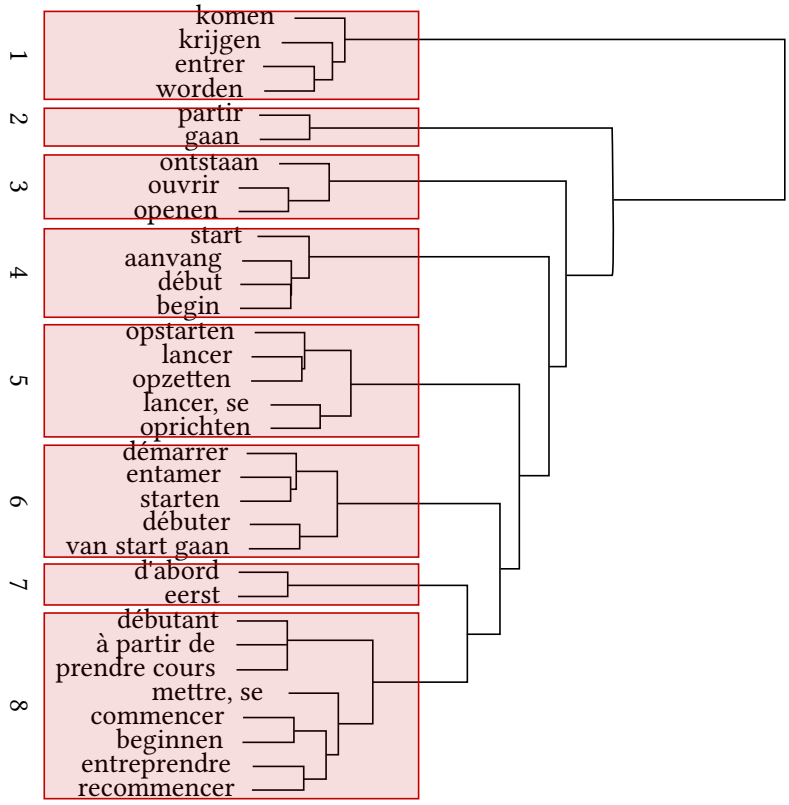


Figure 4.44: Representation of HAC on the MCA for TransDutch$_{FR}$

If the information gathered on the onomasiological level is now reconnected to the semasiological level, some additional insights can again be gained. 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 $\text{TransDutch}_{\text{FR}}$, SPECIFIC ACTION is no longer forming a separate cluster.

Although I cannot make any clear statements about how exactly the clustering of ACTION with SPECIFIC ACTION has come about in $\text{TransDutch}_{\text{FR}}$, it can, however, be stated that it is triggered by a change on the semasiological level, possibly by semasiological levelling out.

4.7 Normalization

4.7.1 Semasiological normalization

I will now focus on 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 $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$. If a same meaning distinction appears in $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$ 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 $\text{TransDutch}_{\text{ENG}}$. Although I showed that this could be interpreted as semasiological shining through (see §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 is not worrisome but rather 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) 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, the cluster SPECIFIC ACTION contains 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. In addition, the distance to the prototype in either variety shows 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$_{ENG}$) a case of onomasiological normalization couold be claimed here (the prototype-based organization of the lexemes within TransDutch$_{ENG}$ is conforming to SourceDutch).

In SourceDutch and TransDutch$_{FR}$, 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$_{FR}$, *begin* (0.05884857) is the closest lexeme to the abstract prototype, but *aanvang* (0.12955053) is now much closer to the abstract prototype than *start* (0.54160901). 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 hold an identical cluster with the lexemes *ontstaan* and *openen.* 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$_{ENG}$, the difference in distance to the abstract prototype slightly decreases (*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, *ontstaan* (0.0593305) is now the closest lexeme to the prototype, and *openen* (0.9492880) is situated further away from the abstract prototype. This argues against onomasiological normalization.

4.8 Conclusion

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

In sum, I found that the fields of translated and non-translated Dutch inchoativity differ from each other on the semasiological level. These semasiological differences are revealed by the differences in the meanings expressed by *beginnen* in translated vs. non-translated Dutch. I also observed differences on the onomasi-

ological level: the prototype-based organization of lexemes within the different meaning distinctions differed in translated Dutch compared to non-translated Dutch.

I 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, I 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.

I detected semasiological normalization for the cluster ONSET OF ABSTRACT PROCESSES. The specific clustering of ACTION and STATE AFTER ONSET in TransDutch$_{ENG}$ (in the REFERENCE CLUSTER) and in TransDutch$_{FR}$ (in separate clusters) was explained alternatively as a difference in degree 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, I will try to come to a better understanding of how such seemingly contradictory mechanisms can act upon a same semantic representation. I will do so by interpreting the results within more broad, cognitive-translational, explanatory frameworks from cognitive translation studies and bilingualism.

5 Cognitive explorations

5.1 Introduction

In the previous chapter, I 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 explained by applying the framework of *translation universals* – which I prefer to consider as *general tendencies* rather than *universals* – on the semantic level. Although the observations could indeed be fitted into the "universals 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, I will therefore look for cognitive explanations for the main observations described in Chapter 4: (i) the overall levelling out on the semasiological level in translated Dutch inchoativity; (ii) the instances of onomasiological shining through in $\text{TransDutch}_{\text{ENG}}$ (the separate clustering of *begin* and *opstarten*); (iii) the semasiological shining through or normalization causing the joint clustering (in $\text{TransDutch}_{\text{ENG}}$) or separate clustering (in $\text{TransDutch}_{\text{FR}}$) of ACTION and STATE AFTER ONSET and (iv) the joint clustering of ACTION and SPECIFIC ACTION in $\text{TransDutch}_{\text{FR}}$ under influence of semasiological shining through.

In this chapter, I will put forward two models that can possibly generate cognitive explanations for these findings. First, I will try to understand the results in the light of Halverson's (2003; 2010; 2013; 2017) Gravitational Pull Hypothesis (§5.2) (hence: GPH). In the subsequent §5.3, I will try to interpret my 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)[1]. 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

[1]This explanation was very briefly introduced in Vandevoorde et al. (2017)

them as explanatory frameworks for empirical studies in TS. Before I try to account for the results using either framework, I will, in the remainder of this section, zoom in on how cognitive explanations can be linked to corpus data (§5.1.1), and more specifically to semantic fields (§5.1.2). In §5.1.3, I will compare the starting points of the two models before I present and apply them to my results in §5.2 and §5.3.

5.1.1 Linking cognitive explanations to corpus data

Before I venture into this search for cognitive explanations, I first need to clarify how evidence from corpus data can be linked to cognitive explanations. In Chapter 2, I substantiated my choice to connect a corpus linguistic methodology with a cognitive linguistic theoretical framework. I equally discussed how the reintegration 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, my 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, it becomes possible (with the corpus-based methodological framework that was set up in Chapter 3) to reveal patterns which are not viewable through process data but which consist of many individual decisions i.e. the outcomes of 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 the current 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). It therefore needs to be clearly established what precisely the created semantic fields represent within a cognitive explanatory framework.

In this study, I am cautious not to consider the visualized semantic fields as representations "of how knowledge or patterns of usage are actually represented in the brain" (Divjak 2010a: 146)[2]. As House (2013: 51) suggested, measurements

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

of observable behavior (in this 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 the measurements – contained in the created semantic fields– can represent within a cognitive explanation (and why they do explain the cognitive processes occurring in the translator's mind), I 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 are stored (Paradis 2007: 199). 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 he or she 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, the semantic fields created in this study can be considered as semantic representations of a generalization (over many translators) of these cumulative results of the systems. This implies that these semantic fields are not thought 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 I cannot possibly answer within the scope of this study. However, by considering these semantic fields as semantic representations of the output of the joint working of the systems, I connect the cognitive explanations which I 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 I 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. (Cook 2003)

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 I will show, it is also suitable to explain translational effects on the semantic level.

Second, the GPH claimsva "multicompetence perspective (Cook 2003), which emphasizes that linguistic cognition in bilinguals is qualitatively different from that in monolinguals" (Halverson 2017: 12). 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" (Halverson 2017: 12), 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). The acceptation of

this null-hypothesis is a prerequisite to apply Paradis' framework to the type of results of this study since the only claim that can be made on the basis of those results is that the *contents* of the representations are accessed (and not the neurological mechanisms themselves).

5.2 Gravitational Pull Hypothesis

In §2.2.3.2, I 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, 2017), 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). The cognitive grammatical model on which the GPH is based is summarized as follows by Halverson (2017: 12):

> 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 the visualizations generated within this study are projected within this account, each of the created semantic fields can be considered 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:

> [F]irst, 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 2017: 12).

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 2017: 13) – 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 2017: 13). High frequency of use leads to entrenchment, which makes the linguistic forms (words/constructions) associated with them "more likely to be selected" (Halverson 2017: 13). Originally, gravitational pull (Halverson 2003) was to be understood as "semasiological salience in the target language" (Halverson 2017: 14). 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 2017: 14). 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 2017: 13, following Geeraerts 2009: 80). Salience effects can also exist on the onomasiological level, where they can be detected by "looking at the range of translations of a given ST item" (Halverson 2017: 28). Within the GPH, salience is operationalized as frequency of use (Halverson 2017: 13).

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 2017: 14). 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 2017: 14). The so-called connectivity, the strength (entrenchment) of a link between two translational equivalents is also thought to potentially influence translation (Halverson 2017: 15). The three above-mentioned phenomena, salience of source language patterns, salience of target language patterns and salient translational connections are thought to cause certain characteristics to become overrep-

resented or underrepresented in translated language compared to non-translated language (Halverson 2017).

In this study, *overlap* was presented in §3.4.3 (together with frequency) as an operationalization of salience in order to substantiate the prototype-based nature of the visualizations. As a consequence, the visualized semantic fields can be employed to assess the *salience* of the revealed patterns. Translational effects of salient source language patterns can be investigated by looking at the SourceField of the source language of a translation (see §4.6.1), translational effects of salient target language patterns by comparing the salient patterns in translated and non-translated target language (see §4.7) and translational effects of salient translational connections can be revealed on the basis of the joint visualization of source and target language lexemes (see §4.6.2).

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

Onomasiological shining through in $\text{TransDutch}_{\text{ENG}}$ (the separate clustering of *begin* and *opstarten*) can be explained as a consequence of connectivity. The visualization in §4.6.2.1 shows 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 in the visualization in §4.6.2.1) and a separate clustering for these lexemes (such as for *begin* and *opstarten*) is to be expected. The fact that there is no such separate clustering for *beginnen-to begin* and *starten-to start* cannot be explained by the GPH.

It is more difficult to interpret semasiological shining through in $\text{TransDutch}_{\text{FR}}$ (the joint clustering of ACTION and SPECIFIC ACTION) as an instance of gravitational pull. Indeed, the joint clustering of ACTION and SPECIFIC ACTION in $\text{TransDutch}_{\text{FR}}$ cannot be explained 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.

I concluded that the joint clustering (in $\text{TransDutch}_{\text{ENG}}$) or separate clustering (in $\text{TransDutch}_{\text{FR}}$) 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 to 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 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 §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 §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, I 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 the visualizations (and, vice versa, the visualizations seem indeed to be suitable instruments to further test the GPH), my post-hoc approach has of course its limitations. The obvious disadvantage is that some of the findings which I 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 my view, also compatible with observable data of "translational behavior". Within cognitive TS, Paradis' theory (2004) has been proposed 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). In §5.3.1, I will outline the main ideas behind Paradis' theory. In §5.3.2, I will apply the model to translation in general, before I use it as an explanatory framework for the results obtained in this study (§5.3.3)[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 Re-

[3]The possibility to apply Paradis' framework to my results was briefly introduced in Vandevoorde et al. (n.d.), an article which is under copyright. Its publisher should be contacted for permission to re-use or reprint the material in any form.

vised Hierarchical Model by Kroll & 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" had yielded inconsistent experimental results (Paradis 2007: 6). 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 5.1 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 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).

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

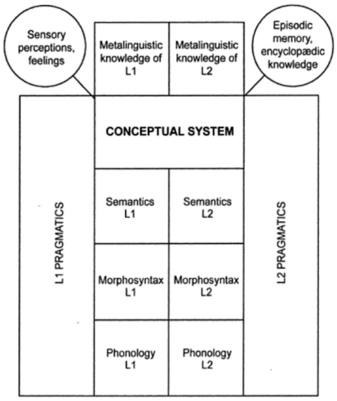


Figure 5.1: Schematic representation of the components of verbal communication (copied from Paradis 2004: 227)

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 lan-

guage 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 5.2 to be a (simplified) representation of the conceptual features activated by the noun *haar*.

filiform	covers body	covers head	in humans	in animals
⊗	⊗	⊗	⊗	⊗

Figure 5.2: 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 5.3) 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 sentence):

filiform	covers body	covers head	in humans	in animals
⊗	○	⊗	⊗	○

Figure 5.3: 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 5.4) will be activated (the activation of 'covers body' and 'in animals' de-activates 'in humans' in this context):

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

filiform	covers body	covers head	in humans	in animals
⊗	⊗	⊗	○	⊗

Figure 5.4: Activated conceptual features for *de hond heeft lang haar*

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:

filiform	covers body	covers head	in humans	in animals
⊗	⊗	○	⊗	⊗

Figure 5.5: Activated conceptual features for *poil*

filiform	covers body	covers head	in humans	in animals
⊗	○	⊗	⊗	○

Figure 5.6: 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: 119-20) (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, 2016: 119-120) (the case of *actrice*).

The importance of form similarity as put forward here is further substantiated by Brysbaert et al.Brysbaert et al. (2014: 140). Although in general bilingual speaker's contexts "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 (cf. also Halverson's "entrenchment of translation pairs" (2017, 15)). Although this direct translation is not a cognate, the quasi-total overlap of conceptual features and/or the association strength (Halverson's 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

Paradis' framework is now applied to the observations about overall semasiological levelling out in translated language; onomasiological shining through in $\text{TransDutch}_{\text{ENG}}$, semasiological shining through or normalization for ACTION and STATE AFTER ONSET in translated language and semasiological shining through in $\text{TransDutch}_{\text{FR}}$ for ACTION and SPECIFIC ACTION. I consider our

visualizations to be 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. Based on the above section, I concluded 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 §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 $\text{TransDutch}_{\text{ENG.}}$. 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 indicate that direct transcoding is taking place in $\text{TransDutch}_{\text{ENG}}$. However, the frequency matrix in appendix C 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 $\text{TransDutch}_{\text{FR}}$ since there are simply fewer 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 I 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 my interpretation is speculative and constitutes only one of the many possible ways to interpret these changes in semantic structure). I 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, my 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, I have made an attempt to explain the main observations of this study on the basis of two cognitively inspired frameworks. I first explored how the GPH could account for the results of this study. I found that the idea of connectivity can explain the observed onomasiological shining through in TransDutch$_{ENG}$ as well as semasiological levelling out. Given my post-hoc approach, it appeared however difficult to connect my 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 I 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 be 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 I already mentioned, a post-hoc application of these frameworks has its obvious limitations. Nevertheless, I hope to have demonstrated the explanatory power of these frameworks, especially when they are combined with methodological instruments such as the visualizations proposed in this 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.

6 Conclusion

6.1 General conclusions

Impelled by the lack of empirical studies involved with meaning variation in translation, I decided to place the study of semantic differences in translated compared to non-translated texts at the center of my concerns. To date, much research in CBTS has focused on lexical and grammatical phenomena in an attempt to reveal presumed general tendencies of translation, but on the semantic level, these general tendencies have rarely been investigated. I therefore set out to answer three central questions.

The first question, how to investigate semantic differences in a translational setting, required a lengthy answer which was covered by the methodology proposed in Chapter 3. Given the attested lack of empirical studies of semantic phenomena in CBTS, no clear hypotheses could be drawn beforehand so that the proposed method necessarily had to be explorative in nature. In addition, CBTS offers very few methodological guidelines for semantic investigations. As a consequence, the first challenge was to develop a methodological technique able to measure semantic similarity of translated and non-translated language. I established a way to visually explore semantic similarity on the basis of representations of translated and non-translated semantic fields of a concept under study. More specifically, I 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. In addition, different types of visualizations were proposed so as to investigate levelling out, shining through and normalization.

The application of the developed method to the case of inchoativity in Dutch allowed me to answer the second question: are there any differences on the semantic level between translated and non-translated texts? Since I did indeed observe differences on the semantic level between translated and non-translated Dutch, the third question required to be answered as well. Based on the additional

visualization techniques proposed in Chapter 3, I made an attempt to link the observed differences to the universal tendencies of levelling out, shining through and normalization, which I considered to be the most suitable ones for semantic research.

I found evidence for the presence of 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. As for semasiological shining through, I found that an influence of the source language possibly provoked the joint clustering of ACTION and STATE AFTER ONSET in $\text{TransDutch}_{\text{ENG}}$, the separate clustering of ACTION and STATE AFTER ONSET in $\text{TransDutch}_{\text{FR}}$ and the joint clustering of ACTION and SPECIFIC ACTION in $\text{TransDutch}_{\text{FR}}$. However, the specific clustering of ACTION and STATE AFTER ONSET in $\text{TransDutch}_{\text{ENG}}$ (into the REFERENCE CLUSTER) and in $\text{TransDutch}_{\text{FR}}$ (into separate clusters) could also be explained as different degrees of target language influence, and hence, as semasiological normalization.

On the onomasiological level, I observed that the prototype-based organization of lexemes within the separate meaning distinctions differed in translated language, compared to non-translated language. Unfortunately, I could not connect my conclusions directly to the idea of onomasiological levelling out, since the number of lexemes in each visualization is kept stable. I 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. For onomasiological shining through, I found that the distinct clustering of *opstarten* and *begin* (as such semasiological phenomena) in $\text{TransDutch}_{\text{ENG}}$ could be explained as an influence of the source language, i.e. shining through on the onomasiological level. Furthermore, the prototype-based organization of *oprichten* and *opzetten* in $\text{TransDutch}_{\text{ENG}}$ showed signs of onomasiological normalization because of the similarity with the prototype-based organization of these lexemes in SourceDutch.

Unsatisfied with the limited explanatory power of the universals paradigm, I tried to explain the main results of this study on the basis of two cognitively inspired frameworks. The proposed cognitive frameworks – the Gravitational Pull Hypothesis and Paradis' neurolinguistic theory of bilingualism – were applied to the 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), I accounted for the separate clustering of *begin* and *opstarten* in $\text{TransDutch}_{\text{ENG}}$

(onomasiological shining through). In addition, by following the reasoning behind translating via the conceptual system (Paradis), I could tentatively explain how the observed instances of semasiological levelling out, semasiological shining through or normalization had come about.

6.2 Retrospective insights

The conclusions about tendencies of levelling out, shining through and normalization are arguably 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. I indeed concluded 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 the 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 me to an important point about the impact of methodological choices on my results. My 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.

Firstly, my decision to select the same lexemes for each visualization was taken to ensure the comparability of the visualizations but had the effect that onomasiological levelling out could not be investigated as such. Secondly, the observation of a frequency threshold of three observations impacted the number of selected lexemes. A frequency threshold of two observations would have resulted in the

iddition of the following lexemes: *aangaan* [to start], *aanvatten* [to commence], *begin*-[initial], *doen* [to do], *lanceren* [to launch], *maken* [to make], *nemen* [to take], *sinds* [since], *start*- [starting]. Thirdly, my choice to base the developed 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. Finally, 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 more fine-tuned interpretations of the semantic fields can be given. A number of alternative methodological possibilities will need to be tested before a deeper level of analysis of the semantic fields can be pursued. 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 the 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.

Due to the lack of previous work on the subject, I was left in the dark about what semantic levelling out, shining through or normalization would look like. This explains the explorative character of this study and my primary concern with the operationalizability of these universal tendencies on the semantic level. In this regard, I did not choose the case of *beginnen*/inchoativity 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 the main observations of this study 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. I 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, the visualizations did not allow me 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 me to differentiate between shining through and normalization, Paradis' idea of translation via the conceptual system offered a possible explanation for the observed phenomena in the translated semantic fields, without creating the need 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 work, I hope to have opened the way for more semantic research in TS. A number of methodological developments presented in this book might constitute a first small step towards more research into semantic differences in translation and more cognitive explanations for translational behavior. I believe to have shown 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 I now applied tentatively and post-hoc.

7 Appendices

Appendix A: First T-image beginnen$_{ENG}$

SLeng * TLdu Crosstabulation

Count

		TLdu	Total
		beginnen	
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

7 Appendices

Appendix B: 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 C: 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 D: 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 E: Second T-image beginnen$_{ENG}$

SLdu * TLeng Crosstabulation (Count)

		TLeng									Total
		beginning	first of all	start	to begin	to open	to set up	to start	to start out	to start up	
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 F: Second T-image beginnen$_{FR}$

SLdu * TLfr Crosstabulation (Count)

		TLfr																	Total
		à partir de	commencer	d'abord	début	débutant	débuter	démarrer	entamer	entreprendre	entrer	lancer	lancer, se	mettre, se	ouvrir	partir	prendre cours	recommencer	
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 G: 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 H: Distances of lexemes to centroids for SourceDutch

Lexeme	1	2	3	4	5	6
aanvang	3.46413967	0.55330205	1.6594735	3.2985467	3.4994764	3.418562e+00
begin	3.69363445	0.08908994	2.2927647	3.5270707	3.8466311	3.636506e+00
beginnen	3.38253885	2.19010475	0.1254173	3.1994866	3.4054110	3.240007e+00
eerst	4.39454617	3.60359130	3.2708235	4.2819591	4.5189627	1.110223e-16
gaan	3.45622880	2.32441153	0.2722819	3.2991556	3.1970538	3.386583e+00
komen	4.48534925	3.89901953	3.4488256	4.5050119	0.1317251	4.602946e+00
krijgen	3.78598542	2.86275180	2.1115304	3.0419090	1.4928849	3.763519e+00
ontstaan	3.46474865	2.21507987	2.0267077	1.3471582	3.7050719	3.566664e+00
openen	4.32177245	3.65865122	3.3377771	0.1718314	4.4902810	4.394572e+00
oprichten	0.02952887	3.63757265	3.2631613	4.1998850	4.3949576	4.379157e+00
opstarten	2.75535998	2.25815302	0.5508145	3.1747593	3.2705318	3.356807e+00
opzetten	0.06749455	3.70027235	3.3478383	4.2210725	4.4466206	4.430260e+00
start	3.58117345	0.20740218	2.0028870	3.4319375	3.7221581	3.549602e+00
starten van	3.19549285	2.23275009	0.1264550	3.1248429	3.3679560	3.293686e+00
start gaan	3.06686816	2.12361039	0.2694401	3.0678892	3.2158486	3.300870e+00
worden	3.55533903	2.48650607	1.1834631	3.4027884	2.1810695	3.467892e+00

Appendix I: Distances of lexemes to centroids for $\text{TransDutch}_{\text{ENG}}$

Lexeme	1	2	3	4	5	6
aanvang	4.1772428	0.6167167	2.25776243	1.9830519	3.4640847	2.201321e+00
begin	4.2660272	2.2744824	2.53884784	2.5465248	3.5484177	2.220446e-16
beginnen	3.6606035	2.8871317	0.06521312	1.3054293	2.8753627	2.535833e+00
eerst	3.4678803	2.4709349	0.51308195	1.0569324	2.8300225	2.085160e+00
gaan	3.6048433	2.8256940	0.11345029	1.2230535	2.8883564	2.600465e+00
komen	3.1391087	2.8694553	1.10378816	0.7203757	1.9726145	2.690092e+00
krijgen	3.7250253	2.9265041	0.11370695	1.3757511	2.8978139	2.561533e+00
ontstaan	3.9346210	3.2177434	1.95464441	1.8270922	1.1745826	2.542763e+00
openen	4.3005328	3.8700998	2.96852224	2.5526285	0.1067802	3.644855e+00
oprichten	0.2004520	4.3190150	3.51758612	2.5663876	4.2131591	4.170110e+00
opstarten	2.5129762	2.3375610	1.53758445	0.4202192	2.8121934	2.554513e+00
opzetten	0.2476172	4.6485414	3.84721417	2.8859388	4.3223569	4.394258e+00
start	4.5318886	0.1284827	2.96275365	2.5973465	3.8828659	2.310246e+00
starten van	3.7323929	2.7788342	0.25003812	1.2945116	2.9107599	2.678813e+00
start gaan	3.6994104	2.7788342	0.37259612	1.2033103	2.8650225	2.266192e+00
worden	3.7258800	2.8383236	0.12738579	1.3202771	2.9009154	2.624651e+00

Appendix J: Distances of lexemes to centroids for TransDutch$_{FR}$

Lexeme	1	2	3	4
aanvang	0.12955053	3.9675981	2.6680740	2.3641785
begin	0.05884857	4.0214808	2.7446571	2.4846664
beginnen	2.36301934	3.4961958	1.1318943	0.6576414
eerst	3.16694244	4.0075011	2.7706002	1.7263104
gaan	3.91766584	4.6097797	3.6749542	2.7439995
komen	2.73282963	3.5973206	1.4586546	1.1315485
krijgen	2.63904093	3.6436850	1.5443559	0.9121243
ontstaan	3.38754995	0.9492880	2.6444725	2.7598750
openen	4.03403545	0.0593305	3.6344425	3.5554558
oprichten	2.72209090	3.5496501	0.2037736	1.6826330
opstarten	2.78419901	3.6195418	0.2664611	1.7438098
opzetten	2.84781317	3.6528091	0.4267752	1.8862396
start	0.54160901	3.7768222	2.1561010	2.0667548
starten van	2.65295171	3.5682872	0.1160007	1.4000293
start gaan	2.60959733	3.5784474	0.4316780	1.2147151
worden	2.61519730	3.5894073	1.2349533	0.9202239

Semantic differences in translation

Change your blurb in localmetadata.tex

ISBN 978-3-96110-072-9