Meta
Journal des traducteurs
Translators' Journal

META

Register, Source Language, and Cognateness Effects on Lexical Choice in Translated Dutch

Lore Vandevoorde

Volume 63, Number 3, December 2018

Traductologie de corpus : 20 ans après

URI: https://id.erudit.org/iderudit/1060166ar
DOI: https://doi.org/10.7202/1060166ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Vandevoorde, L. (2018). Register, Source Language, and Cognateness Effects on Lexical Choice in Translated Dutch. *Meta*, *63*(3), 627–648. https://doi.org/10.7202/1060166ar

Article abstract

In line with recent studies about register and source language effects on translated language (Delaere, De Sutter, *et al.* 2012; De Sutter, Delaere, *et al.* 2012; Kruger and Van Rooy 2012; Delaere and De Sutter 2017), the aim of this paper is twofold: to further investigate the influence of register and source language, and to study any potential influence of the variable "cognateness" on translated language. We focus on specific onomasiological choices (lexical choices) in the semantic field of inchoativity, made by translators into Dutch and attested in corpus observations (Dutch translated texts in the Dutch Parallel Corpus). First, we performed a multinomial regression analysis on our dataset and carried out an Analysis of Deviance to determine whether the predictor variables "source language lexeme" and "register" ("text type") have a significant influence on the response variable (the set of lexemes representing the onomasiological choice range in translated Dutch inchoativity). Doing a second multinomial regression analysis, followed by an Analysis of Deviance, we investigate the influence of the new variable "cognateness" on the translator's onomasiological choice (in the target language). Classification trees were generated as statistics-based visualizations of onomasiological choice in translated Dutch (translated from French and translated from English) within the semantic field of inchoativity. The results of the statistical analyses show that register, source language, and cognateness significantly influence the specific lexical choices made by translators. In addition, the visualizations show how the onomasiological choice for some target lexemes can be predicted on the basis of a single source language lexeme, while other choices are more complex, and will also be determined by the register of the text.

Register, Source Language, and Cognateness Effects on Lexical Choice in Translated Dutch

LORE VANDEVOORDE
Ghent University, Ghent, Belgium
Lore.vandevoorde@ugent.be

RÉSUMÉ

À l'instar de plusieurs études récentes qui se sont penchées sur les effets de registre et de la langue source sur la langue traduite (Delaere, De Sutter, *et al.* 2012 ; De Sutter, Delaere, *et al.* 2012 ; Kruger et Van Rooy 2012 ; Delaere et De Sutter 2017), l'objectif du présent article est d'étudier davantage les influences possibles du registre et de la langue source ainsi que l'influence potentielle de la variable « degré d'apparentement » sur la langue traduite. Nous nous concentrons sur des choix onomasiologiques spécifiques (choix lexicaux) dans le champ sémantique de l'inchoativité, choix effectués par des traducteurs traduisant vers le néerlandais et attestés par des observations sur un corpus (textes traduits vers le néerlandais extraits du Dutch Parallel Corpus). Tout d'abord, nous effectuons une analyse de régression multinomiale sur l'ensemble des données. Ensuite une analyse de déviance est réalisée afin de déterminer si les variables prédictives « lexème » et « registre de la langue source » (type de texte) ont une influence significative sur la variable réponse (l'ensemble des lexèmes qui représentent les choix onomasiologiques éventuels pour exprimer l'inchoativité en néerlandais traduit). Avec une deuxième analyse de régression multinomiale suivie d'une analyse de déviance, nous étudions l'influence de la nouvelle variable « degré d'apparentement » sur le choix onomasiologique du traducteur (dans la langue cible). Des arbres de classification sont ensuite générés en tant que visualisations statistiques du choix onomasiologique en néerlandais traduit (traduit du français et traduit de l'anglais) dans le domaine sémantique de l'inchoativité. Les résultats des analyses statistiques montrent que le registre, la langue source et le degré d'apparentement influencent de manière significative les choix lexicaux spécifiques faits par les traducteurs. De plus, les visualisations montrent comment le choix onomasiologique pour certains lexèmes cibles peut être prédit sur la base du simple lexème de la langue source, alors que d'autres choix sont plus complexes et sont également déterminés par le registre du texte.

ABSTRACT

In line with recent studies about register and source language effects on translated language (Delaere, De Sutter, *et al.* 2012; De Sutter, Delaere, *et al.* 2012; Kruger and Van Rooy 2012; Delaere and De Sutter 2017), the aim of this paper is twofold: to further investigate the influence of register and source language, and to study any potential influence of the variable "cognateness" on translated language. We focus on specific onomasiological choices (lexical choices) in the semantic field of inchoativity, made by translators into Dutch and attested in corpus observations (Dutch translated texts in the Dutch Parallel Corpus). First, we performed a multinomial regression analysis on our dataset and carried out an Analysis of Deviance to determine whether the predictor variables "source language lexeme" and "register" ("text type") have a significant influence on the response variable (the set of lexemes representing the onomasiological choice range in translated Dutch inchoativity). Doing a second multinomial regression analysis, followed by an Analysis of Deviance, we investigate the influence of the new variable "cognateness" on the translator's onomasiological choice (in the target language). Classification trees were generated as statistics-based visualizations of onomasiological choice in translated Dutch (translated

from French and translated from English) within the semantic field of inchoativity. The results of the statistical analyses show that register, source language, and cognateness significantly influence the specific lexical choices made by translators. In addition, the visualizations show how the onomasiological choice for some target lexemes can be predicted on the basis of a single source language lexeme, while other choices are more complex, and will also be determined by the register of the text.

RESUMEN

Al igual que varios estudios recientes sobre el registro y los efectos del idioma de partida en el idioma traducido (Delaere et al. 2012; De Sutter et al. 2012; Kruger y Van Rooy 2012; Delaere y De Sutter 2017), el objetivo de este artículo es investigar la influencia del registro y del idioma de partida en el idioma traducido así como la influencia potencial de la variable «grado de parentesco» en este mismo idioma traducido. Nos focalizamos en las opciones onomasiológicas específicas (opciones léxicas) en el campo semántico de la incoatividad, realizadas por traductores traduciendo hacia el neerlandés y confirmadas en observaciones de corpus (textos traducidos hacia el neerlandés extraídos del Dutch Parallel Corpus). Primero, realizamos un análisis de regresión multinomial seguido por un análisis de desviación para determinar si las variables predictoras «lexema» y «registro del lenguaje de partida» (tipo de texto) tienen una influencia significativa en la variable de respuesta (el conjunto de lexemas representando la oferta de opciones onomásiológicas en el campo de la incoatividad, traducida hacia el neerlandés). Con un segundo análisis de regresión multinomial seguido de un análisis de desviación, investigamos la influencia de la nueva variable «grado de parentesco» en la elección onomasiológica del traductor (en el idioma de destino). Generamos árboles de clasificación con el fin de ofrecer unas visualizaciones estadísticas de elección onomasiológica en neerlandés traducido (traducido del francés y traducido del inglés) dentro del campo semántico de la incoatividad. Los resultados de los análisis estadísticos muestran que el registro, el idioma de partido y el grado de parentesco influyen significativamente en las elecciones léxicas específicas de los traductores. Además, las visualizaciones muestran cómo se puede predecir algunas de las eleccióones onomasiológicas a partir del lexema en la lengua de partida, mientras que otras opciones son más complejas, y también estarán determinadas por el registro del texto.

MOTS CLÉS / KEYWORDS / PALABRAS CLAVE

étude sur corpus, choix onomasiologique, régression multinomiale, inchoativité, arbre de classification
corpus-based, onomasiological choice, multinomial regression, inchoativity, classification tree
estudio basado en datos de corpus, elección onomasológica, regresión multinomial, incoatividad, árbol de clasificación

1. Introduction

One of the major research paradigms in corpus-based translation studies (CBTS) revolves around the idea that there exists a number of *universally viable* characteristics of translated texts. However, a particular characteristic of translated language can only be categorized as *universal* if it is invariable across *all* translation-relevant parameters such as source and target language, specific language pairs, registers, etc. (Chesterman 2004). Recent studies have shown that specific language pairs (Xiao 2010) or registers (Delaere, De Sutter, *et al.* 2012) do have an impact on the presence, directionality and magnitude of differences accounted for in translated texts, compared to non-translated

texts, thus unsettling the universality of translation universals. The body of research refuting the universality of translation universals has become so substantial that it is fair to say that the status of the term *universal* has more or less reached the same status in translation studies (TS) as it has in general linguistics:

> [In linguistics] it has become generally accepted […] 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).

Abandoning the absoluteness of translation universals does not however mean that this avenue of research has come to an end. Much to the contrary, accepting that linguistic differences between non-translated and translated language can be described as general tendencies broadens the scope of research and frees the translation scholar from the (nearly) unattainable goal of distinguishing universally applicable features of translated language: tendencies can exist within certain language pairs, within certain text types or can be triggered by certain translation policies, etc. The possible subtlety of these tendencies makes the endeavor all the more complex, and therefore (often) requires specific research methods, such as advanced statistical techniques, to deal with this complexity.

Two cases in point which add to the intricacy of what makes translated language different from non-translated language, and what we will deal with in this study, are register and source language. More specifically, we want to investigate to what extent register and source language are determinant for the onomasiological choices made by translators.[1] Throughout this paper, we will consistently use the terms onomasiological choice (rather than lexical choice) and semasiological choice which are commonly used in lexical semantic studies (see Section 2.2 for a more in-depth explanation of the two terms).

Studies on the influence of register on onomasiological choice in translation have mostly focused on register variation for binary choices (Delaere and De Sutter 2017). In this study, we make an attempt to model onomasiological choice starting from the idea that a source language word which is (proto)typical for the semantic field to which it belongs (for instance, in this case, French **commencer** and English *to begin* are prototypical verbs in the semantic field of inchoativity) will lead to various, though semantically related, translational outcomes in the target language (in this case, Dutch). In addition, we want to include the variable "cognateness" into the rationale. Evidence from psycholinguistics has shown that cognates are in general produced faster and in a more accurate way than control words that only exist in one of the languages a bilingual person masters. From the point of view of translation studies, the use of cognate translations in texts produced by professional translators is inevitably linked to a (language-dependent) risk of producing false friends, and, on the theoretical level, to the broader study of literal translation (Halverson 2015). The negative associations that both false friends and literal translation typically evoke within translation studies might explain why so little is known about a possible impact of 'cognate preference' on translated texts. By investigating the onomasiological choices within a same semantic field (inchoativity), in texts from different registers, from different source languages (where one language pair (English-Dutch) holds many cognates, and the other one (French-Dutch) does not), we hope to gain a better understanding of the potential influences translated texts undergo.

In the remainder of this introduction, we will give an overview of the main CBTS findings with regards to register and source language influence on translated texts. Furthermore, we will show why cognateness is possibly an interesting factor for modeling translated language. In section 2, we introduce a methodological framework that will allow us to model onomasiological choice within the semantic field of inchoativity. The method consists of (i) a lexeme-selection technique for translational data, (ii) statistical analyses using a multinomial regression analysis and an analysis of Deviance for model selection, and, finally, (iii) a statistical visualization method which allows us to visualize the complex interdependencies between the various variables, which ultimately lead to the translators' onomasiological choices. In section 3, we will thoroughly describe the results of the visual representation of the analysis. Based on previous studies about source-language and register-effects on translated language, our hypothesis is that register will play a determinant role in lexical choice. We will conclude this paper with a discussion on the possible implications of including cognateness as a variable influencing translated texts.

1.1. Influence of register on translated language

The investigation of the influence of register on translated language is a recent phenomenon (for a concise overview of such studies from before 2010, see Neumann (2014)). However, as Kruger and Van Rooy rightfully remark, studies that systematically investigate the relationship between register and translation universals are scarce (Kruger and Rooy 2012: 36). Two studies from 2012 (Kruger and Van Rooy 2012; Delaere, De Sutter, *et al.* 2012) can be considered the first thorough, quantitative investigations of register-effects on translated language. Kruger and Van Rooy investigated occurrence patterns for a number of features which are typically used to investigate the presence or absence of translation universals (type-token ratio, word length, that omission, etc.). They also examined the relationship between these features and register. They hypothesized that they would find significant differences between translated and non-translated English texts for the features linked to translation universals. In addition, less register variation was expected in a translation corpus compared to a comparable corpus of non-translated English texts. While the scholars found limited evidence for the first hypothesis, no support was found for the second hypothesis that translated texts are less register-sensitive than texts originally written in English (although there were clear differences between the registers for most of the features they investigated). They did, however, observe a number of subtle effects, such as increased formality in translated popular writing and an excessive use of appositive linking adverbials in translated academic texts, pointing towards certain translational tendencies.

The study by Delaere, De Sutter, *et al.* (2012) focused on the influence of register on translated language. More specifically, the researchers analyzed how the use of standard or non-standard language in Belgian Dutch translated texts might be affected by the text type (and the source language, see Section 1.2). By pointing out a variable which is typically norm-governed, that is standard versus non-standard language use in the Dutch-speaking region in Belgium (which is largely subject to Netherlandic-Dutch linguistic norms), Delaere, De Sutter, *et al.* (2012) were able to draw a dividing line between registers with and registers without (or with little)

editorial control. Evidence was found for a register-effect and it was observed that text types with a lot of editorial control contained more standard language than text types with limited editorial control. This result aligns with the results of Kruger and Van Rooy (2012), who investigated normalization (adherence to the target language norm) on the basis of frequency of coinages and loanwords as well as frequency of lexical bundles – yielding no significant differences between translated and non-translated texts.

Other systematic investigations of register have been conducted, for instance, by Diwersy, Evert, *et al*. (2014: 202), who concluded that it is "promising to concentrate on individual registers and their contribution to the overall distinction between translations and originals." Lapshinova-Koltunski (2017) investigated the influence of register and translation method. Using hierarchical cluster analysis, the assumption was to find the dataset of translated texts that clusters either according to register or to translation method. Only some of the text types seemed to clearly cluster together into register-specific clusters, while other text types were dispersed amongst more fine-grained clusters. The author concluded that other linguistic properties might be at play and that a more detailed analysis would be required to explain the complex clustering patterns (Lapshinova-Koltunski 2017: 231).

Although the above cited register analyses, in the context of translation, have certainly raised awareness of the importance of the variable "register" in our understanding of the 'DNA' of translated texts, not many firm conclusions can be drawn. Contrary to what one would expect (based on a leveling out hypothesis on the register level), it has become clear that translators are in fact very aware of register differences and that these differences are also reflected in translated texts. Not only has it been pointed out that register is important in guiding translators' choices, but researchers in the field have also been wary of pointing out the complexity of the 'DNA' of translated language. By adding register to the equation, a variable with a broad explanatory potential has now been incorporated into the study of the features of translated language. However, the interdependency between register and other variables, such as linguistic properties (for example, ambiguity) or properties specific to translated texts (for instance, source language), remains unclear. In addition to register, we therefore also want to include the variable "source language" in this study.

1.2. Source language influence on translated language

The body of research on the influence of the source language on translated language is extremely vast, and taken in its largest sense, can be considered to include anything from Schleiermacher's *Methods of Translating* (1813/2004) to Teichs' (2003) concept of shining through, which can be considered the other end of the normalization *continuum* (Hansen-Schirra 2011). Since we are particularly interested in the influence of various variables on translational output, we will only focus on the studies by Delaere, De Sutter, *et al.* (2012), De Sutter, Delaere, *et al*. (2012) and Delaere and De Sutter (2017), who have consistently investigated both source-language and register-effects on translated language. Delaere, De Sutter, *et al.* (2012: 220) concluded that, in addition to a text-type dependent effect, a source-language specific trend was also noticeable, namely that in translated Belgian Dutch, more use was made of standard language compared to non-translated Belgian Dutch. Additional weight for

this conclusion is provided by De Sutter, Delaere, *et al.* (2012: 343), who stated that a translator's linguistic behavior is not only different from a non-translator's behavior, but also varies with text type and source language. A logistic regression analysis further revealed that both text type and source language have a significant impact on onomasiological choice between a formal and a neutral lexeme (De Sutter, Delaere, *et al.* 2012: 343). In a later study, Delaere and De Sutter (2017) again investigated the influence of register and source language, this time on the use of loanwords versus endogenous alternatives in translated and non-translated Belgian Dutch. From their logistical regression analysis, it appears that the source language effect is cancelled out by the register effect. However, since the interaction effect between source language and register could not be included, Delaere and De Sutter (2017) hypothesized that a larger dataset might have revealed interactions between register and source language.

The different multivariate studies by Delaere, De Sutter, *et al.* (2012) show that both register and source language are important factors in shaping onomasiological choice. Although register seems to have a greater impact on onomasiological choice than source language, the authors remain cautious when considering register as the magical explain-it-all variable. The three studies cited here all use lexical profiles of binary choices as a starting point for onomasiological choice, and although they can account for the influence of both register and source language, it remains unclear how a particular choice made by a translator might be influenced by considerations of both source language and register *simultaneously* and *alternately*. We will therefore broaden the hypothesized onomasiological choice from a model with binary choice to a model which covers a choice of 10+ possible onomasiological alternatives in the target language. In addition, we will add a third variable – "cognateness" – which we believe to possibly influence a translator's onomasiological choices.

1.3. Influence of cognateness on translated language

When two words are mutual translations and, in addition to that, are also formally equal or identical, we refer to them as *cognates*.[2] In bilingualism research, the importance of cognateness has been widely investigated and there is overwhelming evidence for the existence of a so-called *cognate facilitation effect* (Costa, Colomé, *et al.* 2000; Schepens, Dijkstra, *et al.* 2012: 157-58, for an overview): bilinguals have faster reaction times and are more accurate when asked to produce cognates compared to control words that only exist in one of the languages they master. Applied to the field of translation, this would imply that translators will also be faster and more accurate when producing cognate translations. In addition, Levý's (1967) so-called Minimax strategy states that translators will give preference to translation solutions which require a minimum of effort for a maximal result. Since the production of a cognate is deemed faster and more accurate than the production of a non-cognate translation, translators are thus expected to choose cognate forms over non-cognate equivalents more often. However, a number of studies in the field of translation have produced results that seem to somewhat contradict this hypothesis. For instance, quasi-experimental research by Shlesinger and Malkiel (2005) led to the conclusion that translators tend to choose a non-cognate translation over a cognate translation when both are (presumably) translationally equivalent. Malkiel (2009) concluded, in a

quasi-experimental study comparing student translations of two source texts from different genres, that it was the source text itself which greatly impacted the use of cognate versus non-cognate translations. A corpus-based study by Vandevoorde, Lefever, *et al.* (2017: 24) added evidence to the importance of taking into account the source language: for Dutch translated from English, the structure of the semantic field of inchoativity – [*to begin,* **beginnen**], a field which displays many cognate translation pairs for English-Dutch – appeared to be directly influenced by the presence of cognate pairs. The authors concluded that cognateness might well be responsible for the different structures of semantic fields in translated and non-translated Dutch inchoativity.

The above findings are far from conclusive about the exact influence of cognateness on translational choices, but do show that a professional translator's behavior might differ from the 'default' bilingual behavior investigated in psycholinguistic research. Indeed, translators might not just "apply" or "succumb to" the cognate facilitation effect and hence behave differently from bilinguals who are not trained language professionals. The findings from the above cited studies do lead to the expectation that cognateness will influence the translator's choices and that there will be an effect of cognateness on translated texts, the exact influence of which is still to be determined. In an attempt to get a grasp of the extent to which translational choices might be influenced by cognateness, we will include, as a variable, the Normalized Levenshtein Distance between the source and the target language lexeme of each corpus observation in our dataset.

2. Methodology

2.1. Data

How do the onomasiological choices made by translators come about in translated texts? From a corpus-based perspective, we can consider translated sentences (with their source language counterparts) – to be present in sentence-aligned parallel corpora for instance – as the final output of (the complex process leading to) translational choices, where each observation of a translated lexeme (in each sentence) can be considered a choice made by the translator.

The data for this study are drawn from the Dutch Parallel Corpus (DPC) (Macken, De Clercq, *et al.* 2011). The DPC is a ten-million-word, sentence aligned, both parallel and comparable corpus. With respect to corpus size, the DPC is, to our knowledge and at the time of writing, the largest available parallel corpus of Dutch. It is furthermore balanced with respect to five text types (external communication, journalistic texts, instructive texts, administrative texts, 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' is the corpus not strictly balanced according to translation direction, but 'only' according to language pair (Paulussen, Macken, *et al.* 2013: 187). However, Delaere and De Sutter pointed out some serious shortcomings with respect to the available text types, endangering the interpretability of the results based on DPC data (Delaere and De Sutter 2017: 88). To overcome this problem, they proposed a new, bottom-up classification of the registers, based on an existing typology and methodologically founded

on research by Biber and Conrad (2009). This led to a restructuring of the DPC into seven registers: broad commercial texts, specialized communication, political speeches, instructive texts, journalistic texts, tourist information and legal texts.[3] For this study, we used the reclassification by Delaere and De Sutter.

2.2. Modeling onomasiological choice with corpus data

In the introduction, we mentioned how onomasiological choice in translated texts has so far been investigated as a binary choice (only two possible translation solutions are usually taken into account). In theory, however, a single trigger lexeme in the source text can lead to an endless list of possible translations (the potential range of onomasiological choices from which the translator can choose). The specific question we are asking runs as follows: if a translator is confronted with a lexeme (in this case, a (proto) typical verb of inchoativity in English or French), what are the possible onomasiological (lexical) choices that are available (in this case, in Dutch), and how will this choice be influenced by factors such as source language, register, and cognateness?

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, Grondelaers, *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). From the perspective of *onomasiology*, and under the assumption that a source language word (for instance, *to start* in "when it *starts* to rain") refers to a concept (in this case, inchoativity), the translator has different words in the target language available (for example, beginnen, starten, aanvangen…) that can be used to express the concept referred to by the source language lexeme in need of translation. In other words, when the translator is confronted with a source language word, he has a (theoretically endless) range of possible translational choices. These potential target language lexemes (the range of choices) are related to the source language word by their meaning: they intend to express the same concept. The lexical realizations used to express a single concept, taken together, can be considered a semantic field.

The idea for this study is to take the semantic field for a concept (in this case, inchoativity) as a starting point, and to consider the lexical realizations for that concept as the range of possible translational outcomes (the range of lexical realizations for the concept of inchoativity will be set to consist of 15 Dutch lexemes). In order to determine a semantic field that can be of use in Translation Studies, the applied method should take into account translational practice, and hence, use translational data. In Vandevoorde, Lefever, *et al.* (2017) and Vandevoorde (forthcoming), we developed a corpus-based technique using the DPC to select candidate-lexemes for a semantic field. The technique, called SMM++, is based on Dyvik's (2004; 2005) Semantic Mirrors Method and uses the idea of back-translation (Ivir 1987) to select a set of lexemes which are thought to be a representation of the semantic field of a (prototypical) lexeme or concept under study.[4] First, all translations of the Dutch lexeme beginnen into French and English were retrieved in the DPC. Then, all translations back into Dutch for the French and English translations of beginnen were looked up again, and only those translations of at least two French or English lexemes

were selected. The latter operation ensured that each of the lexemes selected via this technique is semantically related to the most prototypical expression of the field. For this case study, we use the same semantic field as Vandevoorde (forthcoming), namely the field of inchoativity with Dutch beginnen, French **commencer**, and English *to begin* as the most prototypical expressions of inchoativity. By applying the SMM++, sixteen lexemes expressing inchoativity in Dutch were selected: 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*] and worden [*to become*]. For the current study, we consider these sixteen lexemes as the possible onomasiological choices to express the concept of inchoativity. This means that this study is built on the assumption that when a translator is confronted with a prototypical expression of inchoativity in the source language (either French or English), these sixteen lexemes are the lexical outcomes from which the translator is most likely to choose in the target language (Dutch).

2.3. Datasets for TransDutch$_{FR}$ and TransDutch$_{ENG}$

After applying the lexeme selection technique following the SMM++ and the subsequent selection of the sixteen lexemes, two datasets were created, one for Dutch translated from French (TransDutch$_{FR}$) and one for Dutch translated from English (TransDutch$_{ENG}$). These datasets correspond to the output of the inverse T-image of the SMM++ for TransDutch$_{FR}$ and TransDutch$_{ENG}$ (see Vandevoorde, Lefever, *et al.* 2017 for a detailed account of the selection procedure), but the sets were narrowed down in the following ways:

- Only verbs were selected as French or English source language lexemes.
- A frequency threshold of five observations for each source language verb was set.
- Only data from registers present in both data sets were included.[5]

For TransDutch$_{FR}$, 310 unique corpus observations were selected. Each observation in the dataset for TransDutch$_{FR}$ consists of a source language sentence in French comprising a prototypical verb of inchoativity in French (**commencer**, **débuter**, **démarrer**, **entamer**, **entreprendre**, **lancer**, **se lancer**, **ouvrir**, **partir**) and a target language sentence in Dutch comprising a Dutch inchoative expression (one of the sixteen selected lexemes, as a translation for the French prototypical verb of inchoativity). For TransDutch$_{ENG}$, 490 unique corpus observations were selected. The dataset again corresponds to the output of the inverse T-image of the SMM++, with each observation consisting of a source language sentence in English comprising a prototypical verb of inchoativity in English (*to begin, to open, to set up, to start, to start out, to start up*) and a target language sentence in Dutch comprising one of the sixteen selected lexemes. Next, the two datasets were merged, and a frequency threshold of five observations for each target language lexeme (taking the two source languages together) was applied. The Dutch lexeme aanvang was removed from the dataset (n=3), so that in the final dataset (n=800), only fifteen different Dutch lexemes were present. Finally, for reasons of comparability, only observations for the text types available in both the TransDutch$_{FR}$ and TransDutch$_{ENG}$ datasets were selected for the final dataset, so that only four text types are part of the final dataset: broad

commercial texts, specialized communication, instructive texts and journalistic texts. In this way, the final dataset used for this study consists of 800 unique observations.

For each observation – consisting of a French or English source language sentence and its Dutch translation – we have (i) the annotated inchoative source language lemma in French or English, (ii) the annotated Dutch target language lemma (one of the fifteen Dutch lexemes representing the range of onomasiological choice), (iii) the text type/register. Information about the (iv) cognateness between the source and the target language lexeme for each observation was obtained by calculating the orthographic distance between the two lexemes based on Normalized Levenshtein Distance. This implies that for each translation pair (a source lexeme and its translation), a score is assigned based on the minimum number of insertions, deletions, and substitutions needed to change one lexeme into another one. For instance, the Levenshtein distance (Levenshtein 1966) between English *house* and Dutch huis is 3, since, one deletion (*o* is deleted) and two insertions (i and e are inserted) are needed to arrive from huis to *house*. However, Levenshtein distance depends on word length (the distance between English *toy* and Dutch bal is 3, and between *toy* and huisdier, it is 8, although bal is as different from *toy* as huisdier is from *toy*). In order to obtain similar scores for similar levels of dissimilarity, one can use Normalized Levenshtein Distance (NLD), whereby the Levenshtein distance is divided by the number of letters of the longest string. For *toy* and bal, NLD is then equal to 1:

$$\left(\frac{3\text{ (Levenshtein distance)}}{3\text{ (length of longest string)}}\right)$$

The same goes for the NLD of *toy* and huisdier:

$$\frac{8\text{ (Levenshtein distance)}}{8\text{ (length of longest string)}}.$$

This distance measure (where 1 means complete dissimilarity and 0 complete similarity) can then be converted into a similarity measure by substracting it from one (Divjak and Fieller 2014, 415-16), so that 1 means complete similarity ('full' cognate, orthographic identity) and 0 complete dissimilarity ('full' non-cognate, orthographically completely dissimilar).[6] Following Schepens, Dijkstra, *et al.* (2013: 5), scores for NLD are then calculated[7] as follows:

$$\text{Score} = 1 - \frac{8\text{ (Levenshtein distance)}}{8\text{ (length of longest string)}}$$

2.4. Statistical analyses and visualization

To what extent do the variables "text type," "source language," and "cognateness" influence the translator's choice for a specific target language lexeme? In order to answer this question, we will perform a multinomial logistic regression analysis on our dataset. All analyses for this study were carried out using the statistical software *R*.[8] For the multinomial regression, we used the multinom function from the package nnet.[9] Multinomial regression is used when the response variable is polytomous

(involving more than two categories). In our case, the response variable consists of fifteen categories, namely the fifteen Dutch lexemes. The response variable depends on a set of explanatory variables, in this case: "text type," source language and "cognateness." With a multinomial regression, the log odds ratio (the log of the odds ratio, an exponential function of probabilities) is calculated to find out to what extent the response variable depends on these explanatory variables. Since the resulting coefficients of a multinomial regression are very difficult to interpret, a deviance table of type II tests will be subsequently calculated. As a final step in the analysis, we will use the ctree function from the package partykit[10] to generate classification trees for our final models. With this statistical visualization technique, the decision rules for predicting a categorical outcome can be visualized. Classification trees are also easy to interpret and will facilitate the description of the variation in the data.

3. Results

3.1. Multinomial regression analysis and analysis of deviance with predictor variables "text type" and "source language lexeme"

Before including the newly added predictor variable "cognateness," we first created a model using only the variables "text type" and "source language lexeme." Based on previous research by Delaere, De Sutter, *et al.* (2012), we expect the predictor variable "register" to have a greater impact on the variation in our data than the predictor variable "source language lexeme." Consequently, the first model created contains "text type" and "source language lexeme" as the main effects, as well as their two-way interaction:

TABLE 1

Analysis of deviance (Type II test) for the dataset of inchoativity with predictor variables "text type" and "source language lexeme" (model A)

Response: Dutch target language lexeme			
	LR Chisq	**Df**	**Pr(>Chisq)**
Text type (TT)	196,5622	42	<0.0001***
Source language lexeme (LX)	1121,341	224	<0.0001***
TT:LX	159,1048	672	1
Signif. codes: 0.0001 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Table 1 shows the deviance tests (G-square or Likelihood-Ratio Chi-square) for the two main effects ("text type" and "source language lexeme"), that is the contribution of each predictor to the total variation in the data. Both main effects contribute in a highly significant way to the prediction of the response variable; the interaction effect between "text type" and "source language lexeme" is, however, not significant. We therefore created a second model (Table 2) with "text type" and "source language lexeme" as the main effects, but without the interaction effect:

Table 2

Analysis of deviance (Type II test) for the dataset of inchoativity with predictor variables "text type" and "source language lexeme," without interaction effect (model B)

Response: Dutch target language lexeme			
	LR Chisq	**Df**	**Pr(>Chisq)**
Text type	198,9793	42	p<0.0001***
Source language lexeme	1114,752	210	p<0.0001***
Signif. codes: 0.0001 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Based on these two models, we can conclude that both the text type and the specific source language lexeme are significant predictors of the Dutch target language lexeme that will be chosen by the translator, but the two do not interact.

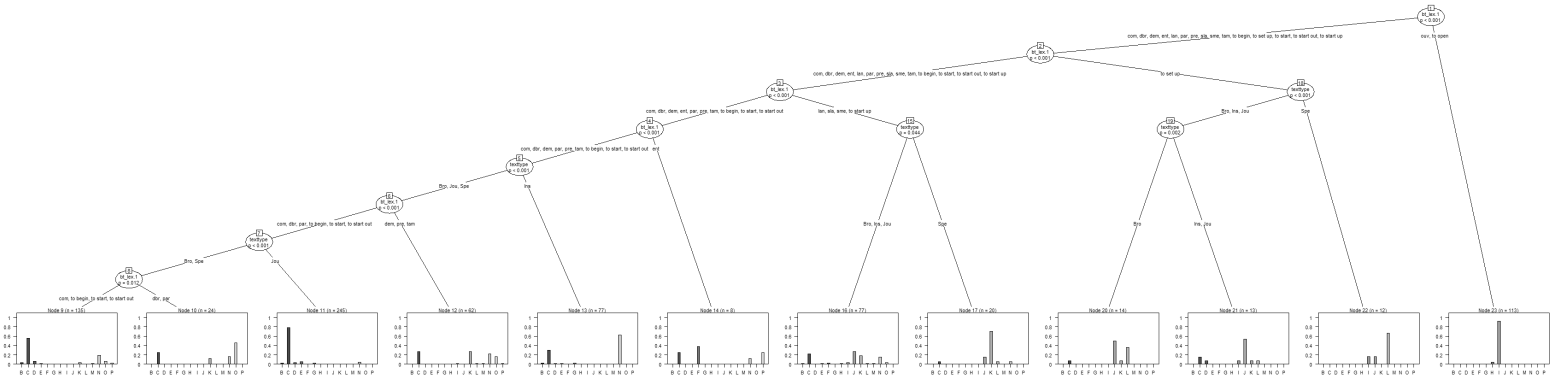
3.2. Visualization of onomasiological choice of model B with classification trees

Figure 1 represents a classification tree for model B (Table 2) with Dutch target language lexemes as response variable, and "source language lexeme" and "text type" as significant predictor variables. In other words, the classification tree is a representation of how onomasiological choice comes about in Dutch translated texts, amongst fifteen Dutch lexemes for inchoativity, when the translator is confronted with an inchoative verb in English or French, and taking into account the significant influence of register and of the specific English or French source language lexeme.

The tree should be read from top right to bottom left. The top right circle is called the *root node* or initial split, every other circle in the tree is called a *leaf node* and represents a further split, which can be read as an "if/then" rule. The initial split of the data is based on the variable "source language lexeme" (node n° 1), and this split is highly significant (p<0.001): if the source language lexeme is **ouvrir** [*to open*] or *to open,* the Dutch translation will be openen [*to open*], and in very few cases ontstaan [*to come into being*]. Next, we move on to node n° 2 (p<0.001): if the source language lexeme is *to set up*, the variation in Dutch translations will depend on the text type. If the text type is specialized communication, then *to set up* is mostly translated as opzetten [*to set up*], and sometimes as openen or oprichten [to establish]. If the text type is broad communication, oprichten will be the most frequently chosen translation, closely followed by opzetten. In instructive and journalistic texts, the most frequently chosen translation of *to set up* is oprichten. Node n° 3 (p<0.001) shows that if the source language word is **lancer** [*to launch*], **se lancer** [*to launch oneself into*], **se mettre** [*to begin*] or *to start up*, the preferred Dutch translation will again depend on the text type. If these verbs appear in specialized communication, these are mostly translated as opstarten, whereas when they are used in broad communication, instructive texts and journalistic texts, the translational variation is more spread: oprichten is most frequently chosen, followed by beginnen, opstarten and starten. The split of node n° 4 (p<0.001) shows that if the source language word is **entrer** [*to enter*], the preferred Dutch translations are komen [*to come*], beginnen [*to begin*], worden [*to become*], and starten [*to start*]. The next node (n° 5, p<0.001) splits off the text type 'instructive texts.' This means that for the remaining verbs – **commencer**

FIGURE 1

Classification tree of 15 Dutch lexemes of inchoativity with predictor variables "source language lexeme" and "text type" (Model B)



bt_lex.1: Predictor variable "source language lexeme"		**texttype: Predictor variable "text type"**		**Categories of response variable Dutch target language lexeme**			
com (**commencer**)	ouv (**ouvrir**)	Bro	broad commercial texts	B	begin	J	oprichten
dbr (**débuter**)	par (**partir**)	Ins	instructive texts	C	beginnen	K	opstarten
dem (**démarrer**)	*to begin*	Jou	journalistic texts	D	eerst	L	opzetten
tam (**entamer**)	*to set up*	Spe	specialized communication	E	gaan	M	start
pre (**entreprendre**)	*to start*			F	komen	N	starten
ent (**entrer**)	*to start up*			G	krijgen	O	van start gaan
lan (**lancer**)	*to start out*			H	ontstaan	P	worden
sla (**se lancer**)	*to open*			I	openen		
sme (**se mettre**)							

[*to begin*], **débuter** [*to begin*, *to start*], **démarrer** [*to start up*], **partir** [*to leave*], **entreprendre** [*to undertake*], **entamer** [*to start*], *to begin*, *to start*, *to start out* – the preferred translations in the instructive text type are beginnen and starten. The sixth node (p<0.001) indicates that if **démarrer**, **entreprendre**, and **entamer** appear in broad commercial texts, journalistic texts or specialized communication, the preferred translations are beginnen, opstarten, starten, and van start gaan. For the remaining French and Dutch lexemes, the preferred translation solution in journalistic texts is beginnen (node n° 7, p<0.001). Finally, within the text types broad commercial texts or specialized communication (node n° 8, p<0.012), **débuter** and **partir** are most often translated as van start gaan. If the source language verb within these two remaining text types is **commencer**, *to begin*, *to start* or *to start out,* then the preferred Dutch translation will be beginnen, and to a lesser extent, **starten**.

3.3. Multinomial regression analysis and analysis of deviance of models including "cognateness" as a predictor variable

As a subsequent step, we included the variable "cognateness" in the model. Since previous research showed that a translator's onomasiological choices might be linked to (levels of) cognateness, we expect the new predictor variable "cognateness" to significantly contribute to the prediction of the variation in our data. It is, however, very difficult to hypothesize how this contribution will relate to the other variables "source language lexeme" and "text type."

The inclusion of a third main effect leads to fourteen possible models (that is fourteen possible combinations of the main effects with one, two or three two-way interactions between the main effects). In order to select the best fitting model out of these fourteen possibilities, we compared the AIC values of the different models. AIC is a "goodness-of-fit measure corrected for model complexity" (Field, Miles, *et al.* 2012: 848). An AIC value is not informative as such, but it can be used for model comparison, with smaller values meaning better-fitting models (Field, Miles, *et al.* 2012: 848). Table 3 gives an overview of the different models with their AIC values (+ refers to 'main effect' and × refers to 'interaction effect').

Table 3

Overview of different models with predictor variables "text type," "source language lexeme," and "cognateness" (NLD), with their AIC values

	MODEL FORMULA	AIC
1	source language lexeme (SL) + NLD	1969.59
2	SL + NLD + SL × NLD	1220.203
3	text type (TT) + SL (model B)	2235.171
4	SL + TT + SL × TT (model A)	2916.266
5	TT + NLD	2757.47
6	TT + NLD + TT × NLD	2647.138
7	TT + SL + NLD	1944.305
8	TT + SL + NLD + NLD × SL	1251.001
9	TT + SL + NLD + NLD × TT	1962.297
10	TT + SL + NLD + SL × TT	2616.032
11	TT + SL + NLD + NLD × SL + NLD × TT	1319.698

12	TT + SL + NLD + NLD × SL + SL × TT	2072.012
13	TT + SL + NLD + NLD × TT + SL × TT	2600.488
14	TT + SL + NLD + NLD × TT +SL × TT + NLD × SL	2156.012

Table 3 shows that the AIC value for Model B (AIC=2235.171, model # 3 from Table 3), which was visualized in Figure 1, is indeed lower than the AIC value for Model A (AIC= 2916.266, model # 4 in Table 3). The comparison of the AIC values in the above table shows in addition that there are two models whose AIC values are much lower than the values for Model B: model # 2 (AIC=1220.203) and model # 8 (AIC=1251.001). In model # 8, the same main effects as in Model B ("text type" and "source language lexeme") are included, and the main effect "NLD" is added as well as the interaction effect between "NLD" and "source language lexeme." In model # 2, only the main effects "source language lexeme" and "NLD" are included as well as the interaction effect between "source language lexeme" and "NLD," but the main effect "text type" is excluded from this model. This model comparison with AIC values seems to show (i) that the inclusion of the main effect "NLD" leads to a better fit of the model (model # 8), but also (ii) that the exclusion of the main effect "text type" leads to an even better fitting model (model # 2). As a final step in the model selection procedure, we compare the values of the Likelihood ratio tests for the two models with the lowest AIC scores (# 2 and # 8) to the null model:

TABLE 4

Comparison of model # 8 and model # 2 to the null model

Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
null model	11186	3092,144973				
SL + NLD + SL × NLD (model # 2)	10766	352,2027359	1 vs 2	420	2739,942237	<0.0001***
TT+ SL + NLD + SL × NLD (model # 8)	10724	299,0011541	2 vs 3	42	53,2015818	0,115232304

Table 4 shows that model # 2 (main effects: "source language lexeme" and "NLD," and interaction effect between "source language lexeme" and "NLD") is significantly better than the null model ($p<0.0001$). Furthermore, the addition of "text type" as a main effect (model # 8, main effects: "text type," "source language lexeme" and "NLD," and interaction effect between "source language lexeme" and "NLD") does not lead to a significantly better model, compared to model # 2. As a consequence, and based on both AIC values and the comparison to the null model, we can conclude that model # 2 is the best fitting model. We can then further conclude that, for this data set, the predictive power of the independent variable "NLD" is greater than that of "text type."

Table 5

Analysis of Deviance (Type II test) for the dataset of inchoativity with predictor variables "source language lexeme," and "cognateness" (NLD) with their interaction effect (model # 2)

Response: Dutch target language lexeme			
	LR Chisq	**Df**	**Pr(>Chisq)**
SL	230,0923	42	<0.0001***
NLD	200,453	14	<0.0001****
SL × NLD	1141.43	210	<0.0001***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Table 5 shows the likelihood-ratio Chi-square (LR Chisq) for the main effects "source language lexeme" and "NLD" and their two-way interaction. Both main effects as well as the interaction effect are highly significant, contributing to the explanation of the response variable. The model in Table 5 (which corresponds to model # 2 from Table 3) is then the final model.

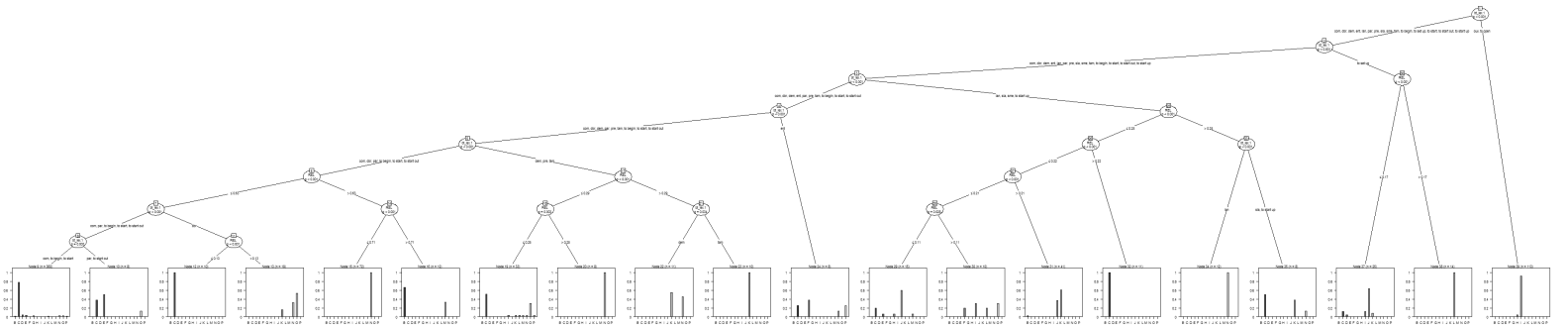
3.4. Visualization of onomasiological choice of model # 2 with classification trees

Figure 2 shows a classification tree for model # 2 with Dutch target language lexemes as response variable, and "source language lexeme," and "NLD" as significant predictor variables. It represents the predictive path towards an onomasiological choice amongst 15 Dutch lexemes of inchoativity in Dutch translated texts. In other words, Figure 2 is a representation of how the translator possibly chooses an inchoative Dutch verb as a translation of an inchoative verb in English or French. The onomasiological predictions within this model are based on the significant influence of the specific English or French source language lexeme as well as the NLD between French or English source language lexeme and the Dutch target language lexeme.

The initial split of the data is based on the variable "source language lexeme" (node n° 1, p<0.0001). Just as for Model B (Figure 1), it is predicted that the Dutch translation openen will be produced if the source language lexeme is either **ouvrir** or *to open*. Importantly, in the case of openen/*to open*/**ouvrir**, neither text type nor NLD seems to have an impact on the translator's onomasiological choice (although the NLD between *to open* and openen is 0.67, indicating they can be considered cognates). Just as for Model B, the prediction from node n° 2 (p<0.001) is that the variation in Dutch translations of *to set up* will depend on the other independent variable in the model, in this case, "NLD." The prediction is that when the NLD between source and target language lexeme is >0.17, the preferred translation will be opzetten, whereas when NLD ≤0.17, the preferred translation will be oprichten. Again similarly to Model B, node n° 3 (p<0.001) indicates that if the source language lexeme is **lancer**, **se lancer**, **se mettre** or *to start up*, the preferred Dutch translation will depend on the other independent variable ("NLD"). If NLD >0.25, the Dutch translation will depend on the source language lexeme: if the source language lexeme is **lancer**, the preferred translation will be starten, if the source language lexeme is **se lancer** or *to start up*, the preferred translations will be beginnen and opstarten and to a lesser extent starten. NLD>0.22 predicts beginnen as the preferred translation, NLD>0.21 predicts oprichten and opstarten, NLD>0.11 predicts komen, openen,

FIGURE 2

Classification tree of 15 Dutch lexemes of inchoativity with predictor variables "source language lexeme," and "cognateness" (NLD) (model # 2)



bt_lex.1: Predictor variable "source language lexeme"		**REL: Predictor variable "cognateness" (NLD)**		**Categories of response variable Dutch target language lexeme**			
com (**commencer**)	ouv (**ouvrir**)			B	begin	J	oprichten
dbr (**débuter**)	par (**partir**)			C	beginnen	K	opstarten
dem (**démarrer**)	*to begin*			D	eerst	L	opzetten
tam (**entamer**)	*to set up*			E	gaan	M	start
pre (**entreprendre**)	*to start*			F	komen	N	starten
ent (**entrer**)	*to start up*			G	krijgen	O	van start gaan
lan (**lancer**)	*to start out*			H	ontstaan	P	worden
sla (**se lancer**)	*to open*			I	openen		
sme (**se mettre**)							

opzetten, and van start gaan, and NLD≤0.11 predicts beginnen and oprichten as preferred translations. Node n° 4 ($p<0.001$) shows exactly the same prediction as node n° 4 in Model B: if the source language word is **entrer**, the preferred Dutch translations are komen [*to come*], beginnen [*to begin*], worden [*to become*] and starten [*to start*]. From node n° 5 ($p<0.001$) the predictions are made that, if NLD>0.29 and the source language lexeme is **démarrer**, the Dutch translations will be opstarten and starten, if the source language lexeme is **entamer**, the Dutch translation will be opstarten. If NLD>0.25 and ≤0.29, the prediction is made that **démarrer**, **entreprendre**, and **entamer** will be translated as starten, if NLD≤0.25, the preferred translations are beginnen and van start gaan. For the remaining source language lexemes (**commencer**, **débuter**, **partir**, *to begin, to start, to start out*), the prediction is made from node n° 6 ($p<0.001$) that, if NLD>0.71, the preferred translations are *begin* and *start,* if NLD>0.63 and ≤0.71, the preferred translation is starten. Node n° 7 ($p<0.001$) shows that the preferred translation for **débuter** is beginnen if NLD<0.13, and opstarten, starten or van start gaan if NLD>0.13. Finally, node n° 8 ($p<0.003$) shows that if **commencer**, *to begin* or *to start* is the source language lexeme, the preferred translation is beginnen, for **partir** and *to start out,* it is beginnen, gaan and van start gaan.

4. Conclusion

4.1. Comparison of the classification trees

The classification trees depicted in Figures 1 and 2 are based on different multinomial regression analyses, carried out with different sets of predictor variables. Although these trees inevitably lead to different interpretations of the predictive path towards the onomasiological variants, there are also some remarkable similarities between the two figures. First, the eight main nodes (node n° 1-8) in each of the two figures are to a large extent similar, splitting up the data into subgroups that are first and foremost based on the variable "source language lexeme" (node n° 5 and n° 7 in Figure 1 and node n° 6 in Figure 2 are the only exceptions). These groups of source language lexemes are as follows:

Table 6

Subgroups of source language lexemes as predicted by the classification trees in Figures 1 and 2 (differences in subgroups in Figure 2 vs Figures 1 are in bold)

Figure 1	Figure 2
to open – **ouvrir**	*to open* – **ouvrir**
to set up	*to set up*
lancer, **se lancer**, **se mettre**, *to start up*	**lancer**, **se lancer**, **se mettre**, *to start up*
entrer	**entrer**
démarrer, **entamer**, **entreprendre**	**démarrer**, **entamer**, **entreprendre**
débuter	**débuter**, **partir**
commencer, *to begin*, *to start*	**commencer**, *to begin*, *to start*, *to start out*
partir, *to start out*	∅

In addition, some of these source language lexemes directly predict the Dutch target language lexeme. For instance, the Dutch target language lexeme openen is

predicted based on the source language lexemes *to open* and **ouvrir**. Neither the text type nor the NLD play a role in the prediction of openen (although openen and *to open* NLD=0.67 can be considered cognates). For **entrer**, the preferred translations are komen, beginnen, worden and starten. Again, neither text type nor NLD seems to play a role here. From both figures, it can be concluded that *to set up* will be translated by either opzetten or by oprichten with specialized communication and NLD>0.17 being predictor of opzetten and broad commercial texts and NLD≤0.17 predicting oprichten.

4.2. Discussion

In this study, we have made an attempt to visualize how onomasiological choice within the field of inchoativity comes about in Dutch translated texts. The two visualizations are both plausible and interpretable representations of how onomasiological choice is determined by variables such as "text type," "source language lexeme," and "cognateness," offering alternate perspectives on how the significant predictors might lead to specific onomasiological choices. Our study inevitably has a number of limitations. For instance, NLD scores are different for each translational pair and each semantic field, so other language pairs and other cross-linguistic semantic fields will display different NLD scores. More studies including "NLD" as a variable will be needed in order to better understand the possible influence of orthographic distance on onomasiological choices in translated texts. Based on the limited evidence of this case study, we could nevertheless tentatively conclude that a higher degree of cognateness for a specific semantic field between a source and a target language will lead to a higher likelihood that NLD will significantly influence onomasiological choices.

Whether one finds Figure 1 or Figure 2 more insightful will depend on the interpreter's point of view. In our opinion, Figure 1 ("text type" and "source language lexeme" as predictor variables) leads to more "applied" insights. The creation of a classification tree with "source language lexeme" and "text type" as predictor variables can be considered an advanced application of use of corpora in translator training: Figure 1 gives insights into text type specific onomasiological preferences as attested in translated texts.[11] From the point of view of translation theory, the model fitting procedure leading to the selection of model # 2 and the classification tree in Figure 2 leads to the important insight that information about the orthographic distance between a source and a target language lexeme contributes in a highly significant way to the explanation of the variation in the data (in this case, influences onomasiological choice). Although it is common to consider as cognates those pairs of words that exhibit an NLD of ≥0.5, Figure 2 shows that differences in NLD, even when NLD is as low as ≤0.11 can predict onomasiological choice. If NLD is indeed a significant predictor of onomasiological choice, this in turn would lead to the important question whether translated texts display a higher ratio of cognates (with the source language from which they are translated) compared to non-translated texts, a question which remains as of now unanswered, but possibly opens up an interesting avenue of research for researchers in CBTS.

ACKNOWLEDGEMENTS

The author would like to thank Stefan Th. Gries for kindly providing us with the *R* script for the calculation of Normalized Levenshtein Distances, and Koen Plevoets for his invaluable help with the statistical analyses. I would also like to express my sincere gratitude to the anonymous reviewers for their pertinent comments.

NOTES

1. The term *translator* is used throughout this paper to refer to the wider notion of the *translation professional,* and therefore includes the (team of) translator(s), editor(s) and any other intermediaries or agents (jointly) involved in the translation process and the coming about of the onomasiological choices that are visible in translated texts. We agree with Kruger that editorial intervention should be taken into consideration when investigating the features of translated language (Kruger 2017). However, the Dutch Parallel Corpus we are using for this study unfortunately does not contain multiple versions – edited and unedited – of the same text. We therefore cannot take this variable into account. In addition, the scope of this study is not to comprehend the role of the different translation professionals in the ultimate decision, but rather to understand how linguistic features might impact upon onomasiological choice.
2. In psycholinguistic research, it is common to use the term *cognates* in this sense, leaving aside the prerequisite of a common etymological origin between two words (which is indeed part of the strict 'linguistic' definition of cognates). In this study, we will use the psycholinguistic definition, leaving out etymological origin.
3. For a more in depth discussion on the register re-classification and examples of content for each text type, see Delaere and De Sutter (2017).
4. In Vandevoorde (2016), Vandevoorde, Lefever, *et al.* (2017) and Vandevoorde (forthcoming), a semasiological perspective was presented, where the created semantic fields were considered "possible and plausible representations of the different meanings of a word under study (beginnen)" (Vandevoorde 2016: 3). In the present study, an onomasiological point of view is taken, meaning that the created semantic fields are thought to represent the different ways of expressing (lexical choices) one and the same concept (in our case, inchoativity). The construction of the semantic fields based on the SMM++ allows for both perspectives (Vandevoorde 2016: 3-4).
5. For instance, the observations from political speeches in English were not included since there were no observations from political speeches in French.
6. Schepens, Dijkstra, *et al.* (2013: 9) consider those pairs of words with a NLD of 0.5 and higher to be cognates.
7. The calculation of the NLD-scores for the 800 observations was carried out using the *R* script developed by Stefan Th. Gries for (Normalized) Levenshtein Distance. The authors want to thank Stefan Th. Gries for kindly providing us with the script.
8. R Core Team (2017): *R*. Version 3.4. Visited 1 September 2017, <http://www.r-project.org>.
9. Ripley, Brian and Venables, William (2 February 2016): Package 'nnet.' *R*. Version 7.3-12. Visited 10 September 2017, <https://cran.r-project.org/web/packages/nnet/nnet.pdf>.
10. Hothorn, Torsten, Seibold, Heidi, and Zeileis, Achim (20 September 2016): partykit: A Toolkit for Recursive Partytioning. *R*. Version 1.1-1. Visited 1 September 2017, <http://partykit.R-Forge.R-project.org/partykit>.
11. The visualization of a classification tree with the three main effects "text type," "NLD," and "source language lexeme" (model # 7), which from this point of view would have been very interesting too, yields a classification tree where only NLD and source language account for all the nodes.

REFERENCES

Biber, Douglas and Conrad, Susan (2003): Register Variation: A Corpus Approach. *In*: Deborah Schiffrin, Deborah Tannen, and Heidi Hamilton, eds. *The Handbook of Discourse Analysis*. Oxford: Blackwell, 175-96.

Costa, Albert, Colomé Angels, and Caramazza, Alfonso (2000): Lexical Access in Speech Production. The Bilingual Case. *Psicológica*. 21:403-37.

Chesterman, Andrew (2004): Hypotheses About Translation Universals. *In*: Gyde Hansen, Kirsten Malmkjaer and Daniel Gile, eds. *Claims, Changes and Challenges in Translation Studies*. Amsterdam/Philadelphia: John Benjamins, 1-13.

Delaere, Isabelle, De Sutter, Gert, and Plevoets, Koen (2012): Is Translated Language More Standardized than Non-Translated Language? Using Profile-Based Correspondence Analysis for Measuring Linguistic Distances between Language Varieties. *Target*. 24(2):203-24.

Delaere, Isabelle and De Sutter, Gert (2017): Variability of English Loanword Use in Belgian Dutch Translations. Measuring the Effect of Source Language and Register. *In*: Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, eds. *Empirical Translation Studies*. Berlin/Boston: Mouton de Gruyter, 81-112.

De Sutter, Gert, Delaere, Isabelle, and Plevoets, Koen (2012): Lexical Lectometry in Corpus-Based Translation Studies. Combining Profile-Based Correspondence Analysis and Logistic Regression Modeling. *In*: Michael Oakes and Ji Meng, eds. *Quantitative Methods in Corpus-Based Translation Studies. A Practical Guide to Descriptive Translation Research*. Amsterdam/Philadelphia: John Benjamins, 325-45.

Divjak, Dagmar and Fieller, Nick (2014): Cluster Analysis. Finding Structure in Linguistic Data. *In*: Dylan Glynn and Justyna Robinson, eds. *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. Amsterdam/Philadelphia: John Benjamins, 405-41.

Diwersy, Sascha, Evert, Stefan, and Neumann, Stella (2014): A Weakly Supervised Multivariate Approach to the Study of Language Variation. *In*: Benedikt Szmrecsanyi and Bernhard Wälchli, eds. *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*. Berlin/Boston: Mouton de Gruyter, 174-204.

Dyvik, Helge (2004): Translations as Semantic Mirrors. From Parallel Corpus to Wordnet. *In*: Karin Aijmer and Bengt Altenberg, eds. *Advances in Corpus Linguistics*. Amsterdam/New York: Rodopi, 311-26.

Dyvik, Helge (2005): Translations as a Semantic Knowledge Source. *In:* Margit Langemets and Priit Penjam, eds. *Proceedings of the Second Baltic Conference on Human Language Technologies*. (HLT'2005: Second Baltic Conference on Human Language Technologies, Tallinn, 4-5 April 2005). Tallinn: Institute of the Estonian Language/Institute of Cybernetics (Tallinn University), 27-38.

Field, Andy, Miles, Jeremy, and Field, Zoë (2012): *Discovering Statistics Using R*. London: Sage Publications.

Geeraerts, Dirk, Grondelaers, Stef, and Bakema, Peter (1994): *The Structure of Lexical Variation. Meaning, Naming, and Context*. Berlin/Boston: Mouton de Gruyter.

Halverson, Sandra L. (2015): Cognitive Translation Studies and the Merging of Empirical Paradigms. The Case of 'literal Translation.' *Translation Spaces*. 4(2):310-40.

Halverson, Sandra L. (2017): Gravitational Pull in Translation. Testing a Revised Model. *In*: Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, eds. *Empirical Translation Studies*. Berlin/Boston: Mouton de Gruyter.

Hansen-Schirra, Silvia (2011): Between Normalization and Shining-Through. Specific Properties of English-German Translations and Their Influence on the Target Language. *In*: Svenja Kranich, Viktor Becher, Steffen Höder, *et al*., eds. *Multilingual Discourse Production. Diachronic and Synchronic Perspec*tives. Hamburg Studies on Multilingualism. Amsterdam/Philadelphia: John Benjamins, 133-62.

Ivir, Vladimir (1987): Functionalism in Contrastive Analysis and Translation Studies. *In*: René Dirven and Vilém Fried, eds. *Functionalism in Linguistics*. Amsterdam/Philadelphia: John Benjamins, 471-81.

Kruger, Haidee, and Van Rooy, Bertus (2012): Register and the Features of Translated Language. *Across Languages and Cultures*. 13(1):33-65.

Kruger, Haidee (2017): The Effects of Editorial Intervention. Implications for Studies of the Features of Translated Language. *In*: Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, eds. *Empirical Translation Studies*. Berlin/Boston: Mouton de Gruyter.

Lapshinova-Koltunski, Ekaterina (2017): Exploratory Analysis of Dimensions Influencing Variation in Translation. The Case of Text Register and Translation Method. *In*: Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, eds. *Empirical Translation Studies*. Berlin/Boston: Mouton de Gruyter.

Levenshtein, Vladimir (1966): Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*. 10(8):707-10.

Macken, Lieve, De Clercq, Orphée, and Paulussen Hans (2011): Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus. *Meta*. 56(2):374-90.

Mauranen, Anna (2008): Universal Tendencies in Translation. *In*: Gunilla Anderman and Margaret Rogers, eds. *Incorporating Corpora: The Linguist and the Translator*. Clevedon/Tonawanda: Multilingual Matters, 32-48.

Malkiel, Brenda (2009): Translation as a Decision Process. Evidence from Cognates. *Babel*. 55(3):228-43.

Neumann, Stella (2014): *Contrastive Register Variation, A Quantitative Approach to the Comparison of English and German*. Berlin/Boston: Mouton de Gruyter.

Paulussen, Hans, Macken, Lieve, Vandeweghe, Willy, *et al*. (2013): Dutch Parallel Corpus: A Balanced Parallel Corpus for Dutch-English and Dutch-French. *In*: Peter Spyns and Jan Odijk, eds. *Essential Speech and Language Technology for Dutch. Results by the STEVIN Programme*. Heidelberg/New York/Dordrecht/London: Springer, 185-99.

Schepens, Job, Dijkstra, Ton, and Grootjen, Franc (2012). Distributions of Cognates in Europe as Based on Levenshtein Distance. *Bilingualism: Language and Cognition*. 15(1):157-66.

Schepens, Job, Dijkstra, Ton, Grootjen, Franc, *et al*. (2013). Cross-Language Distributions of High Frequency and Phonetically Similar Cognates. *PLoS ONE*. 8(5):e63006.

Schleiermacher, Friedrich (1813/2004): On the Different Methods of Translating. (Translated by Susan Bernofsky) *In*: Lawrence Venuti, ed. *The Translation Studies Reader*. 2nd ed. London/New York: Routledge, 43-63.

Shlesinger, Miriam and Malkiel, Brenda (2005): Comparing Modalities: Cognates as a Case in Point. *Across Languages and Cultures*. 6(2):173-93.

Teich, Elke (2003). *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Berlin/New York: Mouton de Gruyter.

Vandevoorde, Lore (2016): On Semantic Differences: A Multivariate Corpus-Based Study of the Semantic Field of Inchoativity in Translated and Non-Translated Dutch. Doctoral dissertation. Ghent: Ghent University.

Vandevoorde, Lore, Lefever, Els, Plevoets, Koen, *et al*. (2017): A Corpus-Based Study of Semantic Differences in Translation. The Case of Dutch Inchoativity. *Target*. 29(3): 388-415.

Vandevoorde, Lore (forthcoming): *Semantic Differences in Translation: Exploring the Field of Inchoativity*. Translation and Multilingual Natural Language Processing. Berlin: Language Science Press.

Xiao, Richard (2010): How Different Is Translated Chinese from Native Chinese? A Corpus-Based Study of Translation Universals. *International Journal of Corpus Linguistics*. 15(1):5-35.