How do translators select among competing (near-)synonyms in translation? A corpus-based approach using random forest modelling

Pauline De Baets and Gert De Sutter

Ghent University

This article investigates how translators choose between multiple competing onomasiological variants to express (verbal) inchoativity in English-to-Dutch translations. Using a corpus-based multifactorial research design, we measure the impact of three well-known socio-cognitive mechanisms on the actual choice, namely the complexity principle, risk aversion, and cognate exposure. We apply the behavioural profile method, which allows us to operationalise these three explanatory mechanisms via ID-tags, and we then use conditional random forest modelling to determine the impact of each mechanism on the choice between four competing verbs of inchoativity. The results of our analyses show that the complexity principle plays a clear role in translated texts, as there is a significant preference for the active construction and for prototypical verbs in passive constructions. Genre-specific risk-averse behaviour as well as cognate avoidance were not observed.

Keywords: cognates, risk aversion, complexity principle, behavioural profile, conditional random forest modelling, verbs of inchoativity, Dutch

1. Introduction

While a translator[1] is translating, often several translation options are cognitively activated to express a certain idea that is encoded in the source text. These options can be lexical, pragmatic, and grammatical in nature, and are referred to in lexical variation studies as onomasiological choices. During the last decades, numerous studies within corpus-based translation studies have shown that lexical and grammatical onomasiological choices made by translators deviate – at least to some extent – from the choices made by writers of original, non-translated texts (for an overview, see Kruger and van Rooy [2012]). However, the adopted onomasiological perspective in corpus-based translation studies is mostly limited to a binary choice (i.e., the choice between two possible translation outcomes). Ongoing research on explicitation in translation that investigates how the translator chooses between an explicit variant and an implicit one is illustrative of this (e.g., Olohan and Baker [2000] and Olohan [2003] on optional *that* and Van Beveren, De Sutter, and Colleman [2018] on optional *om* in Dutch infinitival clauses). In this article, we want to broaden the onomasiological scope by studying multiple competing translation outcomes. Indeed, there are often more than two ways of referring to a certain concept, and by not taking into account this more complex lexical-onomasiological situation, corpus-based translation studies misses opportunities to gain better, more accurate insight into what drives translation behaviour. In this study, we aim to include multiple (prototypical) near-synonymous verbs in Dutch that can be used to express inchoativity (viz. *beginnen* 'to begin', *starten* 'to start', *opstarten* 'to start up', and *van start gaan* 'to launch '). More particularly, we investigate what the underlying motivations are to opt for one verbal lexeme over the others, and to what extent these motivations differ in translated texts compared to non-translated texts. The reason to focus on verbal inchoativity in Dutch is purely

[1] As in all corpus-based studies, the notion of 'translator(s)' should be understood here as the entirety of all actors that intentionally or unintentionally affect linguistic decision-making during the translation process, at one or several stages, giving rise to specific linguistic characteristics in the published version of the translated text, and which cannot be traced back to a specific choice by a specific actor at a specific moment. The concept of 'linguistic choice' is thus to be interpreted as the resultant of a potentially complex interaction between different actors, and not just as the specific choice of a specific translator.

instrumental: it allows us to investigate the patterning of *multiple* onomasiological options as a function of the text status (translated vs. non-translated) and as a function of different underlying motivations (see Section 2). Most languages, including Dutch, have a large set of words and constructions to express inchoativity, which in general signals a change of state, or the onset of a new situation, and often implies a transition from non-action to action (Shi 1990; Piñón 2001; Divjak and Gries 2009; Marín and McNally 2011; Verroens 2011). For our study, we will use a parallel corpus of English-to-Dutch translations and a comparable corpus of authentic Dutch texts, which are part of the ten-million-word bidirectional Dutch Parallel Corpus (Macken, de Clercq, and Paulussen 2011).

At the same time, we also want to contribute to the increased attention to cognitive explanations of translational behaviour, thereby moving beyond explanations in terms of translation universals. Many patterns that were once believed to be unique to translation are now being explored as potential effects of a much more encompassing bilingual language production (Halverson 2003, 2013; House 2008, 2013; De Sutter, Delaere, and Plevoets 2012; Szymor 2018; De Sutter and Lefer 2020; Vandevoorde 2020). We want to continue this line of research by taking into consideration three possible socio-cognitive explanations, as identified by Kruger (2019), for the different onomasiological choices in translated and non-translated texts that are not restricted to translational behaviour alone, but are often used to explain usage patterns in (bilingual) language production in general, and thus have much more explanatory power: the complexity principle (Rohdenburg 1996), risk aversion (Pym 2015), and cognate exposure (Costa, Colomé, and Caramazza 2000; Malkiel 2009a, 2009b). These three explanations are discussed in Section 2.

The orientation towards different socio-cognitive explanations for translation behaviour inevitably goes hand-in-hand with the need to implement advanced statistical techniques, as multiple explanatory variables, serving as proxies for the explanatory mechanisms, need to be

taken into account simultaneously (Halverson 2015, 2017; Gries 2018). These techniques will enable us to decide which of the explanations is the more powerful and which (if any) is non-significant. In order to uncover the explanations that guide translators (compared to non-translating writers) in their onomasiological choice, we combine two innovative methods, namely the behavioural profile approach (Divjak and Gries 2006; Gries and Divjak 2009; Szymor 2015) and random forests (Levshina 2015, 2020). The behavioural profile is a usage-based contextual method that allows us to chart, in a fine-grained way, the multiple syntactic, contextual, and semantic parameters that may influence the choice of a specific lexeme of inchoativity. The method has already been proven useful in comparing and contrasting near-synonymous words (Divjak and Gries 2009; Gries and Divjak 2009; Divjak 2010; Jansegers, Vanderschueren, and Enghels 2015; Szymor 2015). Random forest modelling gives us an overview of the variable importance of each parameter in the onomasiological decision process. The combination of these two methods allows us to determine the exact impact of the three socio-cognitive explanations mentioned above on the onomasiological decision between near-synonymous verbs of inchoativity, and to investigate whether those mechanisms operate differently in translated language than in non-translated language.

The paper is structured as follows: Section 2 gives an overview of previous research in corpus-based translation studies on (near-)synonyms and presents a detailed discussion of the three socio-cognitive explanations. In Section 3, we introduce the methodological framework that underlies our research and in Section 4, we describe the results of our study. In the final section, we summarise the main findings and indicate some interesting possibilities for further research.

2. Previous research: Near-synonymy and socio-cognitive explanations

Near-synonymous words convey roughly the same meaning, but still bear subtle differences. Even if two words refer to the same concept, they are not always perfectly interchangeable as they can name the concept from a different perspective (Edmond and Hirst 2002; Divjak 2010). In general, it is asserted that near-synonyms can differ with respect to any aspect of their meaning (Cruse 1986), namely stylistic variations, expressive variations, denotational variations, and structural variations. As a consequence, when investigating near-synonymy, it is the task of the researcher to uncover which differences exist between near-synonyms and what the semantic and functional relation is between near-synonyms (Divjak and Gries 2006).

In corpus-based translation studies, quite a few studies focus on the choice between near-synonyms. One example is the study of Blum-Kulka and Levenston (1983), which reveals that translators tend to select more familiar target-language words over less familiar synonyms. Egan (2012) focuses on the different possible Norwegian translations for *to start* and *to begin*, and finds that they are very close synonyms. However, he also finds that *starte* is used as a translation of *to start* but never of *to begin*, suggesting a cognate effect. Szymor (2015) is one of the first studies to explore near-synonymy in translation in some depth, combining corpus research with cognitive-linguistic explanations. She introduces the behavioural profile method as an objective method for capturing the word meaning of (near-)synonyms, and shows that their use is different in translated and non-translated language, referring to Halverson's (2003, 2017) gravitational pull hypothesis as a cognitive explanation for these differences. Although her study is innovative, both in method and in theoretical framework, there are also two limitations: first, it is limited to only two deontic modal verbs, and second, there is no in-depth analysis of the differences between near-synonyms in translated and non-translated texts. In a follow-up study in 2018, Szymor looks into more detail into the aspectual differences between translated and non-translated Polish. Szymor (2018) investigates six modal markers, and uncovers that the perfective form of the infinitive is preferred in translated texts, whereas the

imperfective form is more likely in non-translated texts. Szymor (2018) explains these differences by referring to differing degrees of entrenchment and chunking. Finally, Vandevoorde (2020) applies clustering methods in order to learn more about the semantic structure of near-synonyms of inchoativity in translated and non-translated texts. She is the first to closely investigate semantic differences, rather than structural differences between near-synonyms in translated and non-translated texts, finding differences in the organisation of the semantic field of inchoativity in translated texts compared to that of non-translated texts.

However, a stable, fully fledged explanatory framework for these differences in translated and non-translated language is still lacking. After having largely abandoned the translation universals approach introduced by Baker (1993), scholars have been using systemic-functional linguistics (Steiner 1997) and relevance theory (Alves and Gonçalves 2010) to theorise their empirical findings. Recently, there have been attempts to provide more encompassing socio-cognitive explanations for patterns typically found in translated language, building on Halverson (2017) and Kotze (2022). Halverson is the first to start combining theoretical assumptions from cognitive grammar with findings from studies in bilingualism in order to develop cognitive explanations for discriminating features found in translated texts. She finds that many patterns that were believed to be unique to translation are natural effects of bilingual language production (Halverson 2003, 2010, 2013, 2017; House 2013). In two recent papers, Kotze (2022) and Kruger and De Sutter (2018) follow up on that approach by combining explanations in terms of cognitive abilities and restrictions with the internalised role-perception of translators as highly professional language and communication mediators. More particularly, they distinguish between three possible sources of explanation for behavioural differences between translation and non-translation: (cognitive) complexity, (social) risk aversion, and cognate exposure.

2.1 Complexity

The first explanatory principle is commonly referred to as the complexity principle, which states that "in the case of more or less explicit grammatical options, the more explicit one(s) will tend to be favoured in cognitively more complex environments" (Rohdenburg 1996, 151), as more explicit expressions are easier to process. The complexity principle has been empirically verified in corpus-linguistic and psycholinguistic studies, for written and spoken language, and for translated and non-translated language (see Rohdenburg 1996, 2016; Ferreira and Dell 2000; Kruger 2019; Seeber 2013; Kruger and De Sutter 2018; Pijpops et al. 2018). Most of these studies have focused on grammatical alternations, but the complexity principle is applicable to lexical alternations between near-synonyms (i.e., onomasiological variants) as well: the selection of a more prototypical (or a more frequent) synonym lowers cognitive pressure in contexts where it is high since less cognitive effort needs to be invested in prototypical, and hence more entrenched and more accessible, lexemes.

Since translation involves bilingual language processing, it is plausible to assume that translation requires increased cognitive effort compared to monolingual writing (Kruger and De Sutter 2018). This assumption is supported by empirical evidence on pause behaviour (Immonen 2006): translation triggers more and longer pauses at word and clause level compared to original text production, which is linked to increased mental processing. Consequently, we assume that the more frequent and/or more prototypical lexeme will be preferred in translated language. The effect of complexity on translation has been confirmed for grammatical alternations by Kruger (2019), who found that translations in South African English exhibit the lowest frequency of omission of the complementizer *that*, thus suggesting a greater preference for the more explicit (and the more frequent, and more accessible) construction. In a follow-up study, Kruger and De Sutter (2018) show that this tendency holds

in grammatically non-complex contexts and across all registers: "higher cognitive effort involved in translation may reduce the available capacity and lead to reduced sensitivity to register preferences for omission and the selection of the default option" (Kruger and De Sutter 2018, 279). The assumption here is that translators' cognitive load is higher because of bilingual activation and switching costs, and that the translator, therefore, is more sensitive to complexity issues. Based on this, we hypothesise that onomasiological choices in translated language will be affected more by indicators of complexity compared to non-translated language, yielding a higher rate of frequent and prototypical lexemes being selected in translated language.

2.2 Risk aversion

A second explanatory mechanism that has been proposed is risk avoidance, which is a social explanation. Translators appear to invest much effort in those segments of a text where the risk of misinterpretation is higher (Pym 2005), and tend to select the 'safest' option, which minimises the communicative risk. Pym (2008, 326) argues that this is especially true "when there are no rewards for them to do otherwise." Although authors of original texts will also avoid misinterpretations, the stakes are higher for translators, since translation involves communication in a context with greater linguistic and cultural distance between the author and the reader than original texts (Becher 2010; Pym 2015; Kruger 2019). As a consequence, translated texts have been shown to be linguistically more conservative than non-translated texts, which is illustrated by a higher tendency to standardise non-standard source-text segments (for instance, by replacing regional or non-standard varieties by standard ones) or to over-use frequent lexemes or grammatical constructions (e.g., Pym 2005, 2008, 2015; Becher 2010; Delaere, De Sutter, and Plevoets 2012; Delaere and De Sutter 2013; Saridakis 2015;

Kruger and De Sutter 2018; Kruger 2019). Hence, for this study, we hypothesise that translators will have a higher inclination to use the 'safe' and 'most common' verbal lexeme of inchoativity in a specific context (e.g., a genre) in order to minimise communicative risk. We test whether this leads to an over-conventionalisation effect, with the most frequent lexeme in a specific genre being overrepresented in translated texts.

2.3 Cognate exposure

A final explanation that may account for onomasiological choices in translation is the presence of cognates in the source text. Cognates are words that have similar orthographic-phonological forms in two (or more) languages (e.g., the Dutch–English *tourist–tourist*); they also often share the same etymology and show a high degree of semantic overlap (Costa, Colomé, and Caramazza 2000; Gollan and Acenas 2004; Sherkina 2004; Malkiel 2009a; Schepens, Dijkstra, and Grootjen 2012; Balling 2013). In psycholinguistics, it has been shown that cognates have a facilitating effect in bilingual language processing (Carroll 1992; Costa, Miozzo, and Caramazza 1999; Costa, Colomé, and Caramazza 2000; Costa, Santesteban, and Caño 2005): cognates are processed faster, in production and comprehension, both in isolated test contexts (Dijkstra, Grainger, and van Heuven 1999; Costa, Colomé, and Caramazza 2000; Kroll, Dietz, and Green 2000; Costa, Santesteban, and Caño 2005) and in sentence contexts (van Assche et al. 2009). For instance, van Hell and Dijkstra (2002) show that Dutch trilinguals performed better on a lexical decision test when the (Dutch) target words had an English (L2) and French (L3) near-cognate translation equivalent (for instance *banaan–banana–banane*) than when the target words did not have cognate translation equivalents.

If we then apply this cognate facilitating effect to the field of translation, one might expect translators to benefit from the presence of form-similar words as well, as this allows

them to retrieve an accurate translation faster. However, compared to the vast body of research on cognates within psycholinguistics, considerably fewer studies have focused on the effect of cognates on translation. Shlesinger and Malkiel (2005) conducted an experiment which led to the conclusion that translators prefer a non-cognate translation over a cognate translation, even when both are equally good equivalents for the source word. Similarly, Malkiel (2009a) found that translators are hesitant in choosing a cognate translation, and tend to prefer non-cognate translations, thereby contradicting the cognate facilitation effect. This observation may be linked to the rather negative connotation attached to cognates in translation education, as they are associated with false friends (words that share the same form but differ in meaning) (Chamizo Dominguez and Nerlich 2002; Malkiel 2009a, 2009b; Yetkin 2011).

The findings from the above cited studies lead to our third hypothesis that translators will (consciously) resist the cognate facilitation effect and select the non-cognate option in order to avoid losing their credibility as language professionals.

2.4 Summing up

Based on the literature review above, the following three hypotheses are formulated for our empirical study:

1. An increase in complexity will lead to a higher rate of frequent and prototypical lexemes in translated language compared to non-translated language.
2. The most frequent lexeme of a certain genre in non-translated language will be overrepresented in translated language in order to minimise communicative risk.
3. The non-cognate translation option will be chosen more often in translated language (compared to non-translated language) in order to resist the cognate facilitation effect.

3. Methodology

In order to find out which onomasiological options for verbal inchoativity translators prefer in which contexts, and to what extent this preference differs from that of non-translators, and why, we opted for a corpus-based approach. The data for this study are drawn from the Dutch Parallel Corpus (DPC), a bidirectional, multi-genre, sentence-aligned, ten-million-word corpus for the language pairs Dutch–French and Dutch–English (Macken, de Clercq, and Paulussen 2011). We used the genre classification suggested by Delaere (2015), which consists of seven different genres, namely specialised communication, broad commercial texts, journalistic texts, instructive texts, political speeches, legal texts, and tourist information. For the present study, we only focused on the corpus components containing original, non-translated Dutch (which was translated to English) and translated Dutch (which was translated from English). The corpus components with the French data were thus not taken into account in our analyses.

In Section 3.1, we present the method that was used to select the near-synonyms expressing verbal inchoativity, and in Section 3.2, we show how the sentences containing these near-synonyms were enriched by means of the behavioural profile method. Finally, in Section 3.3, we present the statistical methods that were used to analyse the patterns in the data.

3.1 Lexeme selection

To retrieve near-synonyms in the semantic field of inchoativity in an objective and non-intuitive way, we used an extension of the semantic mirroring technique that was originally developed by Dyvik (1998, 2004). It is a corpus-based technique that uses back-and-forth translation to yield semantically related lexemes. The approach is based on the idea that "semantically closely related words ought to have strongly overlapping sets of translations"

(Dyvik 2004, 311) and was later extended by Vandevoorde (2016, 2020) and Vandevoorde et al. (2017). Dyvik starts from an initial lexeme *a* in Language A and extracts all its translations in Language B from a parallel and sentence-aligned corpus to arrive at a first set of translations that is called the first T-image of *a* in Language B. For the next step, the back-translations of that first T-image are looked up to obtain the Inverse T-image of *a* in Language A. Finally, the translations in Language B of the inverse T-image are queried again, resulting in the second T-image, consisting of lexemes that are semantically similar. Following Vandevoorde (2016), we used a minimal overlap criterion (every lexeme should be the translation of at least two source-language lexemes) in order to exclude peripheral lexemes. We set the cumulative frequency threshold to 75%: we took the sum of the most frequent translations until we obtained 75% of the translated data. Finally, we selected only verbs, since (1) they are believed to have a greater breadth of meaning than nouns, (2) their meaning depends more on the linguistic context in which they are used (van Hell and de Groot 1998), and (3) the behavioural profile method requires that all the corpus sentences are coded for the same set of variables (Divjak and Gries 2006; Jansegers, Vanderschueren, and Enghels 2015). Including more word classes would have made it impossible to meet all three requirements.

Previous research (see De Baets, Vandevoorde, and De Sutter 2020) reveals that the two most prototypical expressions of inchoativity in Dutch are *beginnen* 'to begin' and *starten* 'to start', and so we used these two verbs as the initial lexemes for the semantic mirroring technique. All the sentences containing *beginnen* or *starten* were extracted from the corpus and their respective translations in French and English were listed in the first T-image, containing *commencer* 'to begin', *débuter* 'to start', *entamer* 'to initiate', *to start*, and *to begin*. These were then used as the input for a new corpus query, to create the inverse T-image.[2] We again

[2] Note that the French part of the Dutch Parallel Corpus is just used in this step to enable the identification of all relevant verbal lexemes *in Dutch*. After this step, the French data are not taken into further account.

implemented a cumulative frequency threshold to exclude more peripheral translations and we obtained a final set of five lexemes of inchoativity in Dutch, namely *beginnen* 'to begin', *starten* 'to start', *aanvatten* 'to commence', *opstarten* 'to start up', and *van start gaan* 'to launch'. *Aanvatten* turned out to have a very low frequency in the corpus and for that reason, we omitted it from our dataset.

3.2 Behavioural profiles

We extracted all sentences in original Dutch and all sentences in Dutch translated from English that contained one of the four selected verbs of inchoativity. Then, a random sample was drawn (n = 644; see Table 1) which was subsequently annotated for twenty-three different contextual features (also called ID-tags; see Table 2) along the lines of the behavioural profile method (Divjak and Gries 2006, 2009).

INSERT TABLE 1 HERE

Table 1. Absolute frequency of the selected verbs of inchoativity used for this study

	Non-translated Dutch	**Dutch translated from English**	**Total**
beginnen 'to begin'	160	105	**265**
starten 'to start'	101	76	**177**
opstarten 'to start up'	99	30	**129**
van start gaan 'to launch'	54	19	**73**
Total	**414**	**230**	**644**

Taken together, these features represent the unique behavioural profile of each inchoative verb. Although the method was initially designed to capture word meaning (see De Baets,

Vandevoorde, and De Sutter 2020), it can also be used to investigate which features are strongly associated with a certain verb of inchoativity and, consequently, which of the features have the largest impact on onomasiological choices. Table 2 presents all features used in this study.

INSERT TABLE 2 HERE

Table 2: Overview of all ID-tags used in this study

	ID-tags	**Levels of the ID-tags**
Subject-related ID-tags	animacy	animate – inanimate
	concreteness	abstract – concrete
	concreteness level	high – medium – low
	concreteness rating	score 1–5[3]
	countability	countable – uncountable
	number	singular – plural – uncountable
	proper/common name	common name – proper name
	semantics	action – animate – artefact – concrother – dynamic – human – institute – non-dynamic – place – time – undetermined
Verb-related ID-tags	aspect	imperative – infinitive – imperfect – perfect
	mode	imperative – indicative – infinitive
	number	infinitive – singular – plural
	time	future – imperative – infinitive – past – present
	voice	active – passive
Object-related ID-tags	animacy	animate – inanimate – no object
	concreteness	abstract – concrete – no object
	concreteness level	high – medium – low – no object
	concreteness rating	score 1–5
	constituent	adverbial constituent – nominal constituent – sentence – no object
	countability	countable – uncountable – no object
	number	singular – plural – no object

[3] Based on research by Brysbaert et al. (2014); 5 different lists of 6000 words were rated by 75 participants.

	semantics	action – animate – artefact – concrother – dynamic – human – institute – non-dynamic – place – substance – time – undetermined – no object
	type	agent – direct object – indirect object – predicative adjunct – prepositional object – no object
	object2 type	indirect object – predicative adjunct – prepositional object – no object
Contextual ID-tags	temporal indication	duration – starting point – no temporal indication
	modified verb	modified verb – no modified verb
	modifying verb	modifying verb – no modifying verb
	clause type	main sentence – subordinate
	sentence length	very short – short – medium – long – very long
	cognate	cognate – no trigger – trigger ignored
Extra-linguistic ID-tags	genre	broad – fiction – instructive – journalistic – legal – political – special – tourism
	domain	communication – consumption – culture – economy – education – environment – finance – foreign affairs – history – home affairs – institutions – justice – leisure – science – transport – welfare state

In order to tease apart the three sources of explanation mentioned in Section 2, we assume that each of the ID-tags in Table 2 are indicators for one of the explanations.

3.2.1 *Complexity*

The following contrasts are taken as indexical of complexity, with the first-mentioned feature indicating a higher degree of complexity:

- passive vs. active constructions (Rohdenburg 1996; Gleitman et al. 2007);
- long vs. short sentences (Rohdenburg 1996; Arnold et al. 2000; Szmrecsányi 2004; Kruger 2012; Pijpops et al. 2018);
- the presence of an object vs. the absence of objects (Rohdenburg 1996);

- inanimate subjects and objects vs. animate subjects and objects (Rohdenburg 1996; Bonin, Gelin, and Bugaiska 2014);
- abstract subjects and objects vs. concrete subjects and objects (Walker and Hume 1999);
- the presence of a modified or modifying verb vs. the absence of another verb (Ferreira 1991).

As cognitive load is deemed to be higher during translation, we hypothesise that the features of increased complexity will play a more decisive role in influencing onomasiological choice in translated texts than in original texts. We also expect that the prototypical lexemes of inchoativity (namely *starten* 'to start' and *beginnen* 'to begin') will be overrepresented in more complex contexts, and that this effect will be even more striking in translated language.

3.2.2 *Risk aversion*

As in Kruger (2019) and Kruger and De Sutter (2018), we consider the feature *genre* the best indicator to measure the impact of risk aversion: we expect translators to be aware of genre-specific lexical norms, and hence we expect that in general *genre* will have the same effect on the onomasiological choice in translated and in non-translated language. In addition, we expect that translators will be prone to select the most frequent verbal lexeme within a specific genre in the target language.

3.2.3 *Cognate exposure*

Cognates can affect the onomasiological choice in two ways: they either trigger the selection of the cognate word in the target language (the so-called cognate facilitation effect) or they hamper the selection of the equivalent cognate (cognate avoidance). In order to code cognate exposure in Dutch translated from English, we start from the source sentences that correspond to each Dutch target sentence in our sample. Four types of relationships between the Dutch

inchoative verb and the English source sentence were identified. First, if the equivalent word or construction in English signifying inchoativity is not a cognate of one of the four selected Dutch verbs, we coded this as 'non-cognate'. In Example (1), the English source word *commence* is clearly not a cognate of *opstarten* 'to start up' in Dutch:

(1) In addition Boehringer Ingelheim will **commence** a joint research programme including Ablynx scientists. (ST – dpc-aby-002308)
Bovendien zal Boehringer Ingelheim een gezamenlijk onderzoeksprogramma ***opstarten*** *met wetenschappers van Ablynx.* (TT)
'In addition, Boehringer Ingelheim will start up a joint research programme with Ablynx scientists.'

Second, if there is no lexical equivalent of inchoativity in the source sentence, which is, for example, the case when inchoativity is expressed by a progressive structure (*to be* + *ing*-form), we coded this as 'noteq' (i.e., not (lexically) equivalent). Third, if the source-text word or construction is a cognate of one of the four Dutch verbs, but the translator chose another, non-cognate verb, as in Example (2), we coded this as 'cognate ignored'.

(2) This is a natural follow-up of a relationship that **started** 17 years ago when Interbrew transferred its know-how to the Zhujiang Brewery. (ST – dpc-bev-002082)
Dit is een natuurlijke voortzetting van een relatie die 17 jaar geleden werd ***begonnen****, toen Interbrew haar knowhow in de brouwerij Zhujiang inbracht.* (TT)
'This is a natural follow-up of a relationship that was begun 17 years ago when Interbrew imported its know-how in the Zhujiang brewery'

Finally, if the source-text word or construction is a cognate of one of the four Dutch verbs, and the translator chose the cognate verb in the Dutch target sentence, we coded this as 'cognate selected', as in Example (3).

(3) Deliveries will **start** in the last quarter of this year. (ST – dpc-bco-002446)

*De leveringen **starten** in het laatste trimester van dit jaar.* (TT)

'Deliveries start in the last trimester of this year.'

We expect the feature *cognate* to play an influential role, with translators resisting the cognate facilitation effect in order to avoid using a false cognate.

3.3 Statistical analysis

In order to determine which features prompt translators and non-translators to select *beginnen* 'to begin', *starten* 'to start', *opstarten* 'to start up', or *van start gaan* 'to launch' to express an inchoative activity, we used conditional random forest modelling. This method is a so-called ensemble learning method for classifying instances (in our case, inchoative verbs) using the features mentioned in Table 2. This method has been successfully used to discover both linguistic and extralinguistic features that determine the use of near-synonyms or alternating syntactic constructions (Baayen et al. 2013; Levshina 2020). It is especially useful in research designs where the number of predictors (in this case, ID-tags) is high compared to the number of observations. A second advantage is that random forests provide reliable results when predictors are intercorrelated, which is often the case in corpus-based research. We used the R-package 'partykit' (Hothorn, Hornik, and Zeileis 2006; Strobl et al. 2007), with 1500 runs, to perform the random forest modelling. We run two random forest analyses, one on the non-

translated data, and then a second one on the translated data. These two analyses will enable us to compare the impact of each of the features on the two varieties.

4. Results

This section is divided into two parts. First, in Section 4.1 we present an overview of the outcome of the random forest modelling for non-translated Dutch and Dutch translated from English; these will reveal which features (or ID-tags) guide the choice between the four onomasiological alternatives for verbal inchoativity in both varieties. In Section 4.2, we present an in-depth analysis of the exact nature of the most influential features in translated and non-translated Dutch.

4.1 Random forest models for non-translated and translated Dutch

Figure 1 depicts the relative importance of the features that play a role in the onomasiological choice made by the writers of original, non-translated texts. The features (ID-tags) in Figure 1 are ranked according to importance. The figure shows that *voice* is the most influential feature in the decision process between near-synonyms of inchoativity in original texts, followed by *genre*, the *semantic category of the subject*, *sentence length*, and the *type of object* that accompanies the inchoative verb. After those five features, there is a noticeable gap in the graph, which indicates that the remaining features have less influence on the onomasiological choice. We can thus conclude that a diverse set of features determine the choice of verb, including syntactic, semantic, and language-external features. Three complexity-related factors are ranked in the top four most influential features, and in the top ten we observe five

complexity-related features. Section 5.2 presents a more detailed analysis of the effect that each of these features has.

INSERT FIGURE 1 HERE

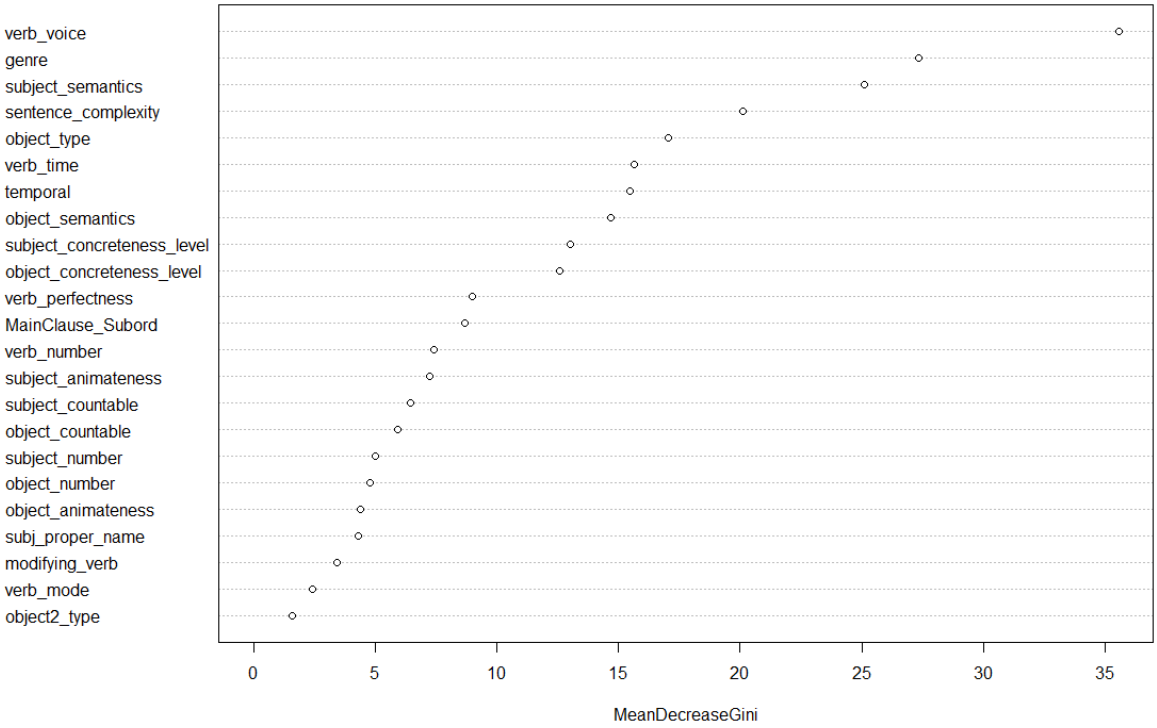

Figure 1. Results of the random forest analysis for non-translated Dutch. Features are ranked from highly important in classifying the lexemes (top) to less important (bottom). Importance is measured in terms of *mean decrease in Gini coefficient*.

Table 3 presents the overall classification accuracy of the random forest model for non-translated Dutch. In general, the model succeeds in classifying the inchoative verbs correctly in 60.39% of cases. In particular, the model performs relatively well for *beginnen* 'to begin' (classification accuracy 76.2%) and *opstarten* 'to start up' (classification accuracy 78%), but has more difficulty in correctly classifying *van start gaan* 'to launch' (classification accuracy 42.6%) and *starten* 'to start' (classification accuracy 27.7%). *Starten* 'to start' is more often

wrongly classified as *beginnen* ‘to begin’ (38.6% of its attestations) than it is correctly classified as *starten* ‘to start’ (27.7%).

INSERT TABLE 3 HERE

Table 3. Classification accuracy of the random forest model for non-translated Dutch

		Predicted cases				
		beginnen	*opstarten*	*starten*	*van start gaan*	**Classification accuracy**
Observed cases	*beginnen* ‘to begin’ (n = 160)	122	7	20	11	0.762
	opstarten ‘to start up’ (n = 99)	7	77	15	0	0.778
	starten ‘to start’ (n = 101)	39	25	28	9	0.277
	van start gaan ‘to launch’ (n = 54)	21	1	9	23	0.426

Figure 2 shows the relative importance of the features that play a role in English–Dutch translations. The same variety of syntactic, semantic, and language-external features are evident in the top ten most influential features, with relatively few differences compared to the random forest model of non-translated Dutch. *Genre* is more influential in translated Dutch, whereas the impact of the *voice* of the inchoative verb, which is the most strongly determining feature in non-translated Dutch, is only the ninth most influential feature in translated Dutch. Furthermore, the *semantic category of the subject* gains somewhat more influence in the decision process between near-synonyms in translation: in translated Dutch it is the second most influential feature, whereas in original Dutch, it is the third most influential feature. When

looking at the complexity-related features, we see that only two features are ranked in the top five most influential features in translated Dutch: the *type of the object* and the *complexity of the sentence*. Compared to non-translated Dutch, the *concreteness level of the subject* and the *object* are less important when deciding between near-synonymous verbs in translation. In non-translated Dutch, the *concreteness level of the object* is ranked as the tenth most influential feature; in translated Dutch, it is in twenty-second position.

INSERT FIGURE 2 HERE

Figure 2. Results of the random forest for translated Dutch. Features are ranked from highly important in classifying the lexemes (top) to less important (bottom). Importance is measured in terms of *mean decrease in Gini coefficient*.

A final interesting observation is the position of *Cognate*, the ID-tag that indicates whether the target lexeme is influenced by the presence of a cognate in the source text. In our dataset, there are three cognate pairs: *beginnen–to begin*, *starten–to start*, and *opstarten–to start up*. The presence of a cognate in the source text is the sixth most influential feature in translated Dutch, signifying that the presence of a cognate in the source text indeed plays a role in the onomasiological choice.

Table 4 presents the overall classification accuracy of the random forest model for translated Dutch. In general, the model succeeds in classifying the inchoative verbs correctly in 69.13% of cases, which is almost 10% better than the classification task in non-translated Dutch.

INSERT TABLE 4 HERE

Table 4. Classification accuracy of the random forest model for translated Dutch

		Predicted cases				
		beginnen	*opstarten*	*starten*	*van start gaan*	**Classification accuracy**
Observed cases	*beginnen* 'to begin' (n = 105)	87	2	15	1	0.829
	opstarten 'to start up' (n = 30)	5	6	19	0	0.200
	starten 'to start' (n = 76)	12	6	57	1	0.750
	van start gaan 'to launch' (n = 19)	7	0	3	9	0.473

The model succeeds very well in predicting *beginnen* 'to begin' and *starten* 'to start', with a classification accuracy of 83% and 75%, respectively. This is in sharp contrast with non-translated Dutch, where the accuracy score for *starten* 'to start' is 28%. The model trained on translated Dutch has more difficulties classifying the less frequent lexemes *opstarten* 'to start up' and *van start gaan* 'to launch' (classification accuracy is 20% and 47%, respectively). *Opstarten* 'to start up' is also the only verb that is more often classified incorrectly than correctly.

5.2 In-depth analysis of the most influential features

In this section, we compare the effect of the most influential features in translated and non-translated texts, namely *voice* and *genre*. *Voice* is the most influential feature for correctly classifying onomasiological choices in non-translated Dutch, but it is not as influential in Dutch translated from English. The detailed analysis in this section will show why that is. *Genre* is very influential in both translated and non-translated Dutch, but that does not mean that translators and non-translators always make the same onomasiological decisions in each of the genres; it is possible that the decisions made per genre are quite different in translated and non-translated Dutch. The in-depth analysis of *voice* (a complexity-related indicator) and *genre* (a risk-aversion-related indicator) will allow us to assess the impact of the complexity principle and risk aversion explanation in translated Dutch. In addition to these two features, we also discuss the presence of *cognates*, in order to assess the impact of cognate exposure in translated Dutch.

5.2.1 *Voice*

Table 5 shows the overall distribution of active and passive sentences containing a verb of inchoativity in translated and non-translated Dutch. We can observe that the active voice is more frequent in general, both in non-translated Dutch (77.3%) and in translated Dutch (85.2%). The difference in distribution of active and passive constructions between translated and non-translated Dutch is statistically significant ($\chi^2(1) = 5.83$, $p < 0.02$), and is in line with the complexity principle: in highly constrained, cognitively more complex situations, such as during translation, language users have a preference for the clause structure which is less complex, namely active constructions.

INSERT TABLE 5 HERE

Table 5. Distribution of verbs of inchoativity in this study in active and passive clauses in translated and non-translated Dutch

	Non-translated Dutch		**Translated Dutch**	
	n	%	*n*	%
Active	320	77.3	196	85.2
Passive	94	22.7	34	14.8
Total	414	100.0	230	100.0

Table 6, however, shows that this preference is also dependent on the specific inchoative verb. In non-translated language, we see that *opstarten* 'to start up' has a very strong preference for the passive voice (64.6%), whereas *beginnen* 'to begin' and *van start gaan* 'to launch' are rarely used in the passive voice (respectively, 3.1% and 3.7%); *starten* 'to start' is used in the passive voice in 23% of its attestations. These varying lexical preferences in non-translated Dutch are statistically significant ($\chi^2(3) = 145.29$, $p < 0.00001$). In translated Dutch, the overall picture is more balanced than in non-translated Dutch (hence the lower overall impact of *voice* in translated Dutch; see Section 4.1), with an overall preference for active constructions. In passive contexts, *opstarten* 'to start up' is still the preferred verb, but it also is less frequent than in non-translated Dutch (40% vs. 64.6%). The same applies to *beginnen* 'to begin': in 2.9% of its attestations, *beginnen* is used in the passive voice. This is slightly less than in non-translated Dutch, where 3.1% of the attestations were found in passive sentences. Another difference is that *starten* 'to start' is used somewhat more often in passive voice in translated than in non-translated Dutch (25% vs. 22.8%), and *van start gaan* 'to launch' is never used in the passive voice. Although the lexical preferences in translated Dutch do not differ as much as in non-translated Dutch, they still are statistically significant ($\chi^2(3) = 34.09$, $p < 0.00001$).

INSERT TABLE 6 HERE

Table 6. Distribution of *beginnen*, *starten*, *opstarten*, and *van start gaan* across active and passive clauses and in translated and non-translated Dutch

		Non-translated Dutch			**Translated Dutch**		
		Active	**Passive**	**Total**	**Active**	**Passive**	**Total**
beginnen	*n*	155	5	160	102	3	105
'to begin'	%	96.9	3.1	100.0	97.1	2.9	100.0
opstarten	*n*	35	64	99	18	12	30
'to start up'	%	35.4	64.6	100.0	60.0	40.0	100.0
starten	*n*	78	23	101	57	19	76
'to start'	%	77.2	22.8	100.0	75.0	25.0	100.0
van start gaan	*n*	52	2	54	19	0	19
'to launch'	%	96.3	3.7	100.0	100.0	0.0	100.0

The results in Table 6 provide some evidence for an explanation in terms of the complexity principle: in translated language, there is a general preference for the cognitively less complex active construction; furthermore, the only verb of inchoativity which has a strong preference for the passive construction in non-translated Dutch, *opstarten* 'to start up', is used more often in active voice in translated Dutch. This suggests that translators, under influence of the cognitively challenging task of mediating a message between two languages, have a strong inclination to use the form of voice which is less complex, even with verbs that usually occur in passive clauses. Additional evidence for the complexity principle in translation is the higher frequency of the prototypical verbs *beginnen* 'to begin' and *starten* 'to start' in passive constructions (52% in translated Dutch vs. 42% in non-translated Dutch) and, vice versa, the lower frequency of the two less prototypical lexemes *opstarten* 'to start up' and *van start gaan* 'to launch' in translated Dutch compared to non-translated Dutch.

5.2.2 *Genre*

The feature that has the largest impact on the onomasiological choice in translated Dutch and the second largest impact in non-translated Dutch is *genre*. Table 7 gives an overall view of the distribution of the data across different genres and text variety (translated vs. non-translated). In both varieties, most of the data in our sample are journalistic in nature (43.5% in non-translated Dutch, 30.9% in translated Dutch), closely followed by broad commercial texts (30.7% in non-translated Dutch, 28.3% in translated Dutch). Touristic texts represent the smallest part with only eight instances, and there are only ten instances of inchoative verbs in legal texts. In the discussion below, we exclude legal texts and touristic texts because of this data sparseness. Finally, it can be seen that the translated part of the dataset contains more specialised texts (22.6% in translated Dutch, and only 8% in non-translated Dutch).

INSERT TABLE 7 HERE

Table 7. Distribution of verbs of inchoativity in this study across genres and translated and non-translated Dutch

	Non-translated Dutch		**Translated Dutch**	
	n	%	*n*	%
Broad	127	30.7	65	28.3
Fiction	8	1.9	14	6.1
Instructive	12	2.9	19	8.3
Journalistic	180	43.5	71	30.9
Legal	9	2.2	1	0.4
Political	37	8.9	8	3.5
Specialised	33	8.0	52	22.6
Tourism	8	1.9%	0	0.0%
Total	414	100.0%	230	100.0%

As *genre* is the testbed for the risk aversion explanation, we will consider the lexical preferences within each genre. Our general assumption is that translators will conform to the typical lexical choices in each genre in non-translated Dutch. If that turns out to be the case, the risk aversion hypothesis is confirmed.

We start with broad commercial texts. *Beginnen* 'to begin' and *starten* 'to start' are used more frequently in translated Dutch (36.9% and 41.5%, respectively) compared to non-translated Dutch (20.5% and 26%, respectively), whereas *opstarten* 'to start up' and *van start gaan* 'to take off' are used more frequently in non-translated Dutch (34.6% and 18.9%, respectively) compared to translated Dutch (12.3% and 9.2%, respectively). In sum, translators tend to use the more prototypical verbs *beginnen* 'to begin' and *starten* 'to start' much more frequently in broad commercial texts, and their choices thus differ from the typical lexical preferences in non-translated broad commercial texts. Although the genre-specific risk aversion hypothesis can be rejected, one could also argue that a more general interpretation of the risk aversion hypothesis can still be confirmed, since the lexemes that are most frequent overall, most prototypical, and most accessible are preferred in translated texts.

The distribution of verbs of inchoativity in translated and non-translated fictional texts is very clear: the only verb that is used is *beginnen* 'to begin', so there is no difference whatsoever between translated and non-translated Dutch. We have to take into account, however, that there are relatively few instances in the fictional texts in our dataset (n = 22), but for the time being, one could cautiously conclude that translated fiction conforms to non-translated fiction, thereby confirming the risk aversion hypothesis.

The situation in instructive texts is quite different. The preferred verb in non-translated Dutch is *beginnen* 'to begin' (50%) and *starten* 'to start' in translated Dutch (73%). Furthermore, it is remarkable that there are very few instances of *beginnen* 'to begin' in translated Dutch (15.8%), and that the less prototypical verb *opstarten* 'to start up' is used in

10.5% of the instances in translated Dutch, but never in non-translated Dutch. The verb *van start gaan* 'to launch' is never used in instructional texts. In sum, we can refute the risk aversion hypothesis, since translated texts do not completely conform to the typical lexical preferences in non-translated Dutch, and they even contain more instances of the less prototypical verb *opstarten* 'to start up' than non-translated texts.

In journalistic translated texts too, there is no conformity to the genre-specific lexical norms: *beginnen* 'to begin' is used much more frequently in translations (71.8%) than in non-translations (53.5%), whereas all other verbs of inchoativity are used less often: *starten* 'to start' is used in 11.3% of the cases in translations compared to 20% in non-translations; *opstarten* 'to start up' is used in 7% of the cases in translations compared to 16.1% in non-translations; *van start gaan* 'to launch' is used both in translations and non-translations in approximately 10% of cases. As in broad commercial texts, translated journalistic texts show different lexical preferences than non-translated journalistic texts, hence refuting the genre-specific risk aversion explanation, but a more general interpretation of the risk aversion hypothesis can still be confirmed, since in translated texts there is a general preference for the lexeme that is the most frequent overall, most prototypical, and most accessible. This confirms the findings of Szymor (2018).

In political texts, the preferred verb in non-translated Dutch is *opstarten* 'to start up' (35.1%) and *beginnen* 'to begin' and *van start gaan* 'to launch' in translated Dutch (37.5%); *beginnen* 'to begin' and *van start gaan* 'to launch' account for only 24.3% and 21.6% of the cases in non-translated Dutch, and *opstarten* 'to start up' is not used at all in translated Dutch. *Starten* 'to start' is used in 18.9% of the cases in non-translated Dutch and 25% of the cases in translated Dutch. These findings lead us to conclude that the risk aversion explanation cannot be maintained for this genre, since in translated texts there is no conformity to the typical lexical

preferences in non-translated Dutch, and these texts contain even more instances of the less prototypical verb *van start gaan* 'to launch' than non-translated texts.

Finally, in specialised texts there is a strong preference for *beginnen* 'to begin' in non-translated Dutch (36.4%) and *opstarten* 'to start up' in translated Dutch (48.1%). *Beginnen* 'to begin' is used in only 19.2% of the cases in translated Dutch, and *opstarten* 'to start up' is used in 33.3% of the cases in non-translated Dutch. *Starten* 'to start' and *van start gaan* 'to launch' are equally frequent in both varieties. Once more, our findings suggest that neither the genre-specific interpretation of the risk aversion hypothesis nor the general interpretation can be maintained.

INSERT TABLE 8 HERE

Table 8. Proportional distribution (%) of *beginnen*, *starten*, *opstarten*, and *van start gaan* across genres and translated and non-translated Dutch

		Broad	**Fiction**	**Instructional**	**Journalistic**	**Political**	**Specialised**
beginnen 'to begin'	Non-translated (*n* = 160)	20.5	100	50.0	53.3	24.3	36.4
	Translated (*n* = 105)	36.9	100	15.8	71.8	37.5	19.2
opstarten 'to start up'	Non-translated (*n* = 99)	34.6	0.0	0.0	16.1	35.1	24.2
	Translated (*n* = 30)	12.3	0.0	10.5	7.0	0.0	26.9
starten 'to start'	Non-translated (*n* = 101)	26.0	0.0	50.0	20.0	18.9	33.3
	Translated (*n* = 76)	41.5	0.0	73.7	11.3	25.0	48.1

van start gaan 'to launch'	Non-translated (n = 54)	18.9	0.0	0.0	10.6	21.6	6.1
	Translated (n = 19)	9.2	0.0	0.0	9.9	37.5	5.8

In conclusion, there is relatively little evidence that risk aversion plays a major role in translating inchoative verbs from English into Dutch: The typical lexical preferences in the respective genres of the target language are mostly ignored in translated texts, and sometimes even non-prototypical verbs are more frequently selected in translated than in non-translated texts.

5.2.3 *Cognate exposure*

In Section 4.1, it was shown that the presence of cognates in the source text does affect onomasiological choices in translated texts (*cognate exposure* was found to be the sixth most influential feature). However, the random forest analysis cannot show whether the presence of a cognate in the source text stimulates the use of a cognate in the target language or hampers it. As mentioned in Section 3, we distinguish between four types of cognate influence: the English source sentence contains the cognate verb that is used in the Dutch target sentence; the English source sentence contains a potential cognate verb, but the Dutch target sentence opted for a non-cognate verb; the English source sentence does not contain a cognate verb; and the English source sentence uses morphological means to express inchoativity (e.g., *ing*-forms) or there is no inchoativity whatsoever. For the analysis below, we did not take the last possibility into account.

Table 9 summarises the exposure to cognate source-text verbs per selected Dutch verb of inchoativity. When we focus on the sentences without a cognate trigger in the source text (the second column in Table 9), it emerges that *beginnen* 'to begin' (31.5%) and *starten* 'to start' (32.4%) are the preferred verbs to express inchoativity, as one would expect from the two

most prototypical verbs. In contexts where a cognate verb of one of three Dutch verbs (*beginnen* 'to begin', *starten* 'to start', and *van start gaan* 'to launch') is available (the third and fourth columns in Table 9; *van start gaan* is the only selected verb of inchoativity that has no cognate in English), *beginnen* is chosen most often (61.4%) when the English cognate is present in the source sentence, and it is chosen in 38.6% of the contexts in which the cognate of another verb of inchoativity is present (viz. *to start* or *to start up*). *Starten* 'to start' is most frequently chosen when its English cognate is present in the source sentence (85%) and it is chosen in 15% of the cases when the cognate of another verb of inchoativity is present. Finally, *opstarten* 'to start up' is hardly chosen when its cognate is present in the source text (22.2%), and it is most frequently chosen when another cognate is available (77.8%).

INSERT TABLE 9 HERE

Table 9. Distribution of *beginnen*, *starten*, *opstarten*, and *van start gaan* in translated Dutch as a function of the absence versus the presence of a cognate in the source text

		No cognate	**Cognate selected**	**Cognate ignored**	**Total *n***
beginnen	*n*	35	43	27	105
'to begin'	%	31.5	61.4	38.6	
opstarten	*n*	21	2	7	30
'to start up'	%	18.9	22.2	77.8	
starten	*n*	36	34	6	76
'to start'	%	32.4	85	15	
van start gaan	*n*	19	0	0	19
'to launch'	%	17.1	0	0	

All in all, then, the findings presented above confirm the general preference in translated texts for the two prototypical verbs in contexts without a direct cognate trigger in the source sentence. Furthermore, the findings also show that cognate words are not avoided in translation,

especially when a verb form of *starten* ‘to start’ and – to a lesser extent – *beginnen* ‘to begin’ is available. This is in line with psycholinguistic observations on the cognate facilitation effect (cognate stimuli in the source text stimulate the use of the cognate translation). On the other hand, when looking at the results for *opstarten* ‘to start up’, it also becomes clear that translators choose the verb of inchoativity very consciously, as the cognate trigger in the source sentence is often ignored there (77.8%) and, in general, it is chosen more often in contexts without a direct cognate trigger.

5. Discussion and conclusion

In this article we aimed to investigate the effect of three well-known socio-cognitive mechanisms on the onomasiological choice between four verbal lexemes of inchoativity in translated Dutch, compared to non-translated Dutch: the complexity principle, risk aversion, and cognate exposure. In order to gain more insight into what drives the choice, we combined two methods, namely the behavioural profile approach and conditional random forest modelling. The results of the general analysis show that in both translated and non-translated language, a diverse set of features affect the onomasiological choice. Alongside some differences, we found a large number of similarities between the two varieties: the language-external feature *genre* is somewhat more important in translated Dutch (compared to non-translated Dutch), whereas the complexity-related feature *voice* is more important in non-translated Dutch. The model based on translated Dutch appears to perform better than the one based on the non-translated data; the model trained on translated Dutch is particularly good at classifying the two prototypical verbs of inchoativity *starten* ‘to start’ and *beginnen* ‘to begin’, whereas the model trained on non-translated data has more difficulties in classifying *starten* ‘to start’.

For the in-depth analysis, we focused on the two most influential features (*voice* and *genre*) and on the presence of cognates in the source text. We observed that the active voice is more frequently used in translated Dutch, which we interpret as evidence for the complexity principle. It also became clear that the preference for the active or passive voice is dependent on the inchoative verb, and that these lexical preferences differ somewhat in translated and non-translated Dutch. Both the overall preference for the active construction and the higher frequency of prototypical verbs in passive constructions is considered to be evidence supporting the hypothesis that translators are more strongly influenced by the complexity principle. This seems to suggest that the cognitively challenging task of translating drives translators to opt for less complex constructions and for more prototypical lexemes.

Genre is the feature with the largest impact on the onomasiological choice in translated Dutch. We expected translators to conform to the typical lexical choices in each genre. However, we found that in general, typical genre-specific lexemes are not selected as often in translated texts compared to non-translated texts. A more general interpretation of risk aversion might still hold, as translators do have an overall preference for using more prototypical verbs of inchoativity.

The last mechanism under scrutiny was the presence of cognates. The random forest analysis shows that the presence of cognates does play a role in the onomasiological choices of translators. Our detailed analysis makes it clear that in the absence of a direct cognate in the source text, the preferred verbs to express inchoativity are *beginnen* 'to begin' and *starten* 'to start'. When there is a cognate trigger in the source text, translators are not hesitant to use the cognate translation, especially when a verbal form of *beginnen* 'to begin' or *starten* 'to start' is available, and much less so when the English cognate of *opstarten* 'to start up' occurs in the source text. This seems to confirm that translators use the outcome of the cognate facilitation effect when translating the most prototypical verbs *beginnen* 'to begin' and *starten* 'to start',

but the observation that the direct translation of the cognate source word of less prototypical verbs is mostly *not* chosen (or is avoided), suggests that the influence of such a cognate facilitation effect is not an automatic process in translators, but a well-considered choice.

The overall picture that emerges from this study is remarkably similar to previous multivariate studies of linguistic features in translated texts (see Kotze [2020] for an overview):

1. In general, the distribution of linguistic features in translated texts is similar to the distribution in non-translated texts.
2. The number, nature, and hierarchy of influencing variables or features that guide language users and translators towards a specific linguistic feature is very similar as well (compare Figures 1 and 2 in this study).
3. In addition to these similarities, some recurring patterns of divergence emerge as well, presumably due to the specific socio-cognitive circumstances in which translators (and interpreters) work (cognitive bilingual switching and the linguistic, social, and cultural expectations of the target audience): linguistic features that are frequent and unmarked in non-translated texts are chosen even more frequently in translated texts; this is commonly referred to as *overuse* of dominant language patterns.
4. Not surprisingly, (lexical) cognates affect the choices of translators (although this effect is far from absolute) and some influencing variables have a stronger or weaker effect in translated texts compared to non-translated texts.

Empirical translation scholars have tried to explain these effect differences by a variety of explanatory mechanisms such as the complexity principle or risk aversion referred to in this study, but it is obvious, also from this study, that much more empirical work is needed to test these mechanisms in other research designs, to flesh them out, and to integrate them in an encompassing coherent theory.

Nevertheless, we hope to have shown that behavioural profiling and random forest modelling are relevant methodological tools for corpus-based translation studies, as they allow for the linguistic characterisation of multiple competing onomasiological alternatives at a very specific level, and measure the impact of the language-internal and -external features that guide the choice between different linguistic options, in whatever language combination. Admittedly, the annotation of twenty-three lexicogrammatical features for four onomasiological variants is very time-consuming and introduces, compared to many other corpus studies, an additional load of complexity in analysing the data and interpreting the resulting patterns. This became particularly clear in Section 4, while trying to make sense of the contradicting results for risk-averse behaviour: some results confirmed risk-averse behaviour among translators, other results refuted this, casting doubt on the methodological approach to risk aversion in this study and on the feasibility of studying risk aversion on the basis of product data alone. Nevertheless, introducing a substantial amount of data complexity in corpus-based research designs is something empirical translation scholars should not shy away from in their endeavour to more accurately understand the specific nature of translation products and their underlying socio-cognitive processes. It is only through intensive, collaborative, and incremental empirical work, in which for instance corpus methods are combined with other reliable empirical methodologies, that significant theoretical progress is to be expected.

Acknowledgments

We are grateful to two anonymous reviewers and the editors for valuable suggestions and comments.

References

Alves, Fabio, and José Luiz Gonçalves. 2010. “Relevance and Translation.” In *Handbook of Translation Studies: Volume 1*, edited by Yves Gambier and Luc van Doorslaer, 279–284. Amsterdam: John Benjamins.

Arnold, Jennifer E., Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. “Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering.” *Language* 76 (1): 28–55.

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova, and Tore Nesset. 2013. “Making Choices in Russian: Pros and Cons of Statistical Methods for Rival Forms.” In *Time and Space in Russian Temporal Expressions*, edited by Laura A. Janda, Stephen M. Dickey, and Tore Nesset, special issue of *Russian Linguistics* 37 (3): 253–291.

Balling, Laura Winther. 2013. “Reading Authentic Texts: What Counts as Cognate?” *Bilingualism: Language and Cognition* 16 (3): 637–653.

Baker, Mona. 1993. “Corpus Linguistics and Translation Studies – Implications and Applications.” In *Text and Technology: In Honour of John Sinclair*, edited by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 17–45. Amsterdam: John Benjamins.

Becher, Viktor. 2010. “Abandoning the Notion of ‘Translation-Inherent’ Explicitation: Against a Dogma of Translation Studies.” *Across Languages and Cultures* 11 (1): 1–28.

Bonin, Patrick, Margaux Gelin, and Aurélia Bugaiska. 2014. “Animates are Better Remembered than Inanimates: Further Evidence from Word and Picture Stimuli.” *Memory & Cognition* 42 (3): 370–382.

Blum-Kulka, Shoshana, and Eddie A. Levenston, 1983. “Universals of Lexical Simplification.” In *Strategies in Interlanguage Communication*, edited by Claus Faerch and Gabriele Kasper, 119–139. Place of publication: Publisher.

Brysbaert, Marc, Michaël Stevens, Simon de Deyne, Wouter Voorspoels, and Gert Storms. 2014. “Norms of Age of Acquisition and Concreteness for 30,000 Dutch Words.” *Acta Psychologica* 150: 80–84.

Carroll, Susanne E. 1992. “On Cognates.” *Second Language Research* 8 (2): 93–119.

Chamizo Domínguez, Pedro J., and Brigitte Nerlich. 2002. “False Friends: Their Origin and Semantics in Some Selected Languages.” *Journal of Pragmatics* 34 (12): 1833–1849.

Costa, Albert, Michele Miozzo, and Alfonso Caramazza. 1999. “Lexical Selection in Bilinguals: Do Words in the Bilingual’s Two Lexicons Compete for Selection?” *Journal of Memory and Language* 41 (3): 365–397.

Costa, Albert, Àngels Colomé, and Alfonso Caramazza. 2000. “Lexical Access in Speech Production: The Bilingual Case.” *Psicológica* 21 (2): 403–437.

Costa, Albert, Mikel Santesteban, and Agnès Caño. 2005. “On the Facilitatory Effects of Cognate Words in Bilingual Speech Production.” *Brain and Language* 94 (1): 94–103.

Cruse, David Alan. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

De Baets, Pauline, Lore Vandevoorde, and Gert dD Sutter. 2020. “On the Usefulness of Comparable and Parallel Corpora for Contrastive Linguistics: Testing the Semantic Stability Hypothesis.” In *New Approaches to Contrastive Linguistics: Empirical and Methodological Challenges*, edited by Renata Enghels, Bart Defrancq, and Marlies Jansegers, xxx–xxx. Berlin: Mouton De Gruyter.

Delaere, Isabelle. 2015. *Do Translations Walk the Line? Visually Exploring Translated and Non-Translated Texts in Search of Norm Conformity*. PhD diss. Ghent University.

Delaere, Isabelle, Gert De Sutter, and Koen Plevoets. 2012. “Is Translated Language More Standardized than Non-Translated Language? Using Profile-Based Correspondence Analysis for Measuring Linguistic Distances between Language Varieties.” *Target* 24 (2): 203–224.

Delaere, Isabelle, and Gert De Sutter. 2013. “Applying a Multidimensional, Register-Sensitive Approach to Visualize Normalization in Translated and Non-Translated Dutch.” *Belgian Journal of Linguistics* 27 (1): 43–60.

De Sutter, Gert, Isabelle Delaere, and Koen Plevoets. 2012. “Lexical Lectometry in Corpus-Based Translation Studies.” In *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research,* edited by Michael P. Oakes and Meng Ji, 325–345. Amsterdam: John Benjamins.

De Sutter, Gert, and Marie-Aude Lefer. 2020. “On the Need for a New Research Agenda for Corpus-Based Translation Studies: A Multi-Methodological, Multifactorial and Interdisciplinary Approach.” *Perspectives* 28 (1): 1–23.

Dijkstra, Ton, Jonathan Grainger, and Walter J. B. van Heuven. 1999. “Recognition of Cognates and Interlingual Homographs: The Neglected Role of Phonology.” *Journal of Memory and Language* 41 (4): 496–518.

Divjak, Dagmar. 2010. *Structuring the Lexicon: A Clustered Model for Near-Synonymy*. Berlin: Walter de Gruyter.

Divjak, Dagmar, and Stefan Th. Gries. 2006. “Ways of Trying in Russian: Clustering Behavioral Profiles.” *Corpus Linguistics and Linguistic Theory* 2 (1): 23–60.

Divjak, Dagmar, and Stefan Gries. 2009. “Corpus-Based Cognitive Semantics: A Contrastive Study of Phasal Verbs in English and Russian.” *Studies in Cognitive Corpus Linguistics*: 273–296.

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

Dyvik, Helge. 2004. “Translations as Semantic Mirrors: From Parallel Corpus to Wordnet.” In *Advances in Corpus Linguistics: Papers from the 23rd International Conference on*

English Language Research on Computerized Corpora (ICAME 23), Göteborg 22–26 May 2002, edited by Karin Aijmer and Bengt Altenberg, 309–326. Göteborg: Rodopi.Edmonds, Philip, and Graeme Hirst. 2002. "Near-synonymy and lexical choice." *Computational linguistics* 28 (2): 105–144.

Egan, Thomas. 2012. "Using Translation Corpora to Explore Synonymy and Polysemy." In *Aspects of Corpus Linguistics: Compilation, Annotation, Analysis*, edited by Signe Oksefjell, Jarle Ebeling, and Hilde Hasselgård, issue of *Studies in Variation, Contacts and Change in English* 12. https://varieng.helsinki.fi/series/volumes/12/

Ferreira, Fernanda. 1991. "Effects of Length and Syntactic Complexity on Initiation Times for Prepared Utterances." *Journal of Memory and Language* 30 (2): 210–233.

Ferreira, Victor S., and Gary S. Dell. 2000. "Effect of Ambiguity and Lexical Availability on Syntactic and Lexical Production." *Cognitive Psychology* 40 (4): 296–340.

Gleitman, Lila R., David January, Rebecca Nappa, and John C. Trueswell. 2007. "On the *Give* and *Take* between Event Apprehension and Utterance Formulation." *Journal of Memory and Language* 57 (4): 544–569.

Gollan, Tamar H., and Lori-Ann R. Acenas. 2004. "What is a TOT? Cognate and Translation Effects on Tip-of-the-Tongue States in Spanish–English and Tagalog–English Bilinguals." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30 (1): 246–269.

Gries, Stefan Th. 2018. "On Over-and Underuse in Learner Corpus Research and Multifactoriality in Corpus Linguistics More Generally." *Journal of Second Language Studies* 1 (2): 277–309.

Gries, Stefan Th., and Dagmar Divjak. 2009. "Behavioral Profiles: A Corpus-Based Approach to Cognitive Semantic Analysis." In *New Directions in Cognitive Linguistics*, edited by Vyvyan Evans and Stéphanie Pourcel, 57–75. Amsterdam: John Benjamins.

Halverson, Sandra L. 2003. “The Cognitive Basis of Translation Universals.” *Target* 15 (2): 197–241.

Halverson, Sandra. 2010. “Cognitive Translation Studies: Developments in Theory and Method.” In *Translation and Cognition*, edited by Gregory Shreve and Erik Angelone, 349–369. Amsterdam: John Benjamins.

Halverson, Sandra L. 2013. “Implications of Cognitive Linguistics for Translation Studies.” In *Cognitive Linguistics and Translation: Advances in Some Theoretical Models and Applications*, edited by Ana Rojo and Iraide Ibarretxe-Antuñano, 33–74. Berlin: Mouton de Gruyter.

Halverson, Sandra L. 2015. “Cognitive Translation Studies and the Merging of Empirical Paradigms: The Case of ‘Literal Translation.’” *Translation Spaces* 4 (2): 310–340.

Halverson, Sandra L. 2017. “Gravitational Pull in Translation: Testing a Revised Model.” In *Empirical Translation Studies: New Methodological and Theoretical Traditions,* edited by Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, 9–46. Berlin: Mouton de Gruyter.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. “Unbiased Recursive Partitioning: A Conditional Inference Framework.” *Journal of Computational and Graphical Statistics* 15 (3): 651–674.

House, Juliane. 2008. “Beyond Intervention: Universals in Translation?” *Trans-kom* 1 (1): 6–19.

House, Juliane. 2013. “Towards a New Linguistic-Cognitive Orientation in Translation Studies.” *Target* 25 (1): 46–60.

Immonen, Sini. 2006. “Translation as a Writing Process: Pauses in Translation Versus Monolingual Text Production.” *Target* 18 (2): 313–336.

Jansegers, Marlies, Clara Vanderschueren, and Renata Enghels. 2015. “The Polysemy of the Spanish Verb *Sentir*: A Behavioral Profile Analysis.” *Cognitive Linguistics* 26 (3): 381–421.

Kotze, H. 2020. “Converging *what* and *how* to find out *why*: An outlook on empirical translation studies”. In *New Empirical Perspectives on Translation and Interpreting*, edited by Lore Vandevoorde, Joke Daems and Bart Defrancq, 333-371. Vancouver: Routledge.

Kotze, Haidee. 2022. “Translation as constrained communication: Principles, concepts and methods.” In *Extending the Scope of Corpus-based Translation Studies*, edited by Sylviane Granger and Marie-Aude Lefer, 67-98. London: Bloomsbury.

Kroll, Judith, F. Dietz, and David Green. 2000. “Language Switch Costs in Bilingual Picture Naming and Translation.” In *Abstracts of the XXVII International Congress of Psychology*, edited by R; Sanchez-Casas: 405.

Kruger, Haidee. 2012. “A Corpus-Based Study of the Mediation Effect in Translated and Edited Language.” *Target* 24 (2): 355–388.

Kruger, Haidee. 2019. “*That* Again: A Multivariate Analysis of the Factors Conditioning Syntactic Explicitness in Translated English.” *Across Languages and Cultures* 20 (1): 1–33.

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

Kruger, Haidee, and Gert De Sutter. 2018. “Alternations in Contact and Non-Contact Varieties: Reconceptualising *That*-Omission in Translated and Non-Translated English Using the MuPDAR Approach.” *Translation, Cognition and Behavior* 1 (2): 251–290.

Levshina, Natalia. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis.* Amsterdam: John Benjamins.

Levshina, Natalia. 2020. “Conditional Inference Trees and Random Forests.” In *A Practical Handbook of Corpus Linguistics*, edited by Magali Paquot and Stefan Th. Gries, 611–643. New York: Springer.

Macken, Lieve, Orphée de Clercq, and Hans Paulussen. 2011. “Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus.” *Meta* 56 (2): 374–390.

Malkiel, Brenda. 2009a. “When *Idioti* (Idiotic) Becomes ‘Fluffy’: Translation Students and the Avoidance of Target-Language Cognates.” *Meta* 54 (2): 309–325.

Malkiel, Brenda. 2009b. “Translation as a Decision Process: Evidence from Cognates.” *Babel* 55 (3): 228–243.

Marín, Rafael, and Louise McNally. 2011. “Inchoativity, Change of State, and Telicity: Evidence from Spanish Reflexive Psychological Verbs.” *Natural Language & Linguistic Theory* 29 (2): 467–502.

Olohan, Maeve, and Mona Baker. 2000. “Reporting *That* in Translated English: Evidence for Subconscious Processes of Explicitation?” *Across Languages and Cultures* 1 (2): 141–158.

Olohan, Maeve. 2003. “How Frequent are the Contractions? A Study of Contracted Forms in the Translational English Corpus.” *Target* 15 (1): 59–89.

Pijpops, Dirk, Dirk Speelman, Stefan Grondelaers, and Freek van de Velde. 2018. “Comparing Explanations for the Complexity Principle: Evidence from Argument Realization.” *Language and Cognition* 10 (3): 514–543.

Piñón, Christopher. 2001. “A Finer Look at the Causative-Inchoative Alternation.” In *Proceedings of SALT 11*, edited by Rachel Hastings, Brendan Jackson, and Zsofia Zvolenszky, special issue of *Semantics and Linguistic Theory* 11: 346–364.

Pym, Anthony. 2005. “Explaining Explicitation.” In *New Trends in Translation Studies: In Honour of Kinga Klaudy*, edited by Krisztina Károly and Ágota Fóris, 29–34. Budapest: Akadémiai Kiadó.

Pym, Anthony. 2008. “On Toury’s Laws of How Translators Translate.” In *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury*, edited by Anthony Pym, Miriam Schlesinger, and Daniel Simeoni, 311–328. Amsterdam: John Benjamins.

Pym, Anthony. 2015. “Translating as Risk Management.” *Journal of Pragmatics* 85: 67–80.

Rohdenburg, Günter. 1996. “Cognitive Complexity and Increased Grammatical Explicitness in English.” *Cognitive Linguistics* 7 (2): 149–182.

Rohdenburg, Günter. 2016. “Testing two processing principles with respect to the extraction of elements out of complement clauses in English.” *English Language & Linguistics* 20 (3): 463–486.

Saridakis, Ioannis E. 2015. “Probabilistic Laws and Risk Aversion in Translation: A Case Study in Translation Didactics.” *Current Trends in Translation Teaching and Learning E (CTTL E)* 2: 196–245.

Schepens, Job, Ton Dijkstra, and Franc Grootjen. 2012. “Distributions of Cognates in Europe as Based on Levenshtein Distance.” *Bilingualism: Language and Cognition* 15 (1): 157–166.

Seeber, Kilian G. 2013. “Cognitive Load in Simultaneous Interpreting: Measures and Methods.” In *Interdisciplinarity in Translation and Interpreting Process Research*, edited by Maureen Ehrensberger-Dow, Susanne Göpferich, and Sharon O’Brien, special issue of *Target* 25 (1): 18–32.

Sherkina, Miriam. 2004. “The Cognate Facilitation Effect in Bilingual Speech Processing: The Case of Russian-English Bilingualism.” In *Proceedings of the Fourth Annual*

Meeting of the Niagara Linguistic Society, edited by Michael Barrie, Mohammad Haji-Abdolhosseini, Nick Pendar, and Jonathon Herd, special issue of *Cahiers linguistics d'Ottawa* 32: 108–121.

Shlesinger, Miriam, and Brenda Malkiel. 2005. "Comparing Modalities: Cognates as a Case in Point." *Across Languages and Cultures* 6 (2): 173–193.

Shi, Ziqiang. 1990. "On the Inherent Aspectual Properties of NPs, Verbs, Sentences and the Decomposition of Perfectivity and Inchoativity." *Word* 41 (1): 47–67.

Steiner, Erich. 1997. "Systemic Functional Linguistics and its Application to Foreign Language Teaching." *Estudios de Lingüística Aplicada* 15 (26): 15–27.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics* 8. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-25

Szymor, Nina. 2015. "Behavioral Profiling in Translation Studies." *trans-kom* 8 (2): 483–498.

Szymor, Nina. 2018. "Translation: Universals or Cognition? A Usage-Based Perspective." *Target* 30 (1): 53–86.

Van Assche, Eva, Wouter Duyck, Robert J. Hartsuiker, and Kevin Diependaele. 2009. "Does Bilingualism Change Native-Language Reading? Cognate Effects in a Sentence Context." *Psychological Science* 20 (8): 923–927.

Van Beveren, Amélie, Gert De Sutter, and Timothy Colleman. 2018. "Questioning explicitation in translation studies: A multifactorial corpus investigation of the *om*-alternation in translated and original Dutch." Paper presented at *Using Corpora in Contrastive and Translation Studies*, Louvain-La-Neuve, September 2018.

Van Hell, Janet G., and Annette M. B. de Groot. 1998. “Conceptual Representation in Bilingual Memory: Effects of Concreteness and Cognate Status in Word Association.” *Bilingualism: Language and Cognition* 1 (3): 193–211.

Van Hell, Janet G., and Ton Dijkstra. 2002. “Foreign Language Knowledge Can Influence Native Language Performance in Exclusively Native Contexts.” *Psychonomic Bulletin & Review* 9 (4): 780–789.

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

Vandevoorde, Lore. 2020. *Semantic Differences in Translation: Exploring the Field of Inchoativity*. Berlin: Language Sciences Press.

Vandevoorde, Lore, Els Lefever, Koen Plevoets, and Gert De Sutter. 2017. “A Corpus-Based Study of Semantic Differences in Translation: The Case of Dutch Inchoativity.” *Target* 29 (3): 388–415.

Verroens, Filip. 2011. *La construction inchoative se mettre à: syntaxe, sémantique et grammaticalisation* [The inchoative construction ‘se mettre à’: syntax, semantics and grammaticalisation]. PhD diss. Ghent University.

Walker, Ian, and Charles Hulme. 1999. “Concrete Words Are Easier to Recall than Abstract Words: Evidence for a Semantic Contribution to Short-Term Serial Recall.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25 (5): 1256–1271.

Yetkin, Nihal. 2011. “Partial False Friends in English–Turkish Translations: Diplomatic Texts.” *Hacettepe University Journal of Faculty of Letters* 28 (1): 207–222.

Address for correspondence

Pauline De Baets

Ghent University

Abdisstraat 1

B-9000 Ghent

Belgium

pauline.debaets@ugent.be

Co-author information

Gert De Sutter

Ghent University

gert.desutter@ugent.be