Definiteness across Languages: from German to Mandarin

David Bremmers, Jianan Liu, Martijn van der Klis and Bert Le Bruyn[1]
Utrecht University
d.j.e.bremmers@students.uu.nl, j.liu4@uu.nl,
m.h.vanderklis@uu.nl, b.s.w.lebruyn@uu.nl

Abstract

We showcase the potential of a data-driven methodology for cross-linguistic research: *Translation Mining*. We introduce the technique and put it to use in the domain of definiteness. We show how Translation Mining confirms existing insights about definiteness in English, German (Schwarz 2009) and Mandarin (Jenks 2018) while at the same time leading to novel insights for Mandarin.

1 Introduction

When approaching a phenomenon α from a language β for the first time, linguists rely on the knowledge about α that was built up based on other languages. This is the standard way of doing cross-linguistic research: we interpret data against existing accounts and – where necessary – extend these or develop new ones.

In this paper, we showcase the potential of a more data-driven methodology: *Translation Mining* (henceforth *TM*). We present the technique and put it to use in the domain of definiteness. In our presentation of the technique, we compare it to the semantic map method (e.g. Haspelmath 1997). The semantic map method is an example of the standard approach sketched above but also shares some features with TM and provides for an insightful comparison.

The paper is organized as follows. Section 1 introduces TM, Section 2 provides a quick review of recent work on definiteness in the languages under consideration and Section 3 presents our application of TM. Section 4 discusses and concludes.

2 From Semantic Maps to Translation Mining

TM can be considered the data-driven variant of semantic maps. We introduce semantic maps (2.1) and then turn to TM (2.2).

2.1 Semantic maps

A semantic map is 'a geometrical representation of functions in "conceptual/semantic space" which are linked by connecting lines' (Haspelmath 2003). Semantic maps are mainly found in

[1]David covered the German data collection and analysis, Jianan the Mandarin one. Martijn designed the Translation Mining interface and contributed to the methodological part of this paper. Bert supervised David and Jianan, brought together the different parts and worked them into the current paper. We thank the members of the Time in Translation project for valuable feedback and gratefully acknowledge financial support of NWO (grant 360-80-070).

typological work: 'the map describes and constrains languages that venture their grammars and/or lexicons into this space, both with respect to diachrony and synchrony.'(Van der Auwera & Plungian 1998). An example of a semantic map is given in Figure 1. We first look into the research that underlies it and then turn to its interpretation.

Semantic maps are developed for specific domains. The map in Figure 1 was developed by Haspelmath (2003) for datives. Researchers start from existing analyses of the domain and cast the widest possible net to identify as many functions as possible. E.g., in the domain of datives, these would include functions like *recipient*, *beneficiary*, etc.

The next step is to establish whether these functions should indeed be distinguished. The criterion for this is empirical: a function is 'put on the map if there is at least one pair of languages that differ with respect to this function' (Haspelmath 2003). An example is the distinction English makes between 'give something *to* someone' and 'buy something *for* someone': many languages allow their dative to appear in both contexts but the fact that English makes a formal distinction between the two warrants the inclusion in the map of the functions *recipient* and *beneficiary*.

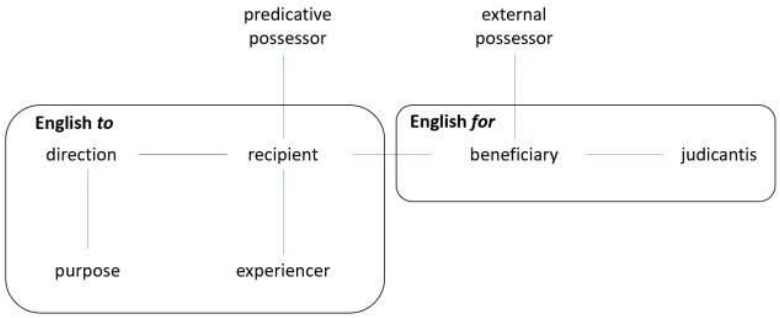

Figure 1: A semantic map of typical dative functions / boundaries of English *to* and *for* (Haspelmath 2003)

Turning to the interpretation of a map as in Figure 1, we find functions connected by lines, and boundaries of specific lexical items/constructions:

Connecting lines are added between two functions – as for *direction* and *purpose* – to indicate that there are lexical items/constructions that combine these functions. The absence of a connecting line between two functions – as for *purpose* and *experiencer* – entails that no lexical items/constructions combine these functions without also conveying the functions that connect them – *direction* and *recipient* in this particular case. The exact geometrical lay-out of the map (e.g. the fact that *direction* and *purpose* are arranged on the vertical axis) does not reflect any claim and in general depends on the graphical creativity of the researcher.

For concreteness, we have added the boundaries of English *to* and *for* but in principle we should be able to add any lexical item/construction that conveys one of the functions that is included in the map. The claim the map makes is that these lexical items/constructions convey functions that are connected on the map, both in diachrony and synchrony. Herein lies the predictive power of semantic maps.

2.2 Translation Mining

TM can be considered the data-driven variant of semantic maps.[2] We briefly compare the two at the levels of data collection and analysis.

Data collection

As we indicated above, classical work in semantic maps relies on existing literature to establish which functions might be relevant for a given domain. The empirical contribution of a semantic map then lies in establishing which functions should indeed be included and how they should be arranged.

The data collection in TM is different. To illustrate, let us switch from the dative domain to the domain of definiteness. Where classical work in semantic maps would look into existing literature and focus on what has been said about definiteness in different languages, TM would aim for neutrality as to what has been claimed before. The way to achieve this is to start from a marker of definiteness – say English *the* – and to select all contexts that contain this marker in the English version of a parallel corpus (i.e. a corpus containing texts and their translations into different languages). The next step is to establish how *the* in these contexts is rendered in the other languages and to repeat the whole procedure for all the equivalents of *the* that are found in the different languages. The output of this data collection is a set of datapoints consisting of contexts with the lexical items/constructions that are used in the different languages of the corpus. An example is given in (1):

(1) 'I'm not having one **in the house**, Petunia!' English: *the*, German: *contracted definite*, Mandarin: *demonstrative*

This datapoint involves the use of a definite in English (*in the house*), a definite that contracts with the preceding preposition in German (*im Haus* as opposed to *in dem Haus*) and a proximal demonstrative in Mandarin (*zhè* CL *fángzi*).

Analysis

The analysis in TM starts with a graphical representation of the data, known as a map. We first present TM maps and then compare them to classical semantic maps.

One of the maps that will come back in this paper is the following:

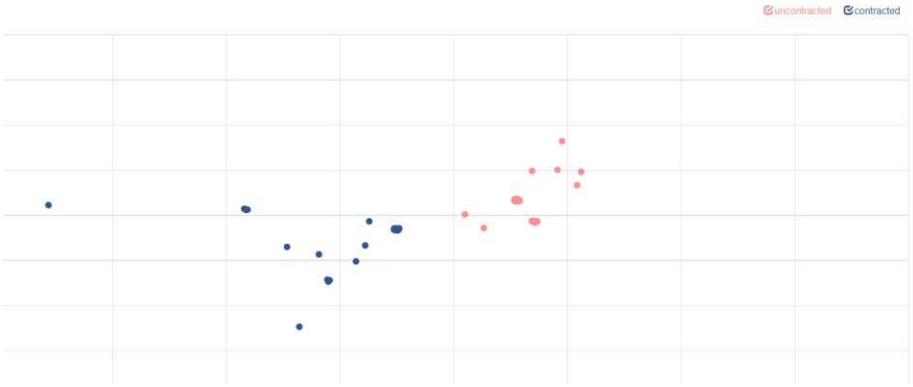

Figure 2: A TM map of definiteness / distribution of German lexical items/constructions

[2]Variants of core components of TM can – among others – be found in Wälchli & Cysouw (2012) and Beekhuizen et al. (2017). For a comparison between classical semantic maps and work like that of Wälchli & Cysouw (2012), see also Georgakopoulos & Polis (2018). See van der Klis et al. (2017, 2019) for an application of TM to the semantic domain of the (*have*)-Perfect.

This map is based on a subset of data that were collected along the lines set out above. We get back to the selection criteria in Section 3. For now, we note that we have used data from English, German and Mandarin and comment on what the map represents and how it has been generated.

Every dot on the map stands for a datapoint like (1). The organization of the dots on the map is created through Multi-Dimensional Scaling (MDS). Intuitively speaking, this algorithm groups contexts that use the same form in a given language. By doing this in parallel for the three languages, a single cross-linguistic pattern of groupings obtains.

With language specific color schemes, TM visualizes variation across languages. A comparison between the English map in Figure 3 and the German one in Figure 2 allows us to establish that the formal distinction between contracted and uncontracted definites has no formal equivalent in English. This is not immediately clear from the static version of the maps: the purple dots in the English map make up very tight clusters and this makes it hard to establish that they outnumber the other forms.

Figure 3: A TM map of definiteness / distribution of English lexical items/constructions

The maps created in TM are however interactive and allow us to zoom in and establish that out of the 96 datapoints on the map, 80 involve the definite article, making it by far the most frequent equivalent of both contracted (30 out of 40) and uncontracted definites (50 out of 56).

Now that we have established how TM maps are generated and what they represent, we can compare them to classical semantic maps. The best way to do so is to think of groupings of dots in TM maps as functions. The classical semantic map fragment that can be derived from Figures 2 and 3 would then look as in Figure 4:

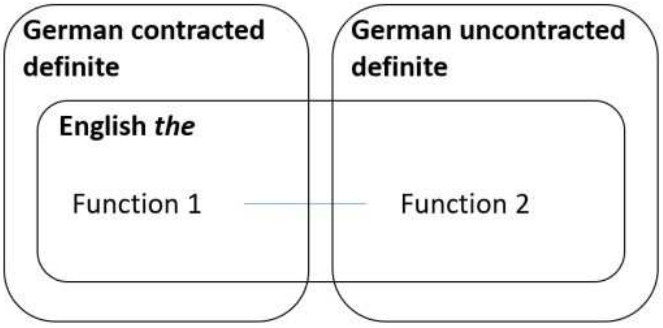

Figure 4: A classical semantic map fragment based on the TM maps of definiteness for German and English

Figure 4 distinguishes two functions. German makes a formal distinction between the two using contracted and uncontracted definites whereas English *the* covers both functions. The crucial point to be made is that no interpretive labels are given to the functions, unlike what we saw for the classical semantic map in Figure 1. Indeed, TM maps are different from classical semantic maps in that they merely record formal distinctions between groups of contexts. Under the assumption that formal distinctions reflect functional ones, TM maps then invite the researcher to investigate differences between groups of contexts and establish which functional distinctions underlie the formal ones. In the classical map method, these two steps are not strictly separated: formal distinctions are used to confirm the relevance of functional ones, not to discover them. In this sense, TM maps and the TM method count as the data-driven variant of classical semantic maps and the classical semantic map method.

3 Definiteness across languages

Putting TM to work for the analysis of a full semantic domain like definiteness across a typologically balanced sample of languages is – at the moment – utopian. We restrict our enterprise in two ways. The first is to limit ourselves to three languages: German, Mandarin and English. The second is to only look at a subset of contexts, *viz.* definites that occur in PPs in German. To justify these restrictions, we first need to give a brief sketch of the literature on definiteness.

The literature typically distinguishes between two types of definiteness, one based on uniqueness, the other on anaphoricity (Russell 1905; Strawson 1950; Kamp 1981; Heim 1982). English is believed to combine the two types in a single lexical item (*the*). German, on the other hand, has been claimed to formally distinguish between the two in the prepositional domain: Schwarz (2009) claims that uniqueness definites can contract with a subset of prepositions whereas anaphoric definites resist contraction. Examples from Schwarz are given in (2) and (3):

(2) Der Empfang wurde **vom/#von dem Bürgermeister** eröffnet.
'The reception was opened **by the mayor**.'

(3) In der New Yorker Bibliothek gibt es ein Buch über Topinambur. Neulich war ich dort und habe **#im / in dem Buch** nach einer Antwort auf die Frage gesucht, ob man Topinambur grillen kann.
'In the New York public library, there is a book about topinambur. Recently, I was there and searched **in the book** for an answer to the question of whether one can grill topinambur.'

The mayor in (2) has not been introduced before but is the unique mayor of the contextually salient town. This is a uniqueness context and the definite contracts with the preposition. In (3), a book is introduced in the first sentence and referred back to in the second. The second sentence constitutes a familiarity context and contraction is not allowed.

The contraction facts from German led Schwarz to coin the terms *weak* and *strong* definiteness. The first applies to contracted (uniqueness) definites, the latter to uncontracted (anaphoric) ones. We adopt this terminology here as it has become standard in studies looking at definiteness distinctions in typologically diverse languages (see Aguilar-Guevara et al. 2019). One such language is Mandarin.
Jenks (2018) argues that Mandarin makes a formal distinction between weak and strong definiteness in its use of bare nouns and demonstratives (examples from Jenks 2018):[3]

[3]This distinction disappears in subject position. See Jenks (2018) for discussion.

(2) a. (#Na/Zhe ge) Taiwan (de) zongtong hen shengqi.
DEM/DEM CL Taiwan DE president very angry
'The president of Taiwan is very angry.'
b. Jiaoshi li zuo-zhe yi ge nansheng he yi ge nüsheng. Wo zuotian yudao
classroom in sit-ASP one CL boy and one CL girl I yesterday meet
#(na ge) nansheng.
DEM CL boy
'There are a boy and a girl sitting in the classroom. I met the boy yesterday.'

(2) shows demonstratives are proscribed in uniqueness contexts (2a) but are mandatory in anaphoric contexts (2b). Bare nouns come out as the mirror image of demonstratives: they are proscribed in anaphoric contexts and mandatory in uniqueness contexts.

With the preliminary data and intuitions about definiteness in German, Mandarin and English in place, we can return to the restrictions we introduced above. To showcase the potential of TM, we need a semantic domain with a solid theoretical basis and a literature that has used this basis to explore cross-linguistic variation. Definiteness qualifies with the literature on English definiteness going back over a century. The recent literature on German and Mandarin further neatly illustrates how new data are approached from existing insights: the distinction between uniqueness and anaphoric definiteness has made its way into the cross-linguistic literature under the label of the weak/strong distinction. The restriction to definites occurring in PPs in German is inspired by Schwarz, who suggests that PPs are the only place where we can find form variation in German.[4] This restriction allows for a more focused data collection.

4 Definiteness in Translation Mining

In this section, we put TM to work for an analysis of the semantic domain of definiteness with the two restrictions we introduced in Section 3: we focus on German, Mandarin and English and restrict our attention to contexts that are rendered with a definite in German PPs. We introduce our corpus, briefly comment on the data collection and then present our results.

4.1 The corpus

The corpus we selected is the first volume of the Harry Potter series and its translations to German and Mandarin. An important asset of this corpus is that the source text is English. Given that English does not make a formal distinction between different types of definiteness, the formal distinctions we find in the German and Mandarin translations are independent from each other and translation biases are kept to a minimum. Other assets of this corpus are its recency and the availability of many other languages.

[4]We independently checked this suggestion through the study of referential expressions in German, English, Dutch and French. We used TM as our methodology and the first three chapters of the novel *L'Étranger* and its translations as our parallel corpus. We found that there is indeed little variation in the expression of definiteness in these languages. Based on Löbner (2011) we had expected some cross-linguistic variation between the definite article and demonstratives but we failed to establish this variation.

4.2 The data collection

Our first step was to extract PPs in German and specifically those in which the preposition contracted with the definite following it or in which the preposition could have contracted. For the uncontracted forms we selected all PPs in the novel. For the contracted forms we limited ourselves to forms in the first three chapters except for those contracted PPs that had an uncontracted counterpart with the same noun. In the latter case we extracted all occurrences in the novel. The goal of our selection procedure was to end up with a dataset with a more or less even distribution of contracted and uncontracted PPs, while at the same time maximizing the likelihood of including minimal pairs.

Once the set of German PPs established, we aligned them with the English original and the Mandarin translation. Alignment was done by two of the authors, one a native speaker of German and the other a native speaker of Mandarin. The number of contexts we ended up with for German-English-Mandarin amounts to 96.

4.3 Results

We discuss the results on the basis of TM maps. We start with German, then move to English and end with Mandarin.

German

The TM map with the color scheme for German was introduced above as Figure 2. Unsurprisingly, we find clear-cut groups of contracted and uncontracted forms. There are 40 contracted cases and 56 uncontracted ones.

Given that TM maps merely record formal distinctions, the map in Figure 2 does not allow us to confirm or refute Schwarz's claim that the groups oppose weak/uniqueness and strong/anaphoric definiteness. With the interactive TM interface we can however zoom in on the contexts themselves and confirm the predictions an analysis along the lines of the weak/strong distinction makes.

(3) **E:** 'I suppose we could take him to the zoo,' said Aunt Petunia slowly, '... and leave him **in the car** ...'
G: 'Ich denke, wir könnten ihn in den Zoo mitnehmen', sagte Tante Petunia langsam, '... und ihn **im Wagen** lassen ...'

(4) [Context: 'As the owls flooded into the Great Hall as usual, everyone's attention was caught at once by a long thin package carried by six large screech owls. Harry was just as interested as everyone else to see what was in this large parcel and was amazed when the owls soared down and dropped it right in front of him, knocking his bacon to the floor.']
E: They had hardly fluttered out of the way when another owl dropped a letter **on top of the parcel**.
G: Sie waren kaum aus dem Weg geflattert, als eine andere Eule einen Brief **auf das Paket** warf.

The car in (3) does not refer back to a previously introduced car but to the unique family car. It consequently counts as a weak definite. The parcel in (4) refers back to the package that was introduced before and counts as a strong definite. As predicted by Schwarz, German relies on a contracted definite in (3) and an uncontracted definite in (4).

In the case of German, TM does not lead to the discovery of new semantic insights in the typology of definiteness but does allow us to check the basic predictions existing analyses

make. It furthermore allows us to establish the opposition between contracted and uncontracted forms in German as a baseline for studying weak/uniqueness and strong/anaphoric contexts in our corpus.[5]

English

The TM map with the color scheme for English was introduced above as Figure 3. The main equivalents of the German definites are the definite article (N=80), the bare singular (N=5) and the demonstrative (N=4). In line with the literature, the comparison with the German TM map shows that the definite article can be found both in weak and strong contexts. The distribution of bare singulars is clearly on the side of the weak/uniqueness definites. This is also in line with the literature that has often hinted at the complementary distribution of bare nouns and weak definites (e.g. Carlson & Sussman 2005; Aguilar-Guevara & Zwarts 2010). The distribution of demonstratives deserves closer scrutiny but the fact that they allow for a deictic and an anaphoric use might explain their appearance in weak and strong contexts.

For English, the contribution of TM is similar to that in German. We do not find new semantic insights. The technique however does allow us to confirm claims in the literature. This reinforces both these claims and the value of the technique.

Mandarin

The TM map with the color scheme for Mandarin is given below as Figure 5. The main equivalents of the German definites are bare nouns (N=79) and demonstratives (N=13).

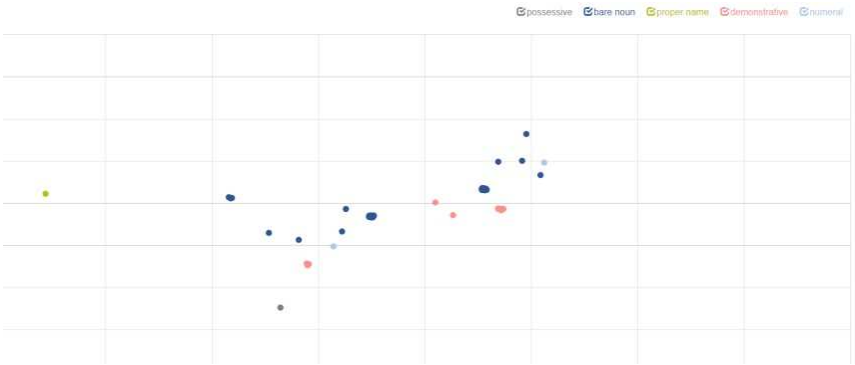

Figure 5: A TM map of definiteness / definiteness / distribution of Mandarin lexical items/constructions

The comparison with the German TM map reveals that the distribution of bare nouns and demonstratives in Mandarin is unexpected on Jenks' analysis. Jenks predicts (i) Mandarin bare nouns to appear in the same contexts as German contracted forms, and (ii) Mandarin demonstratives to appear in the same contexts as German uncontracted forms.

In Figure 6 we compare the TM maps of German and Mandarin side by side. The colors are a visualization of the predictions Jenks makes. Blue covers contexts with contracted

[5]We refer to Bremmers (2019) for an in-depth discussion of the German facts.

forms in German and bare nouns in Mandarin. Red covers contexts with uncontracted forms in German and demonstratives in Mandarin. If Jenks' predictions were borne out, we would have expected the colors to cover the same contexts in the two languages. This is not what we find. Rather, the contracted/uncontracted distinction and the bare noun/demonstrative distinction seem orthogonal to one another.

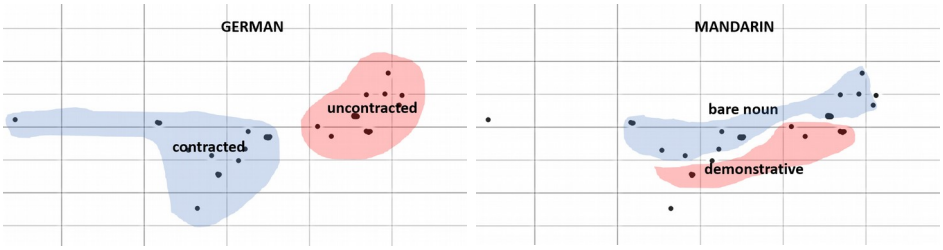

Figure 6: Enhanced TM maps for German and Mandarin

Closer study of the data reveals that the use of demonstratives in contracted contexts (N=3) probably relies on the deictic interpretation of demonstratives and not on their anaphoric use. This is illustrated in (5):

(5) **M:** Pèinī, wǒ juébù ràng tāmen rènhérén jìn **zhè dòng fángzi**.
Petunia I not have them anyone enter this CL house

(5) is the Mandarin version of example (1) and is uttered by a house owner who assures his wife that certain people will never be welcome in their house. The proximal demonstrative that is used refers to the house the two are in at the moment of speech.

With examples like (5) out of the way, the comparison between German and Mandarin becomes simpler. Demonstratives come out – as predicted by Jenks – as equivalents of German uncontracted forms (N=10). The problem that remains is that bare nouns not only occur as the equivalent of German contracted forms (N=34) but also of uncontracted forms (N=45). An example of a bare noun occurring as the equivalent of a German contracted form is given in (6). This is the Mandarin version of example (4) (see example (4) for the broader context):

(6) **M:** Tāmen pūshan-zhe chìbǎng gānggāng fēi zǒu, yòu
They flutter-ASP wings right fly away, and
yǒu yī zhǐ māotóuyīng xié lái yī fēng xìn,
have one CL owl bring come one CL letter
rēng zài **bāoguǒ** shàngmiàn.
throw to parcel on.

Unlike what we found for German and English, the Mandarin facts do not follow from current wisdom on how differences in form relate to differences in function. What they suggest is that weak/uniqueness definites are indeed conveyed by bare nouns but that strong/familiarity definites are conveyed both by bare nouns and demonstratives. These facts invite linguists to have a closer look at what distinguishes between bare nouns and demonstratives in strong/familiarity contexts.

The conclusion that imposes itself is that TM points to there being more functions in the definiteness domain than previously anticipated. In particular, Mandarin seems to come with

two types of strong/familiarity definiteness, one conveyed by bare nouns, the other conveyed by demonstratives.

5 Discussion and conclusion

In this paper, we have introduced Translation Mining as a data-driven way of doing cross-linguistic research. We have compared it to the classical semantic maps method and shown how it can be put to use in the study of definiteness. Our results show that Translation Mining not only allows us to confirm existing intuitions (see our discussion of German and English) but is also able to identify new areas of research (see our discussion of Mandarin).

The division of labor between Mandarin bare nouns and demonstratives in strong/anaphoric definite contexts suggests that there is a dimension of definiteness that has hitherto gone unnoticed. In Bremmers et al. (*ms.*) we argue that this new dimension is a stable one and not an accident of our data collection. We furthermore develop the intuition that bare nouns in anaphoric contexts need to be part of the same narrative sequence as the one their antecedent was introduced in (example (6)).[6] Outside these sequences, a demonstrative is required for anaphoric reference (example (2b)).

We would like to end by noting that it is clear to us that a standard approach to cross-linguistic research would sooner or later also establish a more fine-grained account of the Mandarin data. The value of TM lies in that it allows us to anticipate this result, even with a very restricted application like the one in this paper.

References

Aguilar-Guevara, A. et al. 2019. *Definiteness across languages*. Language Science Press.

Aguilar-Guevara, A. & J. Zwarts. 2010. Weak definites and reference to kinds. In N. Li and D. Lutz (eds.), Semantics and Linguistic Theory (SALT) 20, 179-196.

Beekhuizen, B., J. Watson & S. Stevenson. 2017. Semantic Typology and Parallel Corpora: Something about Indefinite Pronouns. In *CogSci*.

Bremmers, D. 2019. La définitude en français, anglais, allemand et mandarin. Un published BA thesis, Utrecht University. Available at https://dspace.library.uu.nl/handle/1874/384039.

Bremmers, D., J. Liu, M. van der Klis & B. Le Bruyn. 2019. *Translation Mining: definiteness across languages. A reply to Jenks (2018)*. Manuscript, Utrecht University. Available at https://time-in-translation.hum.uu.nl/.

Carlson, G. & R. Sussman. 2005. Seemingly indefinite definites. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 85, 71-85.

Georgakopoulos, T. & S. Polis. 2018. The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass*, 12(2).

Haspelmath, M. 1997. Explaining article-possessor complementarity: economic motivation in noun phrase syntax. *Language*, 227-243.

Haspelmath, M. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In *The new psychology of language*, 277-322. Psychology Press.

Heim, I. 1982. *The semantics of definite and indefinite noun phrases*. Doctoral dissertation, University of Massachusetts, Amherst.

Jenks, P. 2018. Articulated definiteness without articles. *Linguistic Inquiry* 49(3), 501- 536.

[6]The term narrative sequence goes back to the term *narration* in SDRT (e.g. Lascarides & Asher 1993) and involves sequences of (chronologically ordered) events.

Kamp, H. (1981). A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stokhof (eds.) *Formal Methods in the Study of Language*, 277-322. Mathematical Center Tract 135, Amsterdam.

Klis, M. van der, B. Le Bruyn & H. de Swart. 2017. Mapping the perfect via translation mining. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 497-502.

Klis, M. van der, B. Le Bruyn & H. de Swart. 2019. De la sémantique des temps verbaux à la traductologie: une comparaison multilingue de *L'Étranger* de Camus. In E. Corre, D.-T. Do-Hurinville, and H.-L. Dao (eds.) *Linguistic approaches to Tense, Aspect, Modality, Evidentiality, based on the Novel* L'Etranger ("The Stranger") *by Albert Camus, and its Translations.*

Lascarides, A. & N. Asher. 1993. Temporal interpretation, discourse relations, and common sense Entailment. *Linguistics and Philosophy* 16, 437-493.

Löbner, S. 2011. Concept types and determination. *Journal of semantics*, 28(3), 279-333.

Russell, B. 1905. On denoting. *Mind* 14(56), 479-493.

Schwarz, F. 2009. *Two Types of Definites in Natural Language.* Doctoral dissertation, University of Massachusetts, Amherst.

Strawson, P. 1950. On referring. *Mind* 59(235), 320-344.

Van der Auwera, J. & V. Plungian. 1998. Modality's semantic map. *Linguistic Typology* 2(1), 79-124.

Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3), 671-710.