languages

MDPI

Article

Parallel Corpus Research and Target Language Representativeness: The Contrastive, Typological, and *Translation Mining* Traditions

Bert Le Bruyn *, Martín Fuchs, Martijn van der Klis, Jianan Liu, Chou Mo, Jos Tellings and Henriëtte de Swart

Institute for Language Sciences, Utrecht University, 3512 JK Utrecht, The Netherlands; m.fuchs@uu.nl (M.F.); m.h.vanderklis@uu.nl (M.v.d.K.); j.liu4@uu.nl (J.L.); c.mo@uu.nl (C.M.); j.l.tellings@uu.nl (J.T.); h.deswart@uu.nl (H.d.S.)

* Correspondence: b.s.w.lebruyn@uu.nl

Abstract: This paper surveys the strategies that the Contrastive, Typological, and *Translation Mining* parallel corpus traditions rely on to deal with the issue of target language representativeness of translations. On the basis of a comparison of the corpus architectures and research designs of the three traditions, we argue that they have each developed their own representativeness strategies: (i) monolingual control corpora (Contrastive tradition), (ii) limits on the scope of research questions (Typological tradition), and (iii) parallel control corpora (*Translation Mining* tradition). We introduce normalized pointwise mutual information (NPMI) as a bi-directional measure of cross-linguistic association, allowing for an easy comparison of the outcomes of different traditions and the impact of the monolingual and parallel control corpus representativeness strategies. We further argue that corpus size has a major impact on the reliability of the monolingual control corpus strategy and that a sequential parallel control corpus strategy is preferable for smaller corpora.

Keywords: parallel corpora; translation; cross-linguistic variation

check for updates

Citation: Le Bruyn, Bert, Martín Fuchs, Martijn van der Klis, Jianan Liu, Chou Mo, Jos Tellings, and Henriëtte de Swart. 2022. Parallel Corpus Research and Target Language Representativeness: The Contrastive, Typological, and *Translation Mining* Traditions. *Languages* 7: 176. https://doi.org/10.3390/languages7030176

Academic Editors: Juana M. Liceras and Raquel Fernández Fuertes

Received: 7 December 2021
Accepted: 20 June 2022
Published: 7 July 2022

1. Introduction

Cross-linguistic research based on parallel corpora makes two fundamental assumptions: translations are representative of their target language, and they convey the same meaning as their originals. In this paper, we zoom in on the assumption of representativeness, and we refer the interested reader to Le Bruyn et al. (2022) for a recent discussion of the assumption of meaning equivalence.

The assumption of representativeness is, in general, not taken to be absolute: most papers that use parallel corpus data acknowledge that translated language differs from untranslated language and stress the preliminary or qualitative nature of their observations (see, e.g., Bogaards (2022), Grønn and von Stechow (2020)). At the same time, we find that recent studies using parallel corpora neatly succeed in replicating and fine-tuning cross-linguistic patterns that are predicted by the literature (see, e.g., Beekhuizen et al. (2017) on indefinite pronouns, and van der Klis et al. (2021a) on the HAVE-PERFECT), suggesting that the conclusions we can draw from parallel data are more reliable than what is generally assumed. The goal of the current paper is to bring together and evaluate methodological insights into how parallel corpus research allows us to go beyond preliminary observations, while, at the same time, guarding us from overstating the findings. We survey three parallel corpus traditions and identify the strategies each of them uses to deal with the issue of representativeness. We further lay the groundwork for assessing which of the strategies are to be pursued for different types of datasets.

The number of parallel corpus-based papers in this Special Issue (see Bogaards (2022); Corre (2022); Fuchs and González (2022); Gehrke (2022); Mulder et al. (2022); de Swart et al. (2022)) bears witness to the increasing role that parallel corpora play in recent cross-linguistic research on tense and aspect. In the early 2000s, however, the same empirical domain led some members of the linguistic community to turn their backs on parallel corpus research, capitalizing on the issue of representativeness (McEnery and Xiao 1999; McEnery et al. 2006). The growing role of parallel corpora in the recent literature on tense and aspect and the fact that it is one of the empirical domains that led to a fair amount of skepticism earlier on make this Special Issue into an ideal venue for a methodological reflection on the use of parallel corpora.

In this introductory section, we retrace the tense/aspect observations that inspired skepticism towards parallel corpus research (Section 1.1), introduce the three traditions we surveyed (Section 1.2), and present the roadmap of the paper (Section 1.3).

1.1. *Early Skepticism*

McEnery and Xiao (1999) report on the use of aspectual markers in the Mandarin part of the CEPC-health corpus, a parallel corpus containing English healthcare texts and their Mandarin translations. The English and Mandarin parts of the corpus each contain around 35k words.

With Mandarin generally being considered an aspect language, McEnery and Xiao were surprised to find that most verbs in the corpus are not accompanied by explicit aspectual markers such as *le*, *guo*, *zhe*, *zai*, etc. Opposing the verbs marked by *le* and *guo* to unmarked verbs, the authors found that the marked-to-unmarked ratio was 0.043:1. Up till then, the literature had argued that stative verbs do not have to be marked (Xiao 2002) and that narrative discourse may lead to a less frequent use of aspect markers (Chang 1986; Chu 1987; Yang 1995). However, only one-third of the corpus can be characterized as narrative, and even if stative verbs were to make up half of all verbs, the marked-to-unmarked ratio could not be explained by the lack of marking on stative verbs alone. Reflections such as these led McEnery and Xiao to hypothesize that the use of aspect markers in Mandarin translations is not representative of actual Mandarin aspect marking.

To evaluate their hypothesis, the authors compiled the C-health corpus, a corpus of original Mandarin healthcare texts of comparable size to the Mandarin part of the CEPC-health corpus and with the same division between narrative and non-narrative texts. They found that the joint frequency of *le* and *guo* was twice as high as in the CEPC-health corpus (213 vs. 98), a significant difference (LL value = 49.11, $p < 0.001$). The marked-to-unmarked ratio remained low (0.086:1), but it was higher than in the CEPC-corpus, and the increase could be attributed to the significantly higher frequency of markers. McEnery and Xiao concluded that their hypothesis was on the right track and that Mandarin translated texts are not representative of actual Mandarin aspect marking.

We assume that the tense/aspect findings in McEnery and Xiao (1999) were among those that led McEnery et al. (2006) to argue that 'translated language is at best an unrepresentative special variant of the target language' (McEnery et al. 2006, p. 93). This claim was not new at the time and echoed the earlier characterization of translated texts as displaying *translationese* (Gellerstam 1996) and the call to reserve the use of parallel corpora to the study of translation (Lauridsen 1996). What was new, though, was the fact that the claim was made by prominent corpus experts with extensive experience in the domain of tense and aspect, casting a shadow on the use of parallel corpora for cross-linguistic research in general and for the domain of tense and aspect in particular.

1.2. *Parallel Corpus Traditions*

Despite McEnery et al.'s claim, the literature of the past two decades shows that several research traditions have insightfully used parallel corpora for the study of cross-linguistic variation across a variety of empirical domains, including tense and aspect. The current paper is the first to bring together three of these traditions in a bigger methodological comparison, and we argue here that all of them have found ways to deal with the issue of representativeness.

The first tradition dates back to Johansson (1998a, 2007), and we refer to it as the *Corpus-based Contrastive Linguistics tradition* or *Contrastive tradition* for short (Granger and Lefer 2020). The strategy that researchers in this tradition rely on is to consistently complement findings based on parallel corpora with findings based on monolingual corpora. The latter then function as a control on translation biases in the former. The second tradition started with Wälchli (2010) and Wälchli and Cysouw (2012), and we refer to it as the *Typological tradition* (see—among others—Dahl and Wälchli (2016), Beekhuizen et al. (2017), Levshina (2022)). Different from the Contrastive tradition, researchers within the Typological tradition do not rely on monolingual control corpora. At the same time, they do implement an implicit control mechanism in that the questions they ask are not geared towards the analysis of individual items in individual languages, but rather towards higher-level generalizations that hold across a high number of languages and are unlikely to be sensitive to translation biases. The third and final tradition we discuss is a tradition in the making and originates in recent publications by the Utrecht-based *Translation Mining* group, who have applied parallel corpus research in contrastive and comparative linguistics (see—among others—van der Klis et al. (2017, 2021a), Le Bruyn et al. (2019), Bremmers et al. (2021)). We refer to it as the *Translation Mining* tradition. Researchers in this tradition do not rely on monolingual control corpora, but they are nevertheless interested in the analysis of individual items in individual languages. What sets this tradition apart is that it aims for replication across different parallel corpora with different source languages, allowing it to mimic the monolingual control mechanism from the Contrastive tradition across complementary parallel corpus studies.

The rationale behind our selection of parallel corpus traditions is that they represent the three strategies that are a priori available to control for translation biases within corpus research: the Typological tradition restricts the scope of the questions that can be asked, and the Contrastive and *Translation Mining* traditions rely on monolingual and parallel control corpora, respectively. We are confident that other traditions, such as Multiple Parallel Text Analysis (Lu and Verhagen 2016; Lu et al. 2018), Heuristic Translation Mining (Bogaards 2019, 2022) etc., can insightfully be related to one of the traditions discussed here.[1]

1.3. Roadmap

In Sections 2–4, we go over the different traditions in turn and present the strategies that they rely on to deal with the issue of representativeness. To properly ground our discussion, we introduce the corpus architectures of the different traditions, as well as the ways they exploit them, building on datasets and analyses from their respective works in the literature. In Section 5, we lay the foundations for future research to explore the differences between the strategies and highlight the role of corpus size in strategy selection. Section 6 concludes the paper with a general discussion.

2. The Contrastive Tradition

The Contrastive tradition is the most influential one to date and the one underlying most major parallel corpus compilation projects. These projects have led to the English Norwegian Parallel Corpus (Johansson 2007), the Oslo Multilingual Corpus (Johansson 2007), the English–Swedish Parallel Corpus (Altenberg and Aijmer 2000), the English–Portuguese COMPARA corpus (Frankenberg-Garcia and Santos 2003), the English–German CroCo corpus (Hansen-Schirra et al. 2013), and the Dutch Parallel Corpus (Macken et al. 2011). In this section, we analyze the parallel corpus architecture that the Contrastive tradition builds on (Section 2.1), present the measures it typically deploys (Section 2.2), and conclude that it deals with the issue of representativeness by consistently complementing findings based on parallel corpora with findings based on monolingual corpora (Section 2.3).

2.1. The Contrastive Parallel Corpus Architecture

Parallel corpora in the Contrastive tradition all have a similar architecture, including source texts for each of the languages represented in the corpus and translations of these texts to the other languages. This is worked out in Figure 1 for a corpus with two languages.

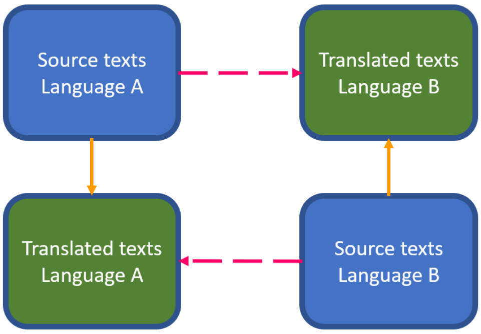


Figure 1. The Contrastive parallel corpus architecture (illustrated for two languages).

The architecture in Figure 1 reflects the way in which researchers in the Contrastive tradition deal with the issue of representativeness in the use of parallel corpora for cross-linguistic variation research. By comparing source and translated texts *between* languages (the dashed pink arrows in Figure 1), researchers can come to hypotheses about how languages relate to one another. These hypotheses are, however, explicitly treated as translation-based, and they are maintained, adjusted, or discarded on the basis of comparisons between source and translated texts *within* each language (the full orange arrows in Figure 1). In Section 2.2, we go over each of these comparisons and illustrate them on the basis of two datasets.

2.2. Putting the Contrastive Parallel Corpus Architecture to Use

2.2.1. Comparisons between Languages

We start with the comparison of source and translated texts *between* two languages. The basic measures in this comparison are measures of translation equivalence. To illustrate, Table 1 provides data about the translation of English *talk* to Norwegian and about the translation of Norwegian *snakke* to English (Hasselgård 2020). These data respect translation direction: the data in the left part of Table 1 are concerned only with English source texts and their translations to Norwegian, and not with translated texts in English and their original Norwegian counterparts. Mutatis mutandis, the same holds for the right part of Table 1.

Table 1. *Talk* and *Snakke* in English<>Norwegian translations.

English > Norwegian			**Norwegian > English**		
talk	*snakke*	204	snakke	*talk*	313
	prate	14		*left*	80
	fortelle	4		*say*	13
	si	3		*mention*	9
	other	34		*other*	61
	Ø	9		Ø	8
	total	268		total	484

The data in Table 1 show that *talk* is translated as *snakke* in 204 out of 268 cases. This means that 76% of the time *snakke* is used as a translation equivalent of *talk*. The data also show that in 313 out of 484 cases Norwegian *snakke* is translated as *talk*, meaning that, approximately 65% of the time, English *talk* is used as a translation equivalent of *snakke*. The formulas underlying these two measures of translation equivalence are given in (1) and (2).

$$(1)\ \frac{C'^{T\beta} \times 100}{C^{S\alpha}} \qquad (2)\ \frac{C'^{T\alpha} \times 100}{C^{S\beta}} \qquad (3)\ \frac{(C'^{T\beta} + C'^{T\alpha}) \times 100}{C^{S\alpha} + C^{S\beta}} \qquad (4)\ \frac{(204 + 313) \times 100}{268 + 484} = 69\%$$

where $C^{S\alpha/\beta}$ stands for the number of times a construction C occurs in Source Texts in language α/β, and $C'^{T\alpha/\beta}$ stands for the number of times construction C′ occurs in Translated Texts in language α/β as the translation of construction C.

Next to the strictly unidirectional measures in (1) and (2), researchers in the Contrastive tradition also use the bidirectional measure presented in (3) and applied to the *talk<>snakke* data in (4). This bidirectional measure is known as *mutual correspondence* and measures how often two constructions occur as translations of one another (Altenberg 1999). In the case of *talk* and *snakke*, this is 69% of the time. We note that the bidirectional nature of mutual correspondence still respects the unidirectionality of the data in Table 1 in the sense that it only looks at how translations relate to source texts and not vice versa: in the same way as in (1) and (2), translated texts only appear in the numerator and not in the denominator. Later, in Section 5.1, we come back to mutual correspondence and propose a more general bi-directional measure, *viz.*, normalized pointwise mutual information (Bouma 2009). The advantage of this more general measure is that, in contrast to mutual correspondence, it can be applied across the different parallel corpus traditions presented in this paper and thus allows for an easy comparison of their respective results.

2.2.2. From Comparisons between Languages to Comparisons within Each Language

As we indicated before, the comparison of source and translated texts *between* languages leads to hypotheses about how languages relate to one another. However, these hypotheses are primarily hypotheses about translation equivalence, and researchers working in the Contrastive tradition rely on the comparison of source and translated texts *within* each language to maintain, adjust, or discard these hypotheses in their move to hypotheses about cross-linguistic similarity. In this additional comparison, the source texts in the corpus are used as monolingual control corpora to filter out the effects of translation in translated texts.

We discuss the move from hypotheses about translation equivalence to hypotheses about cross-linguistic similarity on the basis of the dataset in Table 2 and the way it is discussed in Johansson (1998b, 2007). The dataset is concerned with the Norwegian verbs *hate* (‘hate’) and *elske* (‘love’) and the English verbs *hate* and *love* in the English Norwegian Parallel Corpus (ENPC).

Table 2. (Absolute) frequencies of ‘hate’ and ‘love’ verbs in Norwegian and English translated and untranslated texts.

	Source Texts	**Translated Texts**
Norwegian		
hate	23	34
elske	36	90
English		
hate	67	25
love	100	62

With Johansson, we observe that English *hate* and *love* are almost three times as frequent as Norwegian *hate* and *elske* in source texts but that this difference in frequency is smaller in translations. For Johansson, this leads to two conclusions. The first is based on the comparison of source and translated texts *between* the two languages: the Norwegian verbs have a smaller semantic range than the English verbs. Had their semantic range been identical, the frequencies of Norwegian *hate* and *elske* in translated texts would be practically identical to those of English *hate* and *love* in source texts and vice versa. The second conclusion Johansson draws is based on the comparison of source and translated texts *within* each language: through the influence of translation, the semantic range of Norwegian *hate* and *elske* is broader in translated texts than in the standard language, and the semantic range of English *hate* and *love* is smaller. Had they had a constant semantic range—Johansson's reasoning goes—their frequencies in source and translated texts *within* each language would have been closer to one another. Combining the two conclusions, Johansson argues that the English verbs have a broader semantic range than the Norwegian ones, and that this difference in range is bigger than what the comparison of source and translated texts *between* the two languages suggests.

We will later probe the two assumptions that underlie Johansson's conclusions in more detail in Section 5.2. For now, it suffices to mention them and relate them to the assumptions that underlie the argumentation behind McEnery et al.'s claim that translated language is an unrepresentative special variant of the target language. The first assumption is that source texts are representative of the target language, and the second is that differences between source and translated texts are to be related to the influence of translation. It is these assumptions that led Xiao and McEnery (2004) to compare translated to untranslated texts and argue that translations provide a distorted view of the target language. For Johansson and other researchers in the Contrastive tradition, these assumptions lead them to a corpus architecture that includes source and translated texts for the different languages in the corpus and that is used to tease apart cross-linguistic variation from variation induced by translation.

2.3. The Contrastive Tradition: Conclusion

In this section, we have presented the parallel corpus architecture used in the Contrastive tradition and we have worked out the two types of comparisons that are typically performed on the data. We found that both the corpus architecture and comparisons are geared towards allowing monolingual data to function as a filter for parallel data. We conclude that the Contrastive tradition deals with the issue of representativeness by consistently comparing multilingual data with monolingual data, integrating both in a uniform corpus and research design.

3. The Typological Tradition

In this section, we move to the Typological tradition (Beekhuizen et al. 2017; Dahl and Wälchli 2016; Levshina 2022; Wälchli 2010; Wälchli and Cysouw 2012). This tradition is radically different from the Contrastive tradition, and this might explain why it is not included in the recent overview of parallel corpus research in Granger and Lefer (2020). A striking feature of this tradition is that it acknowledges the issue of representativeness but—at the surface—does not seem to actively control for it. We argue that the Typological tradition does come with safeguards against the influences of translation but that these reside in the details of how it puts parallel corpora to work. To get to the relevant level of detail, we zoom in on one specific study (Section 3.1) and characterize its parallel corpus architecture (Section 3.2), as well as the way it puts it to use (Section 3.3). The study we select is Wälchli and Cysouw (2012), one of the founding studies of the Typological tradition.

3.1. A Specific Study in the Typological Tradition

Wälchli and Cysouw (2012) is a lexical typology study of motion verbs. Their empirical basis is a selection of 360 motion verb contexts from the gospel by Mark and the verbs that are used to render them in a typologically diverse sample of about 100 languages. To be able to interpret the resulting dataset of over 30k datapoints, the authors transform it into a dissimilarity matrix in which the differences in verb choice between pairs of contexts are quantified according to a verb-level version of the Hamming Distance. This dissimilarity matrix is then used as input for Multi-Dimensional Scaling (MDS), which, in turn, constitutes the basis for the interpretation of the dataset. We first introduce the way the dissimilarity matrix is set up and then move to the way Wälchli and Cysouw interpret it and use MDS to help them do so.

3.1.1. The Dissimilarity Matrix

We illustrate the transformation of the initial dataset into a dissimilarity matrix with the sample dataset in Table 3. For convenience, we follow Wälchli and Cysouw in referring to the different motion verb contexts with verse numbers.

Table 3. Sample dataset from Wälchli and Cysouw (2012).

Motion Verb Context (Verse).	English	French	Hungarian	Mapudungun
4:4	*come*	*venir*	*jön*	*aku*
5:1	*come*	*arriver*	*ér*	*puw*
9:33	*come*	*arriver*	*ér*	*puw*
14:3	*come*	*entrer*	*lép*	*aku*

As can easily be established, the motion verbs used in verses 5:1 and 9:33 are identical. If we conceive of differences between contexts in terms of distance, this means that the distance between the two verses is 0. The motion verbs in verses 5:1 and 9:33 do differ from the motion verbs in 4:4 and 14:3, and those, in turn, also differ from each other. The verb-level Hamming Distance between the different contexts is calculated on the basis of a pairwise comparison of their corresponding verb tuples (e.g., <*come*, *venir*, *jön*, *aku*>). Each position in the tuples that contains a different verb (e.g., *venir* vs. *arriver* in the second position of the tuples corresponding to verses 4:4 and 5:1) adds 0.25 (1/#languages) to the distance between pairs of contexts. When we add up these distances, the dataset in Table 3 can be summarized by the dissimilarity matrix in Table 4.

Table 4. Dissimilarity matrix of the sample dataset in Table 3.

	4:4	5:1	9:33	14:3
4:4	0	0.75	0.75	0.50
5:1	0.75	0	0	0.75
9:33	0.75	0	0	0.75
14:3	0.50	0.75	0.75	0

The distance between verses 4:4 and 5:1 is 0.75. This means that only one language makes the same choice of lexical verb in these verses whereas all other languages make a different choice. The distance between verses 4:4 and 14:3 is 0.50, reflecting the fact that two languages select the same verb whereas the other two select different verbs.

3.1.2. Interpreting the Dissimilarity Matrix and MDS

Wälchli and Cysouw's interpretation of the dissimilarity matrix is grounded in the assumption that differences in verb choice across contexts reflect differences in meaning between contexts. The dissimilarity matrix then gives us a fine-grained overview of semantic differences between motion verb contexts. In this sense, it identifies 'typologically relevant features in the grammatical structure of the lexicon', and this is in line with Lehmann's view on lexical typology (Lehmann 1990, p. 163). We note that we use the term *feature* rather loosely here, as Wälchli and Cysouw insist on the fact that they do not identify features but rather category types. The difference in terminology is connected with the similarity-based nature of category types that becomes clear in the remainder of this section. What is important for now is that the identification of category types is central to the approach and that 'the semantics of individual lexical items, their configurations in lexical field or individual processes of word formation' are less central (Lehmann 1990, p. 165).

Interpreting a dissimilarity matrix is not an easy feat, and it helps to visualize it by plotting the distances between the verses, as we do in Figure 2 for the sample dissimilarity matrix in Table 4.

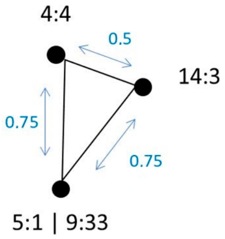


Figure 2. Visual rendition of the distances in the sample dissimilarity matrix in Table 4.

Figure 2 is a visual rendition of the distances in the dissimilarity matrix in Table 4. Verses 5:1 and 9:33 occupy the same position and are at an equal distance from verses 4:4 and 14:3. The latter two verses are at different positions, and the distance between them is two-thirds of the distance between each of them and verses 5:1 and 9:33. A visual rendition of the full dissimilarity matrix of Wälchli and Cysouw can be thought of as a corpus-based representation of the semantic space of motion verbs. In line with the assumption that differences in verb choice across contexts reflect differences in meaning, we can assume that the more distant verses are in this semantic space, the likelier they are to differ in their semantics. Conversely, the closer verses are positioned to one another, the likelier they are to have similar semantics.

We generated Figure 2 by hand, but this is not possible for Wälchli and Cysouw's full dataset. Setting aside the labor intensity it requires, the bigger challenge is that a two-dimensional space similar to that in Figure 2 does not allow us to faithfully render all the distances in Wälchli and Cysouw's full dissimilarity matrix. This already becomes clear when we try to add verse 1:31:

(5) 1:31 <*come*, *s'approcher*, *megy*, *fülkon*>

Verse 1:31 turns out to be at a distance of 0.75 from each of the four verses in Table 3. Within a two-dimensional space, there is no way to faithfully render the respective distances between all five verses. The only way to do so is to add an extra dimension. With 360 verses in total and distances that vary between 0 and 1 with increments of about 0.01, it is likely that we would need many more dimensions than we can easily draw by hand or even conceptualize. Identifying the relevant dimensions can, however, be important from a semantic perspective. Indeed, dimensions generalize over distances between verses and can consequently be thought of as reflecting higher-order differences in meaning. In a highly variable semantic domain such as that of motion verbs, dimensions are then likely to allow us to identify the typologically most relevant features that are at play. Accepting the limitations of the human brain, Wälchli and Cysouw turn to MDS to help them identify dimensions and interpret the semantic distinctions they make.

MDS allows researchers to take a dissimilarity matrix and automatically generate a visualization in a high-dimensional space. For Wälchli and Cysouw's dataset, it thus gives a full representation of the semantic space of motion verbs in the corpus. At the same time, MDS also allows researchers to limit the proliferation of dimensions by weighing the addition of a dimension against the contribution it makes to the faithfulness of the distances in the dissimilarity matrix. From a semantic point of view, it thus allows researchers to statistically evaluate how many and which dimensions are likely to be active in a specific semantic domain. For the domain of motion verbs, Wälchli and Cysouw use the eigenvalues of dimensions to argue that at least the first 30 dimensions are statistically relevant in rendering the distances in the dissimilarity matrix. They also propose an analysis of the 12 statistically most important dimensions. For example, the first would oppose COME verbs to GO verbs. The tenth dimension would further oppose COME verbs to ARRIVE verbs.

These verb types correspond to category types in Wälchli and Cysouw's terminology. As the reader may have noticed, these category types are not discrete but similarity-based: verbs can be more or less COME-, GO-, or ARRIVE-like, but there is no real cutoff point between the different category types.

To give an idea of how the actual output of MDS looks, we present the way it maps the different motion verb contexts of Wälchli and Cysouw in Figure 3. The dimensions we selected are the ones we discussed above: Dimension 1, opposing the COME and GO category types; and Dimension 10, opposing the COME and ARRIVE category types. For concreteness, we have overlayed this visualization with markup identifying individual verbs in English. The blue rectangles stand for motion contexts that take English *come*, the green triangles for motion contexts that take English *go*, and the black dots for motion contexts rendered with a different verb in English.

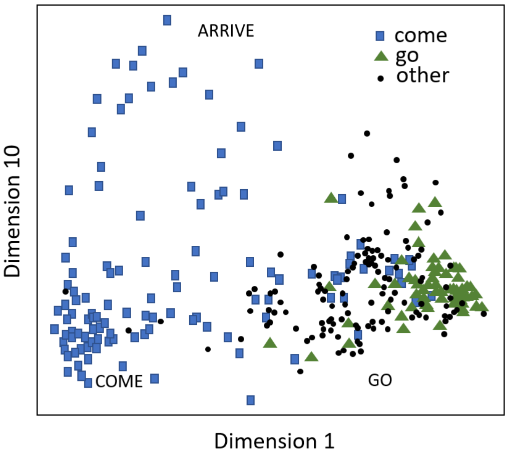


Figure 3. Impression of dimensions 1 and 10 from Wälchli and Cysouw (2012), with markup based on the English version of the gospel by Mark (see also Figure 3 in Wälchli and Cysouw (2012)).

Figure 3 shows how the opposition between English *come* and *go* neatly reflects the opposition between the category types COME and GO in dimension 1. The opposition between the category types COME and ARRIVE in dimension 10 is less pronounced in English: the core of the ARRIVE category type is lexicalized as *come* in the same way as the core of the COME category type.

3.1.3. Summary

In this section, we have shown how Wälchli and Cysouw (2012) use a selection of 360 motion verb contexts from the gospel by Mark to identify typologically relevant category types in the lexical domain of motion verbs. Their approach starts from an inventory of the verbs that are used in these contexts across a typologically diverse sample of about 100 languages. This inventory is then transformed into a dissimilarity matrix that quantifies the differences in lexicalization in the different contexts. The interpretation of this dissimilarity matrix relies on the assumption that differences in lexicalization reflect differences in meaning: contexts that are closer to one another are assumed to be more semantically similar than those that are further away from each other. The authors rely on MDS to help out in the interpretation of the dissimilarity matrix. On the one hand, MDS helps in the visualization of the corresponding high-dimensional semantic space. On the other hand, it helps in identifying the statistically relevant dimensions. Wälchli and Cysouw relate these to category types, a similarity-based version of the classical (discrete) features in lexical typology.

3.2. The Typological Parallel Corpus Architecture

The way in which Wälchli and Cysouw put parallel texts to use is radically different from the way they are put to use in the Contrastive tradition. The literature following in their footsteps varies (see—among others—Beekhuizen et al. (2017), Dahl and Wälchli (2016), Levshina (2022)), but many methodological choices align. In this section and the next, we characterize the parallel corpus architecture underlying this tradition and the way it is put to use. At the same time, we make a comparison with the Contrastive tradition.

Two design features of the parallel corpus architecture in the Typological tradition stand out when compared to the architecture in the Contrastive tradition. The first is that the Typological parallel corpus architecture comes with a far higher number of languages. Even though we did not draw attention to the number of languages in the Contrastive tradition, we typically find their corpora to be limited to two or three languages (e.g., respectively, the English–Norwegian Parallel Corpus and the Dutch Parallel Corpus). As we saw in Section 3.1, Wälchli and Cysouw (2012) rely on a corpus with around 100 languages. Other studies in the Typological tradition rely on corpora with tens of languages (e.g., Levshina 2022) to hundreds of languages (Dahl and Wälchli 2016). This difference with the Contrastive tradition is in line with the respective domains of linguistic research from which the traditions originate. Contrastive linguistics is typically concerned with a small number of languages, whereas typology traditionally builds on far bigger samples.

The second design feature that radically opposes the Typological parallel corpus architecture to the Contrastive one is that it does not come with source texts in all languages. The upshot of this is that no comparison can be made between source and translated texts *within* each language. In Section 2, we argued that this comparison is the foundation of the Contrastive tradition's way of controlling for translation biases. No such control is possible in the Typological tradition. What we find then is that translation data in the Typological tradition are put at the same level as untranslated data and are taken to be directly representative of the languages they are written in. The parallel corpus architecture of the Typological tradition can thus be represented as in Figure 4.

In this architecture, no distinction is made between translated and untranslated texts, and comparisons are directly undertaken for all data in all languages. Judging by this architecture, we would have to conclude that the Typological tradition does not address the issue of representativeness. This conclusion is correct insofar as research in the Typological tradition does not actively control for translation biases through its corpus architecture. In the next section, however, we argue that the issue of representativeness of translation data does not arise in the Typological tradition in the same way that it does in the Contrastive tradition.

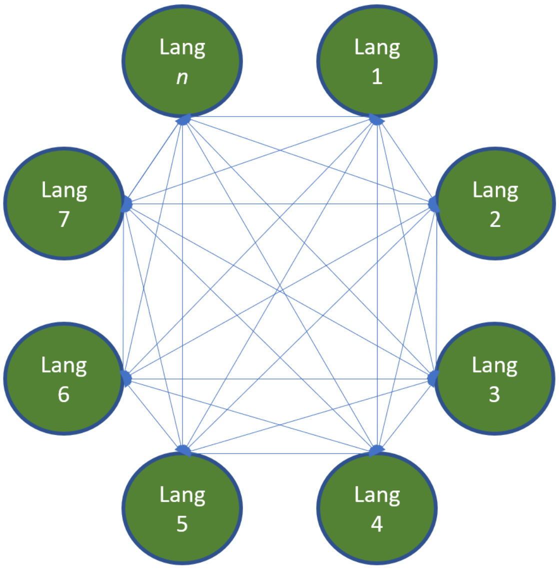


Figure 4. The Typological parallel corpus architecture (illustrated for *n* languages).

3.3. Putting the Typological Parallel Corpus Architecture to Use

Based on the discussion in Section 3.2, one might assume that the fact that no safeguard against the influence of translation is built into the Typological parallel corpus architecture is a trade-off between coverage on the one hand and control over the data on the other. This would make sense, as the Contrastive parallel corpus architecture is very demanding. Indeed, the fact that Contrastive corpora are typically restricted to two or three languages is not an accident, and researchers such as Johansson have indicated that their attempts at consistently implementing the Contrastive parallel corpus architecture for more than three languages have led to mixed results (Johansson 2007). Challenges include copyright issues, but also availability of comparable texts and translations in the different languages. These issues also explain why parallel corpora within the Contrastive tradition are fairly small. For example, the number of words of the English–Norwegian parallel corpus is expressed in the hundreds of thousands of words, a far cry away from the millions and billions of words that are becoming standard practice in monolingual corpus studies. Moving from contrastive to typological research, there is simply no way to implement the architecture of the Contrastive tradition for corpora that have the language coverage required for typological studies.

The practical challenges for extending the control strategy of the Contrastive tradition to the Typological tradition are clear. One might then argue—following Lauridsen (1996)—that it would be better not to use parallel corpora for typological research rather than to build analyses on unreliable data. However, next to the practical considerations, there is an arguably more fundamental reason for the more relaxed way in which researchers in Wälchli and Cysouw's tradition deal with translated texts. This reason is not made explicit in this tradition but relates to the way translation data are put to use. In the Contrastive tradition, researchers compare individual lexical items or constructions at the level of individual languages. This is reflected in the basic correspondence measures they use (see (1) to (3)). In the Typological tradition, the researchers' interest is not in the comparison of individual lexical items or constructions per se. Rather, they use these as a probe into typologically relevant features or category types. This is reflected in the fact that the dissimilarity matrix in Wälchli and Cysouw (2012) compares contexts and not individual expressions. Due to the high number of languages, the influence of each translation on the distances between different contexts is extremely small. Given that the interpretation of the dissimilarity matrix is furthermore statistically driven, and its analysis is limited to the statistically most important dimensions, it is highly unlikely that translation biases in individual translations have a traceable effect on the typological generalizations that researchers within Wälchli and Cysouw's tradition arrive at. Therefore, the use of parallel corpora in the Typological tradition is not a priori methodologically less sound than that in the Contrastive tradition, provided that the focus is on identifying typologically relevant features or category types rather than providing a full analysis of all lexical items included in the corpus.[2]

3.4. The Typological Tradition: Conclusion

In this section, we have presented the Typological parallel corpus tradition. After a detailed presentation of one of the foundational studies (Section 3.1), we discussed the typological parallel corpus architecture and the way it is put to use. We showed how the architecture differs from the one in the Contrastive tradition in the number of languages that are represented and in the fact that translated texts get the same status as untranslated texts (Section 3.2). We further argued that the way in which the Typological tradition deals with translation biases is by not making claims about individual lexical items, limiting the influence of individual translations on the data, and focusing in its analysis on the statistically most important tendencies (Section 3.3). We conclude that the Typological tradition is fundamentally different from the Contrastive tradition and that—through the questions it asks to parallel corpora—it has developed its own strategy to deal with the issue of representativeness.

4. The Translation Mining Tradition

In this section, we turn from the well-established Contrastive and Typological traditions to a tradition in the making. In a number of recent papers, the Utrecht-based *Translation Mining* group has used parallel corpora in a way that presents an interesting mix between the Contrastive and the Typological traditions. On the one hand, we see that it relies on the Typological parallel corpus architecture in putting translated and untranslated texts at the same level. On the other hand, we see that the questions it asks are the same as in the Contrastive tradition and concern the analysis of individual items in individual languages. From our discussion in Section 3, it follows that this particular combination is not a priori without problems. In this section, however, we argue that the *Translation Mining* group still succeeds in dealing with the issue of representativeness. Where the Contrastive tradition does so with a check on translation influence at the level of the corpus itself, we see that this check comes back in an extended research design in the work of the *Translation Mining* group. At the end of this section, we conclude that the *Translation Mining* tradition does deal with the issue of representativeness—be it in a way that differs from the strategies we found in the Contrastive and Typological traditions.

Parallel to our discussion of the two other traditions, we present the *Translation Mining* tradition by discussing its parallel corpus architecture (Section 4.1) and the way it puts it to use (Section 4.2). For concreteness, we take van der Klis et al. (2021a) as our starting point.

4.1. The Translation Mining Parallel Corpus Architecture

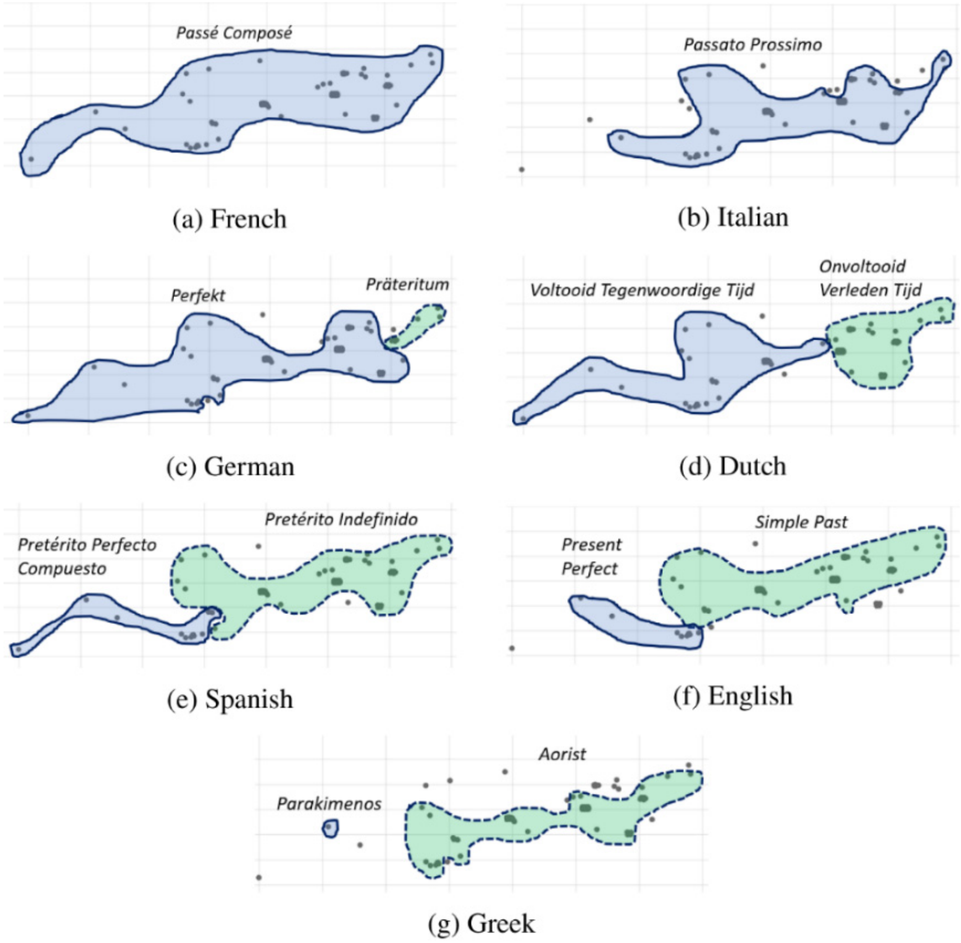
van der Klis et al. (2021a) study the cross-linguistic variation in the use of a specific verb form, *viz*., the combination of *have* (or its counterparts) and a past participle. They refer to this form as the HAVE-perfect (Dahl and Velupillai 2013). Van der Klis et al. take a form-based perspective to cross-linguistic variation and look at how the same form is used in different languages. The languages they look into are French, Italian, German, Dutch, Spanish, English, and (Modern) Greek. The HAVE-perfect form in these languages has received different names, such as *passé composé* in French, *voltooid tegenwoordige tijd* in Dutch, and *parakimenos* in Greek. In the current paper, we stick to the more general HAVE-perfect label.

The corpus van der Klis et al. use consists of all the contexts in which French uses a HAVE-perfect in the first three chapters of Camus' *L'Étranger (n* = 348). In the same way as Wälchli and Cysouw (2012) use MDS to map out distances between contexts based on lexical choices (Section 3.1), van der Klis et al. use it to map out distances based on tense choices. Figure 5 presents the MDS outcomes.

The gray dots in Figure 5a–g stand for contexts from the corpus. The dots are organized in the same way in the different figures, and their organization is based on the first two MDS dimensions. The markup is language-specific and reflects the tense use in the contexts. The blue shapes (full lines) cover contexts with a HAVE-perfect, and the green shapes (dotted lines) contexts with other (past) tenses (*Präteritum*, *Onvoltooid Verleden Tijd*, etc.). The picture that can be read off these maps is that the French and Italian HAVE-perfects have the most extensive use, and that other tenses systematically take over more and more contexts from one language to the next, leading to the implicational hierarchy in (6):

(6) *Implicational hierarchy of* HAVE*-perfect use in past contexts*
French | Italian > German > Dutch > Spanish > English > Greek

We note, with van der Klis et al., that the choice of corpus makes it impossible to check whether there are languages with HAVE-perfects that are not rendered as a HAVE-perfect in French. This issue is picked up by Le Bruyn et al. (2022), who show that none of these languages uses a HAVE-perfect in contexts in which French relies on a past tense. Figure 5a–g and the hierarchy in (6) thus give a fair idea of how the HAVE-perfect competes with past tenses in these different languages.



(a) French

(b) Italian

(c) German

(d) Dutch

(e) Spanish

(f) English

(g) Greek

Figure 5. MDS representations of the use of the HAVE-perfect in the different languages included in the corpus of van der Klis et al. (2021a) (see Figure 3 in van der Klis et al. (2021a)).

In van der Klis et al. (2021a), the MDS representations in Figure 5 and the implicational hierarchy in (6) are the basis for the interpretation of the data in the corpus. As such, translated data are taken at face value, and the parallel corpus architecture is thus the same as that in the Typological tradition, where translated and untranslated data are considered at the same level (see Figure 4).

4.2. The Translation Mining Interpretation of the Typological Corpus Architecture

The *Translation Mining* tradition uses MDS in the same way as the Typological tradition. This might seem to suggest that its aim is to focus on major cross-linguistic tendencies. Even though this is not excluded per se, the type of variation van der Klis et al. (2021a) looks into is far less complex than that in studies such as Wälchli and Cysouw (2012). To give an idea of the difference in complexity, we ran an eigenvalue analysis on the datasets of Wälchli and Cysouw and of van der Klis et al. (2021a).[3] In such an analysis, we can determine how many dimensions should be considered: only those dimensions whose eigenvalues are larger than the absolute value of the smallest eigenvalue are of interest statistically (Wälchli and Cysouw 2012, p. 685).[4] The result of our analysis is that only 2 dimensions are required for van der Klis et al. (2021a)'s data, as opposed to 54 for Wälchli and Cysouw (2012)'s data. For completeness, we also checked whether the difference in complexity was based on the number of languages or on the difference between lexical and grammatical variation. To do so, we ran the eigenvalue analysis on a subset of the languages in Wälchli and Cysouw (2012)'s data, trying to stay as close as possible to the seven languages included in van der Klis et al.'s study. The languages we included were

French, Italian, (Bernese Swiss) German, Swedish (instead of Dutch), Spanish, English, and (Modern) Greek. The result of the analysis is that 24 dimensions are required for a faithful rendition of the dissimilarity matrix data. What this shows is that both the number of languages and the type of variation (lexical vs. grammatical) influence the difference in complexity between the data in van der Klis et al. (2021a) and those in Wälchli and Cysouw (2012). We note that the number of contexts included in both datasets is close enough (348 in van der Klis et al. (2021a) vs. 360 in Wälchli and Cysouw (2012)) to have had little impact on the difference in the number of statistically relevant dimensions.

Wälchli and Cysouw do not present a full interpretation of their dataset but limit themselves to an interpretation of the twelve statistically most relevant dimensions. As such, they solely focus on major cross-linguistic tendencies that are unlikely to be subject to translation biases. Van der Klis et al.'s study is different in that the authors move along the hierarchy in (6) and present detailed analyses of every group of contexts that gets subtracted from the distribution of the HAVE-perfect.[5] Their main claim is that the analysis of these groups of contexts allows them to determine and oppose the relevant semantic ingredients of the HAVE-perfects of the individual languages represented in their corpus. Different from Wälchli and Cysouw, the final aim of van der Klis et al. is thus a full analysis of individual forms in individual languages. In this sense, the *Translation Mining* tradition differs from the Typological tradition and joins the aims of the Contrastive tradition.

In Section 4.1, we concluded that the *Translation Mining* tradition relies on the same parallel corpus architecture as the Typological tradition. In the current section, we have established that the final aim of the *Translation Mining* tradition is more in line with that of the Contrastive tradition in that it does not look for major cross-linguistic tendencies but tries to get a grip on the analysis of individual forms in individual languages. We thus find that the *Translation Mining* tradition presents a mix of the two other traditions: in its parallel corpus architecture, it sides with the Typological tradition, and in its aims, it sides with the Contrastive tradition. From our discussion in Section 3, it follows that this particular mix might presents some problems: the analysis relies on the details of the data, but the architecture prevents researchers from evaluating to what extent these details are sensitive to translation biases. The question that imposes itself then is whether the *Translation Mining* tradition still deals with the issue of representativeness.

The answer to the preceding question is affirmative. To see why, it is important to understand that a parallel corpus study within the *Translation Mining* tradition does not stand on its own. Where the Contrastive tradition builds safeguards against translation biases into its parallel corpus architecture and includes the analysis of parallel and monolingual corpus data in a single study, the *Translation Mining* tradition has a more sequential way of proceeding, and studies the same phenomenon from the perspective of different parallel corpora with different source languages (compare, e.g., van der Klis et al. (2021a) and Le Bruyn et al. (2019)). In this sense, it uses a parallel control corpus strategy, replicating the monolingual control corpus strategy of the Contrastive tradition across different studies. Next to this primary strategy, the *Translation Mining* tradition also relies on native speakers' judgements as an initial check on the acceptability and naturalness of parallel corpus data (see, e.g., Bremmers et al. (2021)), comparisons of different translations of the same source text in a single target language (see, e.g., Bogaart and Jager (2020)), and experiments testing newly found generalizations (see, e.g., Tellings and Fuchs (2021)).

4.3. The Translation Mining Tradition: Conclusion

In this section, we have presented the *Translation Mining* parallel corpus tradition, zooming in on the parallel corpus architecture in van der Klis et al. (2021a) and the way the authors put it to use. We showed how this tradition adopts the parallel corpus architecture of the Typological tradition while sharing the goals of the Contrastive tradition. As for the issue of representativeness, we argued that the *Translation Mining* tradition covers the same safeguards against the influence of translation as the Contrastive tradition, be it across complementary studies. We conclude that the three parallel corpus traditions come with strategies to deal with the issue of representativeness: (i) monolingual control corpora (Contrastive tradition), (ii) limits on the scope of research questions (Typological tradition), and (iii) parallel control corpora (*Translation Mining* tradition).

5. Choosing a Strategy: Some Preliminaries

With three strategies to deal with the issue of representativeness, the next step is to ask under which conditions each of them is more appropriate. For studies interested in analyzing broad typological generalizations, the sparsity of massively parallel texts strongly favors the strategy adopted by the Typological tradition. Practical considerations are also decisive for studies interested in comparing individual items/constructions across multiple languages: we have seen that the Contrastive parallel corpus architecture cannot easily be generalized beyond three languages, so that researchers with a comparative focus on individual items/constructions should resort to the strategy of the *Translation Mining* tradition. Turning to contrastive studies, the choice is one between the representativeness strategies of the Contrastive and the *Translation Mining* traditions, and this is the choice we focus on in this section. Providing a full decision tree to decide between the two lies beyond the scope of this paper, but we here set the stage for future studies to build on. On the one hand, we introduce a generalized version of the measure of mutual correspondence (see Section 2.2) that can be applied to parallel corpora in the two traditions (Section 5.1). This will allow future studies to compare their respective outcomes more easily and evaluate the impact of their respective representativeness strategies. On the other hand, we argue that the assumptions about monolingual and translation corpora that motivate the representativeness strategy of the Contrastive tradition (see Section 2.2) may be correct for big corpora but may not always hold for smaller corpora (Section 5.2). This, in turn, influences the choice between the representativeness strategies of the Contrastive and the *Translation Mining* traditions (Section 5.3).

5.1. Generalizing Mutual Correspondence

In Section 2.2, we presented a number of measures that are used in the Contrastive tradition and illustrated them with English *talk* and Norwegian *snakke* data. The unidirectional measures in (1) and (2) were concerned with how often a given form in the target language is a translation of a given form in the source language. These measures can easily be generalized to apply to any pair of languages, independently of whether they are source or target languages. The same does not hold for the bidirectional measure defined in (3): mutual correspondence. The problem that presents itself is that mutual correspondence is based on data from two independent samples of source and translated texts. Corpora in the *Translation Mining* tradition do not come with such independent samples, and we consequently have to reconsider the rationale behind mutual correspondence to come to a bidirectional measure that can be applied in both traditions, allowing for an easy comparison of their respective findings. We argue that normalized pointwise mutual information (NPMI, Bouma 2009), a measure that originates in Information Theory, provides us with such an alternative rationale. For concreteness, we work with a parallel dataset containing all French indicative verb forms (n = 389) from the first chapter of Camus' *L'Étranger* and their translations to Mandarin. We focus on the French *imparfait* and the way it relates to the Mandarin progressive marker *zai*.

The intuition behind our use of NPMI can best be understood with a small analogy. Imagine person A is tossing a coin and person B is throwing a die. The probability of person A ending up with heads is one out of two, and the probability of person B ending up with six is one out of six. The probability of them ending up with these results in the same turn is the product of the probabilities of each of the results, *viz.*, 1 out of 12. With two fully independent processes, we can thus calculate the probability of ending up with a given pair of results by multiplying the probabilities of the individual results.

Moving to parallel texts, the turns in the coin-and-die example become pairs of expressions that occur as each other's counterparts. We refer to these pairs as *counterpart* *pairs*, or CPs for short. Frequencies give us a handle on the probabilities of individual expressions occurring in CPs. For example, for our Camus data, the probability of finding the French *imparfait* in a CP is 140 out of 389. Likewise, the probability of finding Mandarin *zai* in a CP is 6 out of 389. With the probabilities of these individual expressions in place, we can calculate what the probability of them occurring in the same CP would be if co-occurrence in a CP were random. In the same way as for the coin-and-die example, we simply have to take the product of the individual probabilities; namely, 840 out of 151,321.

Clearly, co-occurrence in a CP is not random, but, given that we can calculate what the probability of two expressions co-occurring in a CP would be if it were, we can compare the actual probability of finding them together in a CP in the corpus to this hypothetical probability. By dividing the actual probability by the hypothetical probability, we then get a measure of how strongly the two expressions are associated with each other across their respective languages. For the *imparfait* and *zai* in our Camus data, the actual probability is 6 out of 389. If we divide this by their hypothetical probability, we end up with 2.78, indicating that the actual probability of finding the *imparfait* in the same CP as *zai* is approximately three times higher than we would expect on the basis of random co-occurrence.

NPMI builds on the actual/hypothetical probability ratio we have introduced, but puts it through two further transformations. The first is to take the (binary) logarithm of this ratio. The main effect of this operation is that the cutoff point between actual probabilities that are higher than the hypothetical ones is moved from 1 to 0. The rationale behind this first transformation is internal to Information Theory, where information is measured in (binary) bits, and random co-occurrence is taken to have no information value. The second transformation consists in dividing the result of the first transformation by the negative value of the logarithm of the actual probability. The latter value equals the result of the first transformation in case the two expressions in question only occur together, i.e., when the ratio is at its highest. The effect of this operation then is to project all values on a scale from -1 to 1, while maintaining 0 as the cutoff point between actual probabilities that are higher than the hypothetical ones. This second transformation thus counts as a normalization and allows for easy comparison across datasets. The final NPMI value for the *imparfait* and *zai* in our Camus dataset is then $\log(2.78)/-\log(6/389) = 0.25$.

With the definition of NPMI in place, we are in a good position to come back to mutual correspondence and discuss the way the two measures relate to one another. We argue that both measures quantify the strength of association between expressions across languages. Mutual correspondence does so by comparing the frequency of expressions in source texts to the frequency of their counterparts in translations. The more frequently their counterparts occur as their translation equivalents, the stronger the association is between the expressions and their counterparts. NPMI follows a different route and compares the actual probability of two expressions occurring as each other's counterparts to the hypothetical probability of the two randomly occurring as each other's counterparts. The more NPMI approaches a value of 1, the stronger the association is between the two expressions. Despite the fact that mutual correspondence and NPMI are clearly mathematically different, we conclude that they do measure the same construct, *viz.*, the association between expressions across languages. NPMI comes out as the more general measure, as it does not rely on two independent samples of source and translated texts. It can consequently be used in parallel corpora from both the Contrastive and the *Translation Mining* traditions and thus allows for easy comparison of their outcomes and of the impact of their respective representativeness strategies.

5.2. Assumptions of the Contrastive Tradition and Corpus Size

In Section 2.2, we pointed out that there are two assumptions that underlie the representativeness strategy of the Contrastive tradition. The first is that source texts are representative of the target language, whereas translated texts are less so; the second is that the differences between source and translated texts *within* a language are to be related to the process of translation underlying the latter. Even though we agree with these assumptions, we also want to warrant against too strict an interpretation, in particular, for smaller corpora similar to the ones used in the different traditions discussed in this paper. The argument we develop is as follows: if (potential) source texts were representative of the target language, we would expect there to be little to no variation between them. We show that this expectation is not borne out and conclude that taking source texts as the ultimate touchstone for target language representativeness in parallel corpus research is not a foolproof strategy.

To make our discussion as concrete as possible, we go back to Johansson's study on *hate* and *love* in the ENPC and remind the reader that Johansson observes that English *hate* and *love* are less frequent in translated texts than in source texts, and that he relates this fact to the influence of translation. On the strongest interpretation of Johansson's reasoning, there should be no independent reason for *hate* and *love* to be less frequent in the translated texts of his corpus. However, this is exactly where smaller corpora are at a disadvantage: unless we have a corpus that is balanced for the phenomenon under study, there is no way to exclude independent factors from intervening in the frequencies of individual expressions. To get a feel for the size of corpus that would be required to be able to abstract away from the influence of such independent factors, we extracted two *hate*/*love* datasets from the Corpus of Contemporary American English (COCA, Davies 2008). Similar to the ENPC used by Johansson, COCA is a balanced corpus, but in contrast to the ENPC, COCA has over a billion words and contains over 20 million words for every year from 1990 to 2019 in the same balanced design as the overall corpus. For comparison, we note that the English and Norwegian source text subcorpora of the ENPC each contain between 600k and 700k words.

The first dataset we extracted opposes the frequencies of *hate* and *love* in the years 1992 (23.8m words) and 1993 (24.5m words). What we find is that *hate* and *love* are clearly more frequent in 1993 than in 1992. We checked the differences for each verb with the same log-likelihood test as the one used by McEnery and Xiao (1999), and found that the distribution of the two verbs is significantly different between the two years (*hate*: LL value = 37.54 ($p < 0.001$), *love*: LL value = 37.21 ($p < 0.001$)). What this dataset shows then is that, even in a far bigger corpus than the ENPC, there is no way to guarantee that there are no independent fluctuations in the frequencies of individual expressions. The relevance of this observation lies in the fact that COCA is a monolingual corpus, and therefore the fluctuations cannot be related to translation.

The second dataset we extracted moves to an even higher level of aggregation and opposes the frequencies of *hate* and *love* in the years 1990–1994 to those in the years 1995–1999. Where the data in Table 5 were still concerned with two subcorpora of around 20m words, we now move to two subcorpora with over 100m words (139m and 147m words respectively). The data are presented in Table 6.

Table 5. Frequencies of *hate* and *love* in the years 1992 and 1993 in COCA (relative frequencies per 1m words between brackets).

	Hate	Love
1992	1374 (57)	4671 (196)
1993	1762 (72)	5430 (221)

Table 6. Frequencies of *hate* and *love* in the years 1990–1994 and 1995–1999 in COCA (relative frequencies per 1m words between brackets).

	Hate	Love
1990–1994	7671 (55)	24,656 (177)
1995–1999	8377 (56)	31,979 (216)

The difference we found in Table 5 for *hate* in the 1992 and 1993 subcorpora has clearly become smaller, especially if we were to focus on relative frequencies. It is still significant, though (*hate*: LL value = 4.12 ($p < 0.05$)). For *love*, moving to this higher level of aggregation makes little difference, even in relative frequencies, and the difference between the two subcorpora in Table 6 remains highly significant (*love*: LL value = 584.19 ($p < 0.001$)). This second dataset thus further strengthens our claim that independent fluctuations in the frequencies of individual expressions are difficult to avoid and that these need not be related to translation in any way. We conclude that the comparison between source and translated texts can inform us about the influence of translation, but that this comparison should be handled with care. This holds for big corpora and *a fortiori* for the smaller corpora used in the traditions discussed in this paper. Corpus size is, of course, relative to the phenomenon under study, and lexical phenomena are likelier to require bigger corpora than more grammatical phenomena.

5.3. Corpus Size and Choosing a Representativeness Strategy

Our discussion in Section 5.2 warrants against an overreliance on the comparison between source and translated texts *within* a language to control for the influence of translation. What the data from COCA show is that the variation we find might well be due to factors that have little or nothing to do with translation. The question that imposes itself is how to best deal with this extra complication; in particular, for smaller corpora. The answer—we argue—lies in the extended research design of the *Translation Mining* tradition.

In Section 4.2, we already pointed out that the *Translation Mining* tradition does not rely on one corpus but replicates the parallel vs. monolingual perspective of the Contrastive tradition across studies of multiple parallel corpora with different source languages. The advantage of this approach is that it maintains the parallel vs. monolingual perspective but, at the same time, forces researchers to pay attention to the individual characteristics of each corpus and invites them to systematically reflect on different sources of variation. Predictions based on one corpus are checked on the next, and hypotheses on why predictions are borne out or not are systematically evaluated.

A further research design feature that we did not treat in Section 4.2 but does play a role in teasing apart different types of variation in the *Translation Mining* tradition is related to an architectural feature of its corpora, whose relevance can be best highlighted here. Major corpus compilation projects from the Contrastive tradition include fragments of different source texts from different authors and translations from different translators. Corpora in the *Translation Mining* tradition are markedly different in the sense that they are typically built around a single source novel and one translation per language at a time. The rationale behind this move is that it allows researchers to keep constant as many variables as possible, while actively looking for variation between subparts of the corpus. In Le Bruyn et al. (2019) and van der Klis et al. (2021b), this strategy leads to the opposition between dialogue and narrative discourse in their analysis of the tense use in the first volume of the *Harry Potter* series by J.K. Rowling and its translations to a number of Western European languages. The inclusion of this opposition is a direct consequence of the difference in the use of the HAVE-perfect they found between Chapter 1 and Chapters 16/17. On closer analysis, this difference turned out to be due to the fact that Chapter 1 contains little-to-no dialogue, whereas dialogue abounds in Chapters 16/17. This active search for variation within a single corpus typically turns up variation that is independent of translation, and complements the attention for different types of variation across multiple corpora that we noted above. Together, they allow research within the *Translation Mining* tradition to lead—across multiple studies—to the critical mass required to support claims about cross-linguistic variation, on the one hand, and about the influence of translation, on the other.

6. General Discussion and Conclusions

This paper surveyed the strategies that the Contrastive, Typological and *Translation Mining* parallel corpus traditions rely on to deal with the issue of representativeness, and laid the foundations for future research to come to a motivated choice of strategy for a given dataset.

In Sections 2–4, we compared the corpus architectures and general research designs of the three traditions and concluded that they have each developed their own representativeness strategy: (i) monolingual control corpora (Contrastive tradition), (ii) limits on the scope of research questions (Typological tradition), and (iii) parallel control corpora (*Translation Mining* tradition). In Section 5, we argued that different datasets favor different strategies and zoomed in on the question of whether monolingual or parallel control corpora is to be preferred for datasets contrasting individual items/constructions across two or three languages. We introduced normalized pointwise mutual information (NPMI) as a bi-directional measure of cross-linguistic association that is independent of the architecture of parallel corpora, allowing for an easy comparison of the outcomes of different traditions and the impact of the monolingual and parallel control corpus strategies. We further argued that corpus size has a major impact on the reliability of the monolingual control corpus strategy and submitted that a sequential parallel control corpus strategy might be preferable for smaller corpora.

The variety of representativeness strategies we have surveyed shows that present-day parallel corpus research is very much aware of the need to control for target language representativeness. In this sense, the parallel corpus community has clearly responded to the objections against the use of parallel corpora in cross-linguistic research that were raised by—among others—McEnery and Xiao (1999) and McEnery et al. (2006). At the same time, our discussion also sheds new light on the data that led to these objections. In Section 5, we argued that corpus size impacts on the reliability of the monolingual control corpus strategy, but it crucially also impacts on the reliability of the findings that led McEnery and Xiao (1999) and McEnery et al. (2006) to argue that there are crucial differences between translated and untranslated texts. We remind the reader that McEnery and Xiao (1999) found that the aspect markers *le* and *guo* were twice more frequent in a 35k word monolingual Mandarin corpus than in a 35k word corpus with translated Mandarin texts (cf. Section 1.1). Even though the authors made sure that the two corpora matched in genre, the size of the corpora is simply too small to draw firm conclusions. In line with this view, we note that recent work within translation studies shows that with corpora of about one million words the difference in frequency between translated and untranslated texts goes down to five percent for *le* and drops below one percent for *guo* (Xiao and Hu 2015). Differences thus remain and should be controlled for, but they are significantly more nuanced than the earlier literature suggests.

We conclude that target language representativeness cannot be taken as a given and has to be controlled for in parallel corpus research. However, the fact that multiple strategies are available allows researchers to carry out methodologically sound cross-linguistic research on a variety of parallel datasets and across multiple empirical domains, including tense and aspect. We hope the reflections in this paper inspire parallel corpus research to further explore and develop its representativeness strategies, allowing it to properly assess the scope of its conclusions and go beyond qualifying its findings as merely 'preliminary' or 'qualitative'.

Author Contributions: Conceptualization, B.L.B.; investigation, B.L.B., M.F., M.v.d.K., J.L., C.M., J.T., H.d.S.; writing—original draft preparation, B.L.B.; writing—review and editing, B.L.B., M.F., M.v.d.K., J.L., C.M., J.T., H.d.S.; funding acquisition, B.L.B., H.d.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NWO (Dutch Research Council), grant number 360-80-070.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank the reviewers for their kind and constructive comments. Our thanks also go to Bernhard Wälchli for providing access to the raw data from Wälchli and Cysouw (2012) and to NWO for financial support.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

[1] We note that there is also a vibrant literature on the use of parallel corpora in natural language processing and translation studies. Working out how these literatures relate to the different traditions that are more geared towards linguistic analysis lies beyond the scope of this paper, though.

[2] A reviewer correctly points out that the existence of multiple source texts for the different translations included in Wälchli and Cysouw (2012) complicates the interpretation of the data. We leave this complication aside as it is connected to the bible corpus Wälchli & Cysouw use and not to the overall parallel corpus architecture of the Typological tradition.

[3] We thank Bernhard Wälchli for providing us with the data from Wälchli and Cysouw (2012).

[4] We computed eigenvalues in R with the function *cmdscale()* and the parameter *eig* set to *TRUE*.

[5] We use the term *group of contexts* deliberately here as van der Klis et al. do not run cluster analyses on their data. We refer to van der Klis and Tellings (2022) for an overview of the different ways of running cluster analyses independently or alongside MDS.

References

Altenberg, Bengt. 1999. Adverbial Connectors in English and Swedish: Semantic and Lexical Correspondences. *Language and Computers* 26: 249–68.

Altenberg, Bengt, and Karin Aijmer. 2000. The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies. In *Corpus Linguistics and Linguistic Theory Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20) Freiburg im Breisgau 1999*. Edited by Christian Mair and Marianne Hundt. Amsterdam: Brill, pp. 15–33.

Beekhuizen, Barend, Julia Watson, and Suzanne Stevenson. 2017. Semantic Typology and Parallel Corpora: Something about Indefinite Pronouns. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Edited by Glenn Gunzelmann, Andrew Howes, Thora Tenbrink and Eddy Davelaar. Austin: Cognitive Science Society, pp. 112–7.

Bogaards, Maarten. 2019. A Mandarin map for Dutch durativity: Parallel text analysis as a heuristic for investigating aspectuality. *Nederlandse Taalkunde* 24: 157–93. [CrossRef]

Bogaards, Maarten. 2022. The Discovery of Aspect: A heuristic parallel corpus study of ingressive, continuative and resumptive viewpoint aspect. *Languages* 7: 158. [CrossRef]

Bogaart, Jade, and Heleen Jager. 2020. La variation Étrange Dans L'Étranger. La Competition du Parfait et du Passé Dans les Traductions Néerlandaises de L'Étranger. Bachelor's thesis, Utrecht University, Utrecht, The Netherlands.

Bouma, Gerlof. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* 30: 31–40.

Bremmers, David, Jianan Liu, Martijn van der Klis, and Bert Le Bruyn. 2021. Translation Mining: Definiteness across Languages—A Reply to Jenks (2018). *Linguistic Inquiry*, 1–18. [CrossRef]

Chang, Vincent Wu. 1986. The Particle LE in Chinese Narrative Discourse: An Investigative Description. Ph.D. thesis, University of Florida, Gainsville, FL, USA.

Chu, Chauncey. 1987. The semantics, syntax, and pragmatics of the verbal suffix *zhe*. *Journal of Chinese Language Teachers Association* 22: 1–41.

Corre, Eric. 2022. Perfective marking in the Breton tense-aspect system. *Languages*, 7.

Dahl, Östen, and Viveka Velupillai. 2013. The Perfect. In *The World Atlas of Language Structures Online*. Edited by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology, Chapter 68.

Dahl, Östen, and Bernhard Wälchli. 2016. Perfects and iamitives: Two gram types in one grammatical space. *Letras de Hoje* 51: 325–48. [CrossRef]

Davies, Mark. 2008. The Corpus of Contemporary American English (COCA). Available online: https://www.english-corpora.org/coca/ (accessed on 30 June 2022).

de Swart, Henriëtte, Jos Tellings, and Bernhard Wälchli. 2022. *Not . . . Until* across European Languages. *Languages* 7: 56. [CrossRef]

Frankenberg-Garcia, Ana, and Diana Santos. 2003. Introducing COMPARA, the Portuguese-English parallel corpus. In *Corpora in Translator Education*. Edited by Federico Zanettin, Silvia Bernardini and Dominic Stewart. Manchester: Routledge, pp. 71–87.

Fuchs, Martín, and Paz González. 2022. Perfect-Perfective Variation across Spanish Dialects: A Parallel Corpus Study. *Languages* 7: 166. [CrossRef]

Gehrke, Berit. 2022. Differences between Russian and Czech in the Use of Aspect in Narrative Discourse and Factual Contexts. *Languages* 7: 155. [CrossRef]

Gellerstam, Martin. 1996. Translations as a source for cross-linguistic studies. *Lund Studies in English* 88: 53–62.

Granger, Sylviane, and Marie-Aude Lefer. 2020. Introduction: A two-pronged approach to corpus-based crosslinguistic studies. *Languages in Contrast* 20: 167–83. [CrossRef]

Grønn, Atle, and Arnim von Stechow. 2020. The Perfect. In *The Wiley Blackwell Companion to Semantics*. Edited by Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann and Thomas Ede Zimmermann. Hoboken: John Wiley & Sons, Inc. [CrossRef]

Hansen-Schirra, Silvia, Stella Neumann, and Erich Steiner. 2013. *Cross-Linguistic Corpora for the Study of Translations*. Berlin: De Gruyter.

Hasselgård, Hilde. 2020. Corpus-based contrastive studies: Beginnings, developments and directions. *Languages in Contrast* 20: 184–208. [CrossRef]

Johansson, Stig. 1998a. On the role of corpora in cross-linguistic research. In *Corpora and Cross-Linguistic Research: Theory, Method, and Case Studies*. Edited by Stig Johansson and Signe Oksefjell. Amsterdam: Rodopi, pp. 93–103.

Johansson, Stig. 1998b. Loving and hating in English and Norwegian: A corpus-based contrastive study. In *Perspectives on Foreign and Second Language Pedagogy. Essays presented to Kirsten Haastrup on the Occasion of Her Sixtieth Birthday*. Edited by Dorte Albrechtsen, Birgit Henriksen, Inger M. Mees and Erik Poulsen. Odense: Odense University Press, pp. 93–103.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora*. Amsterdam: John Benjamins.

Lauridsen, Karen. 1996. Text corpora and contrastive linguistics: Which type of corpus for which type of analysis? *Lund Studies in English* 88: 63–72.

Le Bruyn, Bert, Martijn van der Klis, and Henriëtte de Swart. 2019. The Perfect in dialogue: Evidence from Dutch. *Linguistics in the Netherlands* 36: 162–75. [CrossRef]

Le Bruyn, Bert, Martijn van der Klis, and Henriëtte de Swart. 2022. Variation and stability: The present perfect and the tense-aspect grammar of western European languages. In *Beyond Time 2*. Edited by Astrid De Wit, Frank Brisard, Carol Madden-Lombardi, Michael Meeuwis and Adeline Patar. Oxford: Oxford University Press.

Lehmann, Christian. 1990. Towards lexical typology. In *Studies in Typology and Diachrony: Papers Presented to Joseph H. Greenberg on His 75th Birthday*. Edited by William A. Croft, Suzanne Kemmer and Keith Denning. Amsterdam: John Benjamins, pp. 161–85.

Levshina, Natalia. 2022. Semantic maps of causation: New hybrid approaches based on corpora and grammar descriptions. *Zeitschrift für Sprachwissenschaft* 41: 179–205. [CrossRef]

Lu, Wei-Lun, and Arie Verhagen. 2016. Shifting viewpoints: How does that actually work across languages? An exercise in parallel text analysis. In *Viewpoint and the Fabric of Meaning*. Edited by Barbara Dancygier, Wei-lun Lu and Arie Verhagen. Boston/Berlin: De Gruyter Mouton, pp. 169–90.

Lu, Wei-Lun, Arie Verhagen, and I-Wen Su. 2018. A Multiple-Parallel-Text Approach for Viewpoint Research Across Languages. In *Expressive Minds and Artistic Creations: Studies in Cognitive Poetics*. Edited by S. Csábi. Oxford: Oxford University Press, pp. 131–57.

Macken, Lieve, Orphée De Clercq, and Hans Paulussen. 2011. Dutch parallel corpus: A balanced copyright-cleared parallel corpus. *Meta: Journal des Traducteurs/Meta: Translators' Journal* 56: 374–90. [CrossRef]

McEnery, Tony, and Richard Xiao. 1999. Domains, text types, aspect marking and English-Chinese translation. *Languages in Contrast* 2: 211–29. [CrossRef]

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London and New York: Routledge.

Mulder, Gijs, Gert-Jan Schoenmakers, Olaf Hoenselaar, and Helen de Hoop. 2022. Tense and aspect in a Spanish literary work and its translations. *Languages* 7.

Tellings, Jos, and Martín Fuchs. 2021. *Sluicing and Temporal Definiteness*. Manuscript. Utrecht: Utrecht University.

van der Klis, Martijn, Bert Le Bruyn, and Henriëtte 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*. Edited by Mirella Lapata, Phil Blunsom and Alexander Koller. Valencia: Association for Computational Linguistics, pp. 497–502.

van der Klis, Martijn, Bert Le Bruyn, and Henriëtte de Swart. 2021a. A multilingual corpus study of the competition between PAST and PERFECT in narrative discourse. *Journal of Linguistics* 58: 423–57. [CrossRef]

van der Klis, Martijn, Bert Le Bruyn, and Henriëtte de Swart. 2021b. *Reproducing the Implicational Hierarchy of* PERFECT *Use*. Manuscript. Utrecht: Utrecht University.

van der Klis, Martijn, and Jos Tellings. 2022. Multidimensional scaling and linguistic theory. *Corpus Linguistics and Linguistic Theory*. Advance online publication. [CrossRef]

Wälchli, Bernhard. 2010. Similarity semantics and building probabilistic semantic maps from parallel texts. *Linguistic Discovery* 8: 331–71. [CrossRef]

Wälchli, Bernhard, and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50: 671–710. [CrossRef]

Xiao, Richard. 2002. A Corpus-Based Study of Aspect in Mandarin Chinese. Ph.D. thesis, Lancaster University, Lancaster, UK.

Xiao, Richard, and Xianyao Hu. 2015. *Corpus-Based Studies of Translational Chinese in English-Chinese Translation*. Berlin/Heidelberg: Springer.

Xiao, Richard, and Tony McEnery. 2004. *Aspect in Mandarin Chinese*. Amsterdam: John Benjamins.

Yang, Suying. 1995. The Aspectual System of Chinese. Ph.D. thesis, University of Victoria, Victoria, BC, Canada.