Mapping the PERFECT via Translation Mining

Martijn van der Klis
Digital Humanities Lab
Utrecht University
M.H.vanderKlis@uu.nl

Bert Le Bruyn
UiL OTS
Utrecht University
B.S.W.LeBruyn@uu.nl

Henriëtte de Swart
UiL OTS
Utrecht University
H.deSwart@uu.nl

Abstract

Semantic analyses of the PERFECT often defeat their own purpose: by restricting their attention to 'real' perfects (like the English one), they implicitly assume the PERFECT has predefined meanings and usages. We turn the tables and focus on form, using data extracted from multilingual parallel corpora to automatically generate semantic maps (Haspelmath, 1997) of the sequence 'HAVE/BE + past participle' in five European languages (German, English, Spanish, French, Dutch). This technique, which we dub *Translation Mining*, has been applied before in the lexical domain (Wälchli and Cysouw, 2012) but we showcase its application at the level of the grammar.

1 Introduction

The PERFECT is a diachronically and linguistically unstable category (Lindstedt, 2000) and is subject to widespread cross-linguistic variation. We zoom in on the HAVE PERFECT that Dahl and Velupillai (2013) trace back to a transitive possessive construction, and manifests itself mainly in Western European languages. Despite extensive literature on the PERFECT, the goal of providing a proper semantics has not been reached (Ritz, 2012).

We propose to use semantic maps (Haspelmath, 1997) for this purpose. Semantic maps are geographical layouts that graphically represent how meanings of grammatical functions are related to each other. While current formal semantic approaches to the PERFECT (e.g. Portner (2003)) are driven by sets of predefined usages exemplified by prototypical instantiations, we aim to generate semantic maps directly from data.

We believe multilingual parallel corpora are an excellent source for this. Translation equivalents provide us with form variation across languages in contexts where the meaning is stable. Parallel corpora have been frequently used in the domain of lexical semantics (e.g. Dyvik (1998)). We showcase a method (adapted from Wälchli and Cysouw (2012)) to create semantic maps directly from multilingual parallel corpora, and adapt it to the level of grammar. We focus on a set of five European languages (German, English, Spanish, French, Dutch), although the methodology can easily be adapted to include more languages.

Linguists commonly distinguish the three core PERFECT meanings in (1):

(1) a. Mary has visited Paris.
(*her past visit is relevant now*) [experiential]
b. Mary has moved to Paris.
(*she currently lives in Paris*) [resultative]
c. Mary has lived in Paris for five years (now).
(*she moved there five years ago*) [continuative]

The resultative meaning in (1b) is thought to constitute the core of the PERFECT. However, (2) (taken from the subtitles of "Body of Proof") shows that the same meaning of a past event and a result with current relevance can be conveyed by a PAST, PERFECT or PRESENT.

(2) a. In case you hadn't noticed, we just got a confession. [en-PAST]
b. Falls es ihnen entging, er hat gestanden.
If it you escaped, he has confessed.
[de-PERFECT]
c. Si vous ne l'avez pas remarqué, on a
If you not it have noticed, we have
des aveux.
confessions.
[fr-PRESENT]

Taking (1) as a starting point for cross-linguistic variation, and ignoring other tense-aspect forms (as in (2)) would lead to a skewed view on variation and on the PERFECT itself. As Ritz (2012)

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 497–502,
Valencia, Spain, April 3-7, 2017.

states, the PERFECT is the 'shapeshifter' of tense-aspect categories, and adapts its meaning to fit into a given system. Our final goal is to provide a compositional semantics of the PERFECT across languages that takes the variation in (2) and (2) into account. The competing, form-based methodology that we outline in the next section constitutes the stepping stone that enables us to reach this goal.

2 Methodology

To construct semantic maps directly from data extracted from multilingual parallel corpora, we apply an existing method in the lexical domain (Wälchli and Cysouw, 2012) at the level of grammar. We dub our method *Translation Mining*. In the following paragraphs, we lay out the method in detail.

2.1 Step 1) Extraction of PERFECTs

In the first phase, we extract fragments containing verbs phrases that match the 'HAVE/BE + past participle' pattern from the EuroParl corpus (Tiedemann, 2012). To do so, we modify an existing algorithm by van der Klis et al. (2015), that takes care of three difficulties in extracting these forms from corpora: (1) words between the auxiliary verb and the past participle, (2) lexical restrictions for BE in French, German and Dutch and (3) a reversed order in subordinate clauses in German and Dutch.

The algorithm searches each of the five languages under investigation (German, English, Spanish, French and Dutch) for PERFECTs and will then return the aligned sentences in the other languages. This yields five-tuples of fragments consisting of at least one PERFECT. Note that this approach is necessary to find the triplet in (2), because only in German a PERFECT is involved. This scheme therefore allows for competing forms within a language to enter the realm of investigation. Also, taking five languages into account will create a broader perspective on the semantics of the PERFECT than monolingual research would do.[1]

2.2 Step 2) Word-level alignment of verb phrases

After extracting fragments containing a PERFECT in step 1, we asked a single human annotator (a BSc student proficient in all languages under investigation) to mark the corresponding verb phrases in the aligned fragments. To facilitate the annotator's job we created a web application (dubbed *TimeAlign*) that allows users to see two aligned fragments (a "source" and a "translation") and to mark the corresponding verb phrase in the target language.[2] The annotator can also signal

[1]The source code of this algorithm can be found on GitHub: https://github.com/UUDigitalHumanitieslab/perfectextractor.

[2]The source code of this application can be found on GitHub: https://github.com/UUDigitalHumanitieslab/timealign. The application has been built in Django, a Python web framework (https://www.djangoproject.com/).

Figure 1: The annotation interface used in step 2. The annotator can select (by clicking on words) a suitable translation for the marked words in the source fragment, or use the checkboxes to mark the source as not being a PERFECT or as the translated fragment as an incorrect translation of the source fragment.

Generic tense	DE	EN	ES	FR	NL
PERFECT	Perfekt	present perfect present perfect continuous	pretérito perfecto compuesto	passé composé	vtt
PRESENT	Präsens	present	presente	présent	ott
PAST	Präterium	simple past	pretérito imperfecto pasado receinte pretérito perfecto simple	imparfait passé récent	ovt
PAST PERFECT	Plusquamperfekt	past perfect	pretérito pluscuamperfecto	plus-que-parfait	vvt
OTHER	Futur I/II	-	participio	futur antérieur	-

Table 1: Possible tenses in step 3 for each language, categorized in a more generic tense category. We also allow to attribute 'other' if none of the tenses fit.

when the target fragment is not a correct translation of the source, or when the verb phrase in the source is in fact not a PERFECT (see Figure 1).

Fragments without a PERFECT in the source and incorrect translations are removed from the dataset. The remaining pairs are merged back into five-tuples. Step 2 thus yields five-tuples of verb phrases, at least one of which (the source) is a PERFECT.

2.3 Step 3) Tense attribution

In the third step, we assign tenses to the verb phrases marked in the translations (see step 2). For the tense labelling, we opted for the categories displayed in Table 1. The tenses are automatically or manually assigned, depending on the level of detail of part-of-speech tags per language. The tense attribution for English, French and Dutch is straightforward: we used the part-of-speech tagging of the EuroParl corpus to retrieve the label.[3] However, for German and Spanish we opt for manual tense attribution, because the part-of-speech-tagging of the auxiliary verbs in EuroParl was too coarse-grained.

2.4 Step 4) Dissimilarity matrix

The tense attribution process of step 3 yields five-tuples of aligned tense attributions (see Table 2 for an example outcome). We design a simple distance function: we define five-tuples to be similar (distance = 0) if all the tense attributions match up, if not, we add 1 for each mismatch and divide the sum by 5. We use the distance function on the five-tuples to create a (dis)similarity matrix. Table 3 shows an application of the distance function and the resulting matrix.

We decided to remove five-tuples from the results in which one of the translations was missing or contained a non-verbal translation. We believe including these examples in the current pilot study, with a limit dataset and only five languages in total, would have a negative effect on our analyses. We will address this issue in future research.

[3]The source code of this algorithm is part of TimeAlign, see link above.

#	DE	EN	ES	FR	NL
1	Perfekt	present perf.	passé comp.	pretérito perf. comp.	vtt
2	Präterium	simple past	passé comp.	pretérito perf. comp.	vtt
3	Perfekt	present perf.	passé récent	pasado receinte	vtt

Table 2: Example set of tense attributions.

	#1	#2	#3
#1	-	2/5	2/5
#2	2/5	-	4/5
#3	2/5	4/5	-

Table 3: Dissimilarity matrix for the example tense attributions in Table 2.

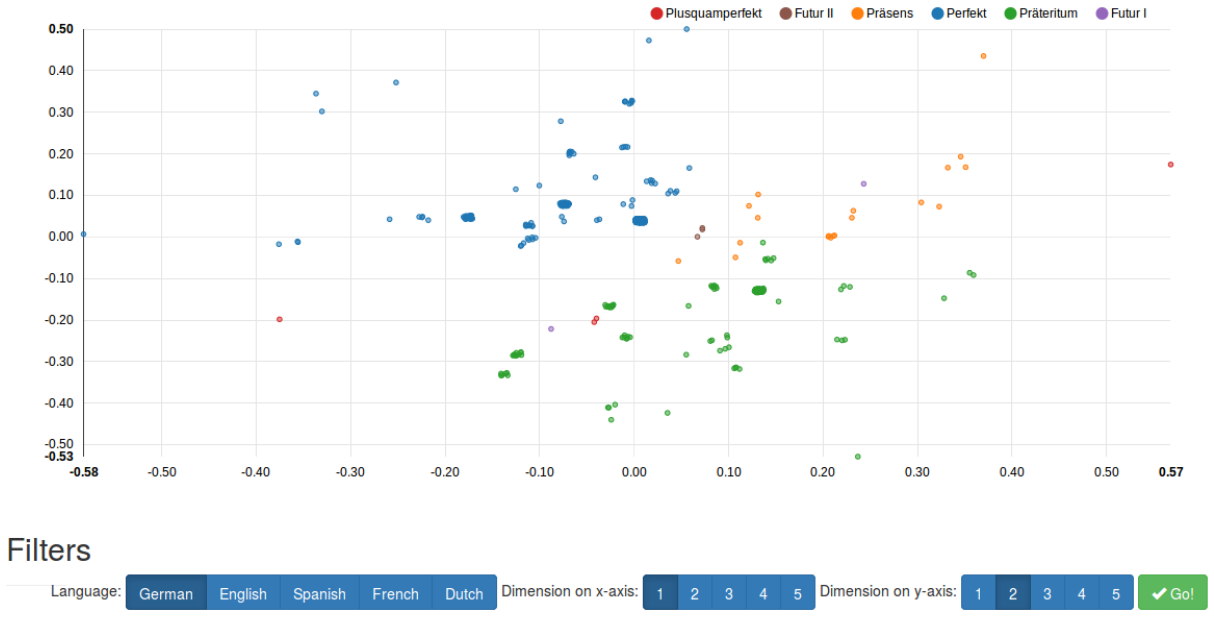

Figure 2: Visualization of the dissimilarity matrix via multidimensional scaling. The points are labelled using the tenses of the selected language. Users can also change the dimensions shown. Clicking on a point allows to inspect a single five-tuple (example shown in Figure 3).

2.5 Step 5) Visualization via multidimensional scaling

The resulting matrix from step 4 is then plotted using multidimensional scaling (MDS)[4]. On top of that, we created an interactive visualization (see Figure 2).

This visualization shows which space the various tenses (PERFECT and other) occupy on the map, and thus enables researchers to see how tenses interact within a language. The visualization also allows for comparison between languages, because it uses a color labeling that remains constant between languages (e.g. the German *Perfekt* has the same color as the English *present perfect*). Furthermore, being able to filter tenses allows to focus on one specific tense or interaction between specific tenses. The researcher can also choose to show other dimensions of the MDS algorithm, which facilitates interpretation. Hovering over a point on the map directly shows you the five-tuple the point is based on, and clicking on a point will yield a new page in which you can inspect the underlying data (see Figure 3 for an example).[5]

Compared to Wälchli and Cysouw (2012), our main contributions in this methodology are (1) the web application to allow for easier annotation and (2) the interactive visualization of the MDS algorithm, which allows for researchers to more easily compare PERFECT usage within and across languages, as well as interpret dimensions.

[4]We use the MDS algorithm from the *scikit-learn* package (Pedregosa et al., 2011), a Python package for machine learning, and visualized the results using the *nvd3* package (http://nvd3.org/).

[5]The source code of this visualization is part of TimeAlign, see link above.

	DE	EN	ES	FR	NL
PERFECT	360	347	371	481	438
PRESENT	19	18	47	20	20
PAST	124	146	89[7]	8[8]	36
PAST PERFECT	4	1	3	2	18
other	5	-	2	1	-

Table 4: Descriptive statistics of tense attributions in all five languages.

3 Premilinary results

In this pilot study we analyzed a small part of the Q4/2000 portion of the EuroParl corpus.[6] Running the *Translation Mining* methodology on this corpus yielded 512 complete five-tuples in total.

We first observe the descriptive statistics in Table 4 that result from mapping the language-specific tense labelling in step 3 to more generic tenses (e.g. *simple past* to PAST, see Table 1). As is commonly reported in literature (see de Swart (2007) and references therein), the French *passé composé* takes responsibility for a wide range of PERFECT uses. In German and English one tends to use PAST for quite a few contexts where French would use the *passé composé*. In Spanish, the *presente* also competes with the PAST in this respect.

[6]Specifically, the files 00-12-11.xml, 00-12-14.xml and 00-12-15.xml, totaling 106k tokens for the English translation.

[7]This consists of 79 fragments labelled as *préterite perfecto simple*, 6 as *pasado reciente* and 4 as *préterite imperfecto*.

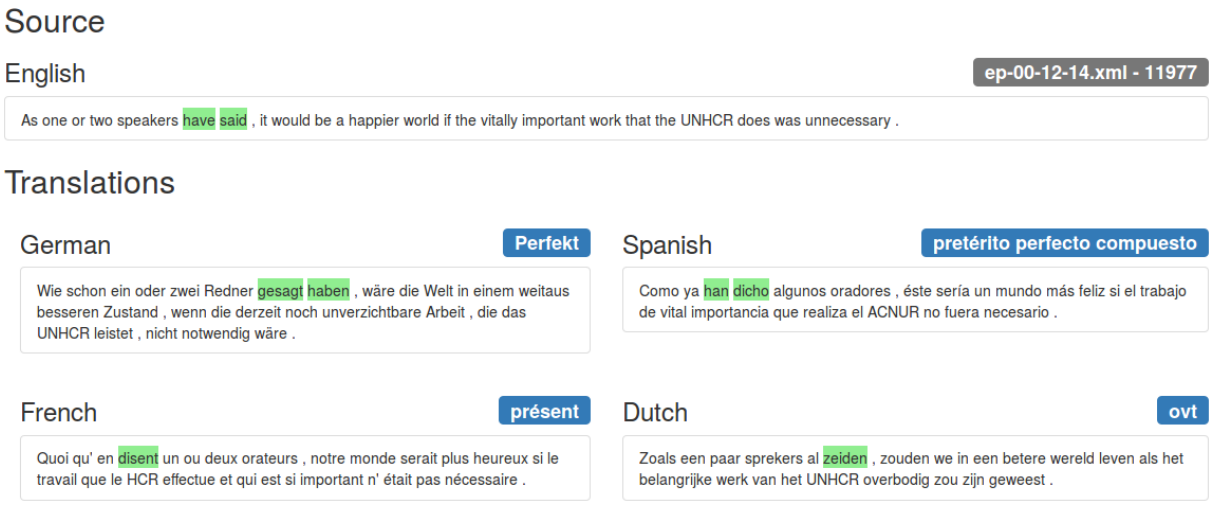

Figure 3: Detailed view of a five-tuple of fragments. The "source" fragment shows the extracted sentence from step 1 with the PERFECT marked in green. The "translations" are the aligned fragments with manually annotated verb phrases from step 2 and semi-automatically annotated tenses from step 3.

Moving from descriptive statistics to the MDS visualization, we look at dimensions governing the competition between languages. The German data (depicted in Figure 2) is most obvious in this respect, where the x-axis shows a transition from PERFECT to unmarked (aspectual perspective), and the y-axis from PRESENT to PAST (temporal orientation). However, this attribution is not so easily translated into other languages, even though in each language we do find clear clusters of PERFECT use.

In the visualization, we can also look at outliers to find cases where one language is different from the other languages. We can confirm e.g. that English requires a PAST with a locating time adverbial, whereas German, Dutch and French tolerate a PERFECT in this configuration. Spanish patterns with English (see Schaden (2009)) in this respect. An example of this phenomenon can be found in (3) below.

(3) a. [de] Frau Präsidentin, wir **haben** am 4. Dezember **abgestimmt**.

b. [en] Madam President, we **voted** on 4 December.

c. [es] Señora Presidenta, **votamos** el pasado 4 de diciembre.

d. [fr] Madame la Présidente, nous **avons voté** le 4 décembre.

e. [nl] Mevrouw de Voorzitter, op 4 december **hebben** wij hierover **gestemd**.

Another interesting outlier is the RECENT PAST, available for French and Spanish. This periphrastic tense signals recency and is expressed in German, English and Dutch through the use of a PERFECT combined with an additional time adverbial: *gerade*, *just*, *kortgeleden* respectively, see (4) below. A tentative conclusion could be that the RECENT PAST is a dimension of the PAST or of the PERFECT, but in both cases this recency requires additional marking.

(4) a. [de] Der Gerichtshof **hat** nämlich *gerade* die Richtlinie aus dem Jahr 1998, die Werbung und Sponsoring fr Tabakerzeugnisse verbietet, **aufgehoben**.

b. [en] The fact is that the Court of Justice **has** *just* **repealed** the 1998 Directive banning advertising and sponsorship of tobacco products.

c. [es] El Tribunal de Justicia, efectivamente, **acaba de anular** la directiva de 1998 que prohibía la publicidad y el patrocinio de los productos del tabaco.

d. [fr] La Cour de justice, en effet, **vient d'annuler** la directive de 1998 interdisant la publicité et le parrainage en faveur des produits du tabac.

e. [nl] Het Hof van Justitie **heeft** *kortgeleden* de richtlijn van 1998 betreffende het verbod op reclame en sponsoring in de tabakssector **geannuleerd**.

[8]This consists of 7 fragments labelled as *passé recent* and 1 as *imparfait*.

4 Discussion

The interactive maps allowed us to reproduce earlier research (e.g. de Swart (2007), Schaden (2009)), but also to draw new conclusions on the tense/aspect role of the PERFECT across languages. Our methodology can be applied to a wide range of grammatical phenomena. There are some remaining issues though.

First of all, interpreting the results of the MDS algorithm is more qualitative than quantitative. While the visualization helps researchers to form ideas on the role of the PERFECT, these intuitions will need to be supported by statistics. We are currently looking into applying Analysis of Similarities (ANOSIM, Clarke (1993)) on the (dis)similarity matrices to pair this with the MDS visualization.

A second limitation is that the EuroParl corpus contains only political dialogue, and therefore might not cover the whole range of PERFECT use. We should also check for register variation. Our plan is to repeat our methodology on the OpenSubtitles2016 corpus (Lison and Tiedemann, 2016), as well as to find (or create) a multilingual parallel corpus of literary texts.

Lastly, we think the distance function we now use might be too simplistic. It considers all tense differences to be equal, even though it is quite clear that e.g. a PRESENT is semantically more distant from a PAST PERFECT than a PERFECT. Also, there is no cross-language comparison. We plan to experiment with the distance function to finetune our results.

References

K. R. Clarke. 1993. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18(1):117–143.

Östen Dahl and Viveka Velupillai. 2013. The perfect. In Martin Haspelmath, editor, *The World Atlas of Language Structures Online*.

Henriëtte de Swart. 2007. A cross-linguistic discourse analysis of the perfect. *Journal of pragmatics*, 39(12):2273–2307.

Helge Dyvik. 1998. A translational basis for semantics. In Stig Johansson and Signe Oksefjell, editors, *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*, pages 51–86. Rodopi, Amsterdam.

Martin Haspelmath. 1997. *Indefinite pronouns*. Clarendon Press, Oxford.

Jouko Lindstedt. 2000. The perfect – aspectual, temporal and evidential. In Ö. Dahl, editor, *Tense and Aspect in the languages of Europe*, pages 365–384. De Gruyter, Berlin.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Paul Portner. 2003. The (temporal) semantics and (modal) pragmatics of the perfect. *Linguistics and Philosophy*, 26(4):459–510.

Marie-Eve Ritz. 2012. Perfect tense and aspect. In Östen Dahl, editor, *The Oxford Handbook of Tense and Aspect*, pages 881–907. Oxford University Press, Oxford.

Gerhard Schaden. 2009. Present perfects compete. *Linguistics and Philosophy*, 32(2):115–141.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1246.

Martijn van der Klis, Bert Le Bruyn, and Henriëtte de Swart. 2015. Extracting present perfects from a multilingual corpus. Presentation at Computational Linguistics in the Netherlands 26.

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