Mapping the PERFECT via Translation Mining

Martijn van der Klis, Bert Le Bruyn, Henriëtte de Swart

CLIN 27, Leuven, 10 February 2017



UiL OTS
UTRECHT INSTITUTE OF LINGUISTICS

NWO

The PERFECT – the basics

- Morpho-syntax: *auxiliary (HAVE/BE) + past participle*
- Appears a.o. in German, English, Spanish, French, Dutch
- Core meaning: *past event + present state*.

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

The PERFECT – current approaches

- Current approaches tend to look at the PERFECT in **isolation**:
  - They ignore variation within a language:
    - Competition with other tense-aspect forms in the grammar (in particular PAST and PRESENT)
  - They ignore variation between languages:
    - Same meaning can be conveyed by other tense-aspect forms

The PERFECT – our proposed analysis

- **Semantic maps** (Haspelmath 1997)
  - Allows to see variation within and between languages
- **But**: generate these directly from the data
- **Final goal**: compositional semantics of the PERFECT across languages

Data – multilingual parallel corpora

- Translation equivalents provide us with ***form variation*** across languages in contexts where the ***meaning is stable***.
- Genre variation through different corpora
  - This pilot: EuroParl (Tiedemann 2012)
  - OpenSubtitles (Tiedemann 2016)
  - Literary corpora (following de Swart 2007)

Translation Mining

- **Extraction of PERFECTS** from multilingual corpora
- **Word-level alignment** of verb phrases
- **Tense attribution** for translations
  - Discard 'other' translations (nominalizations, 'free' translations)
- Create a **dissimilarity matrix**
- **Multidimensional scaling** to project variation onto a semantic map

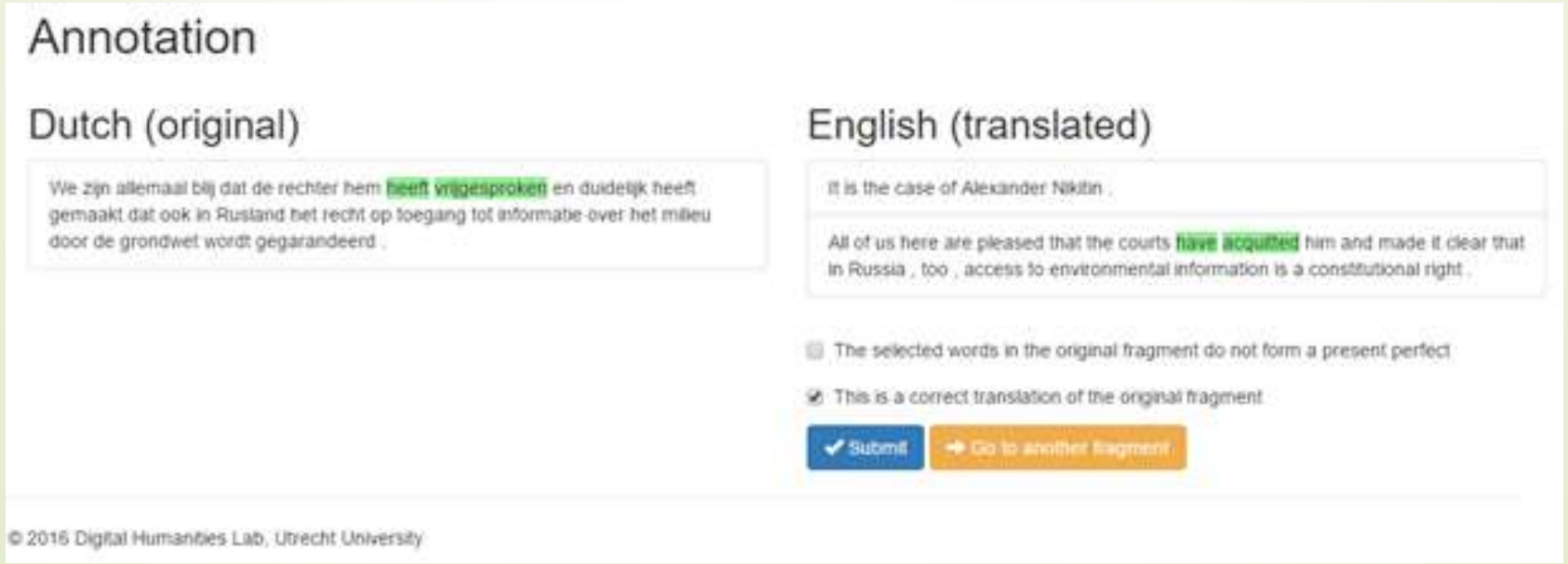
- *Methodology inspired on Wälchli & Cysouw 2012*

Step 1) Extraction of PERFECTS

- Stand-alone application: **PerfectExtractor**
- Simple principle:
  - Look for an auxiliary verb (HAVE/BE)
  - Find a neighboring past participle
- Takes care of words in between, lexical restrictions, reversed order, passive present perfects.
- Uses config files to discern between languages and corpora
- Source code on GitHub: https://github.com/UUDigitalHumanitieslab/time-in-translation

Step 2) Word-level alignment

- Web application: **TimeAlign**
- Allows for manual alignment of translations of PERFECTS.
- Source code on GitHub: https://github.com/UUDigitalHumanitieslab/timealign



Step 3) Tense attribution

- To allow for calculating distances between fragments, we need to assign a tense to every selected translation.
  - **Algorithm** for English, French and Dutch
  - **Manually** for German and Spanish
    - POS-tags do not discern between PRESENT and PAST tense for auxiliary verbs

Step 4) Dissimilarity matrix

- **Distance function**: we define a pair of source and translations to be maximally similar (d=0) if all tenses match up.

#	German	English	French	Spanish	Dutch
1	Perfekt	Present perfect	Passé composé	Pretérito perfecto conpuesto	Voltooid tegenwoordige tijd
2	Präterium	Simple past	Passé composé	Pretérito perfecto conpuesto	Voltooid tegenwoordige tijd
3	Perfekt	Present perfect	Passé récent	Pasado receinte	Voltooid tegenwoordige tijd

- $d(1, 2) = 2/5$, $d(1, 3) = 2/5$, $d(2, 3) = 4/5$

Step 4) Dissimilarity matrix

- Using our defined distance function, we can create a dissimilarity matrix.

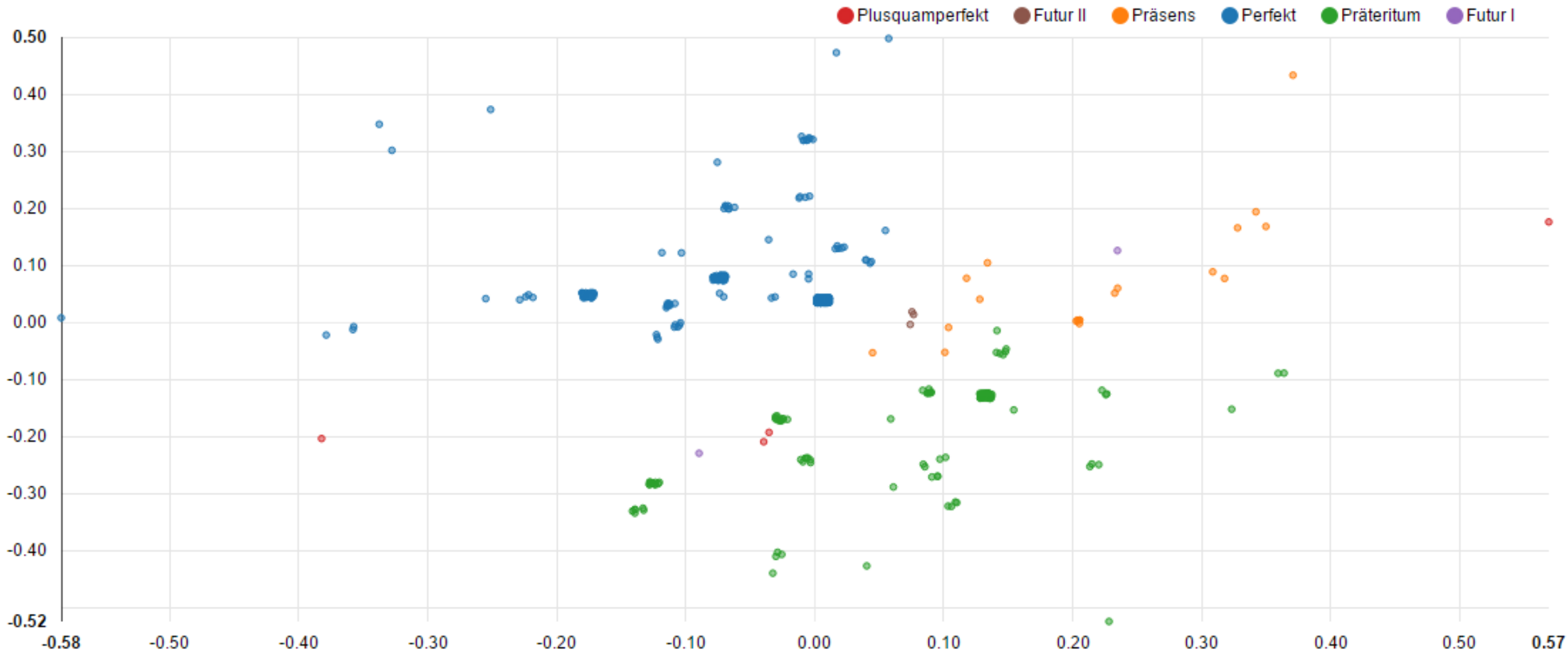
	#1	#2	#3
#1	-	2/5	2/5
#2	2/5	-	4/5
#3	2/5	4/5	-

Step 5) Multidimensional Scaling

- **Multidimensional scaling (MDS)** tries to create a low-dimensional representation of the data, while respecting distances in the original high-dimensional space.

Step 5) Multidimensional Scaling

- Web application: **TimeMapping**
  - Visualisazes the results of MDS on a scatter plot, we use the attributed tenses as labels.
- Friendly user interface:
  - Click/unclick tenses for filtering
  - Choose which dimensions of the MDS to show
  - Choose which language to show
  - Drill-through to raw data
- Source code on GitHub: https://github.com/UUDigitalHumanitieslab/timealign



Filters

Language: German English Spanish French Dutch

Dimension on x-axis: 1 2 3 4 5

Dimension on y-axis: 1 2 3 4 5

✔ Go!

Interpreting the maps

- Switching dimensions reveals two main contrasts:
  - Temporal orientation: PRESENT vs. PAST (high/low)
  - Aspectual perspective: PERFECT vs. SIMPLE (left/right)
- We are working on statistical analysis using ANOSIM (Clarke 1993) or non-parametric MANOVA (Anderson 2001)

Qualitative research (1)

- The maps show more **central** and more **peripheral** uses of the PERFECT
- We can investigate **outliers**
  - E.g. confirm 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). Example on next page.

Dutch

ep-00-12-11.xml - 3552

Mevrouw de Voorzitter , op 4 december hebben wij hierover gestemd .

Translations

German

Perfekt

Frau Präsidentin , wir haben am 4. Dezember abgestimmt .

English

simple past

Madam President , we voted on 4 December .

Spanish

presente

Señora Presidenta , votamos el pasado 4 de diciembre .

French

passé composé

Madame la Présidente , nous avons voté le 4 décembre .

Qualitative research (2)

- We can look at **special verb forms**, and find out how other languages deal with them in a different tense-aspect grammar.
  - E.g. English *Present Perfect Continuous* has been claimed to translate as a PRESENT in French (cf. Nishyama & Koenig 2010) and Dutch.
  - But: German maintains a PERFECT in sentences that contain a *seit* ('since') adverbial. Note the special construction in Spanish.

German

ep-00-12-14.xml - 10491

Seit Monaten haben die Nachrichten von der Elfenbeinküste eine begründete Unruhe in der europäischen Öffentlichkeit und entsprechende Besorgnis in unserem Parlament hervorgerufen .

Translations

English

present perfect continuous

Mr President , for months the news from Côte d ' Ivoire has been causing justified alarm within European public opinion and the corresponding concern in our Parliament .

Spanish

pretérito perfecto compuesto

Señor Presidente , desde hace meses las noticias de Costa de Marfil han venido causando justificada alarma en la opinión pública europea y la correspondiente preocupación en nuestro Parlamento .

French

présent

Monsieur le Président , les nouvelles de la Côte-d'Ivoire suscitent depuis des mois l' inquiétude justifiée chez les citoyens européens et l' inquiétude correspondante au sein de notre Parlement .

Dutch

ott

Mijnheer de Voorzitter , sinds enkele maanden zorgen de berichten uit Ivoorkust terecht voor onrust in de Europese publieke opinie en voor overeenkomstige bezorgdheid in ons Parlement .

Qualititative research (3)

- Another **special construction** is the RECENT PAST in French/Spanish.
- Where French and Spanish use a RECENT PAST verb form, English, German and Dutch use a PERFECT.
- Languages who use a PERFECT to refer to a recent past use an additional time adverbial: *just, gerade, kortgeleden*.
- Tentative conclusion: RECENT PAST can be a dimension of the PAST or of the PERFECT, but in both cases recency requires additional marking.

Dutch

ep-00-12-11.xml - 3661

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

Translations

German

Perfekt

Der Gerichtshof hat nämlich gerade die Richtlinie aus dem Jahr 1998 , die Werbung und Sponsoring für Tabakerzeugnisse verbietet , aufgehoben .

English

present perfect

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

Spanish

pasado reciente

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

French

passé récent

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 .

Future research

- **Lexical annotation**: Annotate verbs for aspectual class, identify locating time adverbials, identify aspectually sensitive adverbials (*since, for, already, just, always,* negation), and investigate interaction with PAST/PRESENT/PERFECT.
- **Discourse annotation**: Investigate narrative and non-narrative tense use of the PERFECT.
- **Syntactic annotation**: Investigate the interaction of (i) voice (active/passive, see CLIN 2015 presentation) and (ii) clause type (main/subordinate, e.g. conditional) with PAST/PRESENT/PERFECT.

Thanks for your attention!

- Any questions or feedback?
- More on our research programme on our website: http://bit.ly/timeintranslation