Stative verbs as edge cases in the Perfect construction

Martijn van der Klis    Bert Le Bruyn    Henriëtte de Swart

UiL OTS, Utrecht University

August 31, 2018
Intro - the **Perfect** construction

**Morpho-syntax**: *auxiliary (HAVE) + past participle*

*Examples:*
- Mary *has read* Camus.
- Appears in most West-European languages
- Core meaning: past event with current relevance
Intro - the **Perfect** construction

Morpho-syntax: *auxiliary* (HAVE) + *past participle*

- e.g. Mary **has read** Camus.

  - Appears in most West-European languages
  - Core meaning: past event with current relevance

But: distribution of **Perfect** differs between languages, e.g.:

(1) **Was hast du gemacht?**
*lit.* *What have you done?*
Intro - the **Perfect** construction

Morpho-syntax: *auxiliary (HAVE) + past participle*

* e.g. Mary **has read** Camus.
  * Appears in most West-European languages
  * Core meaning: past event with current relevance

But: distribution of **Perfect** differs between languages, e.g.:

(1) **Was hast du gemacht?**
  * lit. What **have you done?**
  * What **did you do?**
Intro -stative verbs in the Perfect

From English, we know states are special: in the Perfect, they raise a (potential) continuative reading. However, this is language-specific (de Swart, 2016):

Mary has lived in Tallinn for five years.


Marie lebt seit fünf Jahren in Tallinn
Intro -stative verbs in the Perfect

From English, we know states are special: in the Perfect, they raise a (potential) continuative reading. However, this is language-specific (de Swart, 2016):

(2) Mary has lived in Tallinn for five years.

*lit. Marie hat seit fünf Jahren in Tallinn gelebt.*
From English, we know states are special: in the Perfect, they raise a (potential) continuative reading. However, this is language-specific (de Swart, 2016):

(2) Mary **has lived** in Tallinn for five years.

*lit.* Marie *hat seit fünf Jahren in Tallinn gelebt.*

Marie **lebt** seit fünf Jahren in Tallinn
Intro - stative verbs in the **Perfect**

Stative verbs also appear less frequent in the **Perfect** per se. Evidence comes from a collostructional analysis on the British National Corpus, fiction section (van der Klis, 2018):
Intro - stative verbs in the Perfect

Stative verbs also appear less frequent in the Perfect per se. Evidence comes from a collostructional analysis on the British National Corpus, fiction section (van der Klis, 2018):

- **Attracted verbs**
  - hear
  - see
  - happen
  - change
  - finish

- **Repelled verbs**
  - feel
  - want
  - know
  - think
  - say
Schaden (2009) postulates a dichotomy for the Perfect in the European languages:

- English and Spanish are alike:
  - Not licensed with past time adverbial
    (* I have read a book yesterday)
  - Not licensed in narrative context
    (* ...and then I have seen Mary)
Intro - cross-linguistic variation

Schaden (2009) postulates a dichotomy for the Perfect in the European languages:

▶ English and Spanish are alike:
  • Not licensed with past time adverbial
    (* I have read a book yesterday)
  • Not licensed in narrative context
    (* ...and then I have seen Mary)

▶ French and German are alike:
  • Can appear with past time adverbial
  • Can appear in narrative context
Intro - cross-linguistic variation

Schaden (2009) postulates a dichotomy for the **Perfect** in the European languages:

- English and Spanish are alike:
  - Not licensed with past time adverbial
    (* I have read a book *yesterday*)
  - Not licensed in narrative context
    (* ...and then I have seen Mary*)

- French and German are alike:
  - Can appear with past time adverbial
  - Can appear in narrative context

Main questions in this talk:

- Are this indeed the important differences? Do we find differences between these language pairs?
- Is this a dichotomy, a scale, or something else? What about other languages?
Our data

We use parallel corpora: translation equivalents provide us with **form variation** across languages in contexts where the meaning is stable.

Advantages:
▶ more down-to-earth register (as opposed to Bible/Europarl)
▶ high quality translations (as opposed to OpenSubtitles)
▶ allows to study phenomena in dialogue vs. discourse

Disadvantages:
▶ small datasets (so frequent phenomena required)
▶ possible translator effects
▶ copyright issues
Our data

We use parallel corpora: translation equivalents provide us with **form variation** across languages in contexts where the **meaning is stable**.

While there are several alternatives available, we opt to use translations of novels.
Our data

We use parallel corpora: translation equivalents provide us with **form variation** across languages in contexts where the **meaning is stable**.

While there are several alternatives available, we opt to use translations of novels.

Advantages:

▶ more down-to-earth register (as opposed to Bible/Europarl)
▶ high quality translations (as opposed to OpenSubtitles)
▶ allows to study phenomena in dialogue vs. discourse
Our data

We use parallel corpora: translation equivalents provide us with **form variation** across languages in contexts where the **meaning is stable**.

While there are several alternatives available, we opt to use translations of novels.

Advantages:

- more down-to-earth register (as opposed to Bible/Europarl)
- high quality translations (as opposed to OpenSubtitles)
- allows to study phenomena in dialogue vs. discourse

Disadvantages:

- small datasets (so frequent phenomena required)
- possible translator effects
- copyright issues
Our data - L’Étranger parallel corpus

We use the first three chapters of Albert Camus’ novel L’Étranger as our data. Why?

- internal monologue: passé composé can be used in French, but not in most other languages: PERFECT stretched to its max
- confirmation of earlier research (e.g. de Swart (2007))
Our data - L’Étranger parallel corpus

We use the first three chapters of Albert Camus’ novel *L’Étranger* as our data. Why?

- **internal monologue**: *passé composé* can be used in French, but not in most other languages: **PERFECT** stretched to its max

- confirmation of earlier research (e.g. de Swart (2007))

Languages in L’Étranger corpus:

- **Romance** French, Italian, Spanish
- **Germanic** German, Dutch, English
- **other** Breton, Estonian, Farsi, Greek, Hebrew, Mandarin, Russian
Annotation

Annotation is done in a web application (dubbed *TimeAlign*). Steps:

1. algorithm extracts all *passé composé* forms automatically
2. annotators select corresponding verb forms in translation
3. annotators assign tense-aspect labels (so e.g. *simple past* or *Perfekt*)
<table>
<thead>
<tr>
<th>language</th>
<th>fragment</th>
<th>TA-label</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>fr</td>
<td>Aujourd’hui, maman est morte.</td>
<td>passé composé</td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>Heute ist Mama gestorben.</td>
<td>Perfekt</td>
<td></td>
</tr>
<tr>
<td>nl</td>
<td>Vandaag is moeder gestorven.</td>
<td>voltooid tegenwoordige tijd</td>
<td></td>
</tr>
<tr>
<td>es</td>
<td>Hoy, mamá ha muerto.</td>
<td>pretérito perfecto compuesto</td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>Mother died today.</td>
<td>simple past</td>
<td></td>
</tr>
</tbody>
</table>
We refer to a TA-labeling for a single fragment as a **tuple**.
Results - descriptive statistics

Descriptive statistics for all *passé composé*-forms in the first three chapters:

<table>
<thead>
<tr>
<th>Tense</th>
<th>fr</th>
<th>de</th>
<th>nl</th>
<th>es</th>
<th>en</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perfect</strong></td>
<td>375</td>
<td>351</td>
<td>45</td>
<td>19</td>
<td>12</td>
</tr>
<tr>
<td><strong>Past</strong></td>
<td>-</td>
<td>23</td>
<td>325</td>
<td>355</td>
<td>354</td>
</tr>
<tr>
<td><strong>Present</strong></td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><em>other tenses</em></td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>6</td>
</tr>
</tbody>
</table>

Note:

- We discarded ‘other’ translations (nominalizations, periphrastic constructions, etc.) and only considered complete tuples.
As we are dealing with parallel data, we can also count tuple frequencies:

<table>
<thead>
<tr>
<th>de</th>
<th>nl</th>
<th>es</th>
<th>en</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAST</td>
<td>PAST</td>
<td>PAST</td>
<td>PAST</td>
<td>20</td>
</tr>
<tr>
<td>PERFECT</td>
<td>PAST</td>
<td>PAST</td>
<td>PAST</td>
<td>297</td>
</tr>
<tr>
<td>PERFECT</td>
<td>PERFECT</td>
<td>PAST</td>
<td>PAST</td>
<td>25</td>
</tr>
<tr>
<td>PERFECT</td>
<td>PERFECT</td>
<td>PERFECT</td>
<td>PAST</td>
<td>6</td>
</tr>
<tr>
<td>PERFECT</td>
<td>PERFECT</td>
<td>PERFECT</td>
<td>PERFECT</td>
<td>10</td>
</tr>
</tbody>
</table>

All other possible combinations: less than 5 occurrences. This hints at a **subset relation** rather than a dichotomy.
Alternative technique - multidimensional scaling

Wälchli & Cysouw (2012) provide a technique to generate semantic maps directly from parallel corpus data using multidimensional scaling (MDS). We showcase this technique on our data.
We define a tuple to be maximally similar \((d = 0)\) if all tenses match up.

<table>
<thead>
<tr>
<th>#</th>
<th>fr</th>
<th>de</th>
<th>nl</th>
<th>es</th>
<th>en</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Perfect</td>
<td>Perfect</td>
<td>Perfect</td>
<td>Perfect</td>
<td>Perfect</td>
</tr>
<tr>
<td>2</td>
<td>Perfect</td>
<td>Perfect</td>
<td>Perfect</td>
<td>Past</td>
<td>Past</td>
</tr>
<tr>
<td>3</td>
<td>Perfect</td>
<td>Past</td>
<td>Past</td>
<td>Perfect</td>
<td>Perfect</td>
</tr>
</tbody>
</table>

In this table, \(d(1, 2) = 0.4\), \(d(1, 3) = 0.4\) and \(d(2, 3) = 0.8\).
Applying our defined distance function, we can create a dissimilarity matrix:

<table>
<thead>
<tr>
<th></th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>-</td>
<td>0.4</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>#2</td>
<td>0.4</td>
<td>-</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>#3</td>
<td>0.4</td>
<td>0.8</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MDS - dissimilarity matrix

Applying our defined distance function, we can create a dissimilarity matrix:

<table>
<thead>
<tr>
<th></th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>⋯</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>-</td>
<td>0.4</td>
<td>0.4</td>
<td>⋯</td>
</tr>
<tr>
<td>#2</td>
<td>0.4</td>
<td>-</td>
<td>0.8</td>
<td>⋯</td>
</tr>
<tr>
<td>#3</td>
<td>0.4</td>
<td>0.8</td>
<td>-</td>
<td>⋯</td>
</tr>
<tr>
<td>⋮</td>
<td></td>
<td></td>
<td></td>
<td>⋮</td>
</tr>
<tr>
<td>⋮</td>
<td></td>
<td></td>
<td></td>
<td>⋮</td>
</tr>
</tbody>
</table>

We use **multidimensional scaling** (MDS) to visualize this dissimilarity matrix. MDS tries to create a low-dimensional representation of the data, while respecting distances in the original high-dimensional space.
MDS - demonstration

Demo time!
MDS - conclusions

Conclusions from multidimensional scaling:

- **subset relation** - no dichotomy - between Western European languages
- **clear distinctions** between language pairs that were presumed to be close together
MDS - conclusions

Conclusions from multidimensional scaling:

▶ **subset relation** - no dichotomy - between Western European languages
▶ **clear distinctions** between language pairs that were presumed to be close together

Our interpretation of the distinctions:
MDS - zooming in

Let’s zoom in on the difference between French/Italian and German:

- French can have stative verbs in the **Perfect** tense
- (cognitive) stative verbs go with a **Past** tense in German (and nl/es/en)

Example:

(3) **fr** J’ai voulu voir mamam tout de suite.
    **de** Ich **wollte** sofort zu Mama.
    **nl** Ik **wilde** moeder meteen zien.
    **es** Yo **quería** ver a mamá inmediatamente.
    **en** I **wanted** to see mother straight away.
Let’s annotate all verbs for stativity (Maienborn, 2015); and compare with the tense choice in German:

<table>
<thead>
<tr>
<th>Verbs</th>
<th>Präteritum</th>
<th>Perfekt</th>
</tr>
</thead>
<tbody>
<tr>
<td>dynamic</td>
<td>6</td>
<td>345</td>
</tr>
<tr>
<td>static</td>
<td>19</td>
<td>23</td>
</tr>
</tbody>
</table>
Zooming in - Association analysis

Let’s annotate all verbs for stativity (Maienborn, 2015); and compare with the tense choice in German:

<table>
<thead>
<tr>
<th></th>
<th>Präteritum</th>
<th>Perfekt</th>
</tr>
</thead>
<tbody>
<tr>
<td>dynamic</td>
<td>6</td>
<td>345</td>
</tr>
<tr>
<td>stative</td>
<td>19</td>
<td>23</td>
</tr>
</tbody>
</table>

Typical state verbs in the Präteritum:
- être ‘to be’
- avoir ‘to have’
- paraître ‘to seem’
- falloir ‘to must’
- comprendre ‘to understand’
- vouloir ‘to want’
- croire ‘to believe’
- trouver ‘to find’
Let’s annotate all verbs for stativity (Maienborn, 2015); and compare with the tense choice in German:

<table>
<thead>
<tr>
<th></th>
<th>Präteritum</th>
<th>Perfekt</th>
</tr>
</thead>
<tbody>
<tr>
<td>dynamic</td>
<td>6</td>
<td>345</td>
</tr>
<tr>
<td>stative</td>
<td>19</td>
<td>23</td>
</tr>
</tbody>
</table>

Typical state verbs in the Präteritum:

- _être_ ‘to be’
- _avoir_ ‘to have’
- _paraître_ ‘to seem’
- _falloir_ ‘to must’
- _comprendre_ ‘to understand’
- _vouloir_ ‘to want’
- _croire_ ‘to believe’
- _trouver_ ‘to find’
Results - comparing all tenses

We repeated the annotation procedure for all indicative VPs in the first chapter of L’Étranger.

Conclusions:

▶ Clear dimensions: dimension 1: Past vs. Present, dimension 2: Perfect vs. unmarked
▶ Variation between languages mainly in the competition between Perfect and Past
▶ Only one new Perfect introduced: continuative use with stative verbs in English (as expected)

(4) en Have you been here long?
   fr Il y a longtemps que vous êtes là?
de Sind Sie schon lange hier?
nl Hoe lang bent u al hier?
es ¿Hace mucho tiempo que está usted aquí?
Conclusion

We improved Schaden (2009):

- Clear differences between languages
- No dichotomy, but rather a spectrum of \textit{Perfect} use
- Stative verbs lead to differences on both ends of the spectrum
Conclusion

We improved Schaden (2009):

▶ Clear differences between languages
▶ No dichotomy, but rather a spectrum of Perfect use
▶ Stative verbs lead to differences on both ends of the spectrum

Advantages of using MDS:

▶ hypothesis generation (and confirmation via additional annotation)
▶ white box method

Disadvantages:

▶ difficult to interpret dimensions
▶ overlapping contexts
On our to-do list

▶ Analyse languages that do not have a **Perfect** (e.g. Mandarin, Russian)
▶ More annotation layers, regression analysis of factors deciding between **Perfect** and **Past**
▶ Analyse competition between **Perfect** and **Present** (novel written in present tense)
▶ Repeat analysis for different genres (e.g. news articles, subtitles etc.)
On our to-do list

- Analyse languages that do not have a **Perfect** (e.g. Mandarin, Russian)
- More annotation layers, regression analysis of factors deciding between **Perfect** and **Past**
- Analyse competition between **Perfect** and **Present** (novel written in present tense)
- Repeat analysis for different genres (e.g. news articles, subtitles etc.)

- Your idea here?
On our to-do list

- Analyse languages that do not have a **Perfect** (e.g. Mandarin, Russian)
- More annotation layers, regression analysis of factors deciding between **Perfect** and **Past**
- Analyse competition between **Perfect** and **Present** (novel written in present tense)
- Repeat analysis for different genres (e.g. news articles, subtitles etc.)

- Your idea here?

Thanks! Stay tuned via time-in-translation.hum.uu.nl
Bonus slide: hierarchical analysis

Cluster Dendrogram

cdiss
hclust(*, "ward.D2")