CLARIAH meets Time in Translation - analysing sinds through WP3 tools

Martijn van der Klis

Time in Translation
UiL OTS, Utrecht University

CLARIAH Toogdag
Hilversum, 5 April 2019
Intro - Time in Translation

In the NWO-funded Time in Translation project, we aim to create a cross-linguistics semantics of the (present) Perfect. There is overlap, but also clear variation across Germanic (and Romance) languages in use of the Perfect.
In the NWO-funded Time in Translation project, we aim to create a cross-linguistics semantics of the (present) Perfect. There is overlap, but also clear variation across Germanic (and Romance) languages in use of the Perfect.

(1) a. John and Mary have used CLARIAH tools.
    b. Jan en Marie hebben CLARIAH tools gebruikt.
In the NWO-funded Time in Translation project, we aim to create a cross-linguistics semantics of the (present) Perfect. There is overlap, but also clear variation across Germanic (and Romance) languages in use of the Perfect.

(1) a. John and Mary have used CLARIAH tools.
b. Jan en Marie hebben CLARIAH tools gebruikt.

(2) a. John worked with OpenSoNaR yesterday.
b. Jan heeft gisteren met OpenSoNaR gewerkt.
In the NWO-funded Time in Translation project, we aim to create a cross-linguistics semantics of the (present) **Perfect**. There is overlap, but also clear variation across Germanic (and Romance) languages in use of the **Perfect**.

1. a. John and Mary *have used* CLARIAH tools.
   b. Jan en Marie *hebben* CLARIAH tools *gebruikt*.

2. a. John *worked* with OpenSoNaR yesterday.
   b. Jan *heeft* gisteren met OpenSoNaR *gewerkt*.

3. a. Mary *has known* GrETEL since version 1.0.
   b. Marie *kent* GrETEL sinds versie 1.0.
In this talk, we show that CLARIAH WP3-tools allow to accelerate *monolingual* linguistic research. Nevertheless, they currently fall short for *multilingual* analysis.
In this talk, we show that CLARIAH WP3-tools allow to accelerate **monolingual** linguistic research. Nevertheless, they currently fall short for **multilingual** analysis.

We use **SINCE**-adverbials (as in (3)) as a case study. In the literature on the **Perfect**, **SINCE**-adverbials play an important role, because in English, *since* appears exclusively with a *perfect* (present, past or future) in the main clause. There are however significant differences in use of **SINCE**-adverbials between the Germanic languages.
There are two main sources of variation within since-adverbials. Firstly, in English, *since* appears exclusively with a *perfect*. In German and Dutch, *seit* and *sinds* can appear in the *simple tenses* as well (recall (3)).
Intro - variation in SINCE-adverbials

There are two main sources of variation within SINCE-adverbials. Firstly, in English, *since* appears exclusively with a *perfect*. In German and Dutch, *seit* and *sinds* can appear in the *simple tenses* as well (recall (3)).

Secondly, German *seit* allows *durational phrases* as its complement. English uses *for* instead. Dutch can do without an adverb, but if an adverb is used, rather *nu* ‘now’ or *al* ‘already’ than *sinds*.

(4) a. Ich habe *seit* drei Stunden auf dich gewartet.
b. I have been waiting for you *for* three hours.
c. Ik heb (*nu/al/*sinds*) drie uur op je gewacht.
There are two main sources of variation within *since*-adverbials. Firstly, in English, *since* appears exclusively with a *perfect*. In German and Dutch, *seit* and *sinds* can appear in the *simple tenses* as well (recall (3)).

Secondly, German *seit* allows *durational phrases* as its complement. English uses *for* instead. Dutch can do without an adverb, but if an adverb is used, rather *nu* ‘now’ or *al* ‘already’ than *sinds*.

(4)  

a. Ich habe *seit* drei Stunden auf dich gewartet.  
b. I have been waiting for you *for* three hours.  
c. Ik heb *(nu/al/*sinds)* drie uur op je gewacht.

In this talk, we add more data for Dutch *sinds*. 
Dutch *sinds* - tense

Let’s use GrETEL to search for occurrences of Dutch *sinds*, and specify we want a single verb in the main clause. Result:

**Example-based Search**

Based on: "Dit is sinds 1998 het geval."

**209 Results for “lasy”**

<table>
<thead>
<tr>
<th>#</th>
<th>ID</th>
<th>Component</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>WS-U-E-A-000000036.p.20.s.5</td>
<td>WS</td>
<td>Sinds 2002 is hij weer Kamerlid.</td>
</tr>
<tr>
<td>5</td>
<td>WS-U-E-A-0000000005.p.37.s.7</td>
<td>WS</td>
<td>In Amsterdam zijn sinds kort 3 eurocafé's.</td>
</tr>
<tr>
<td>6</td>
<td>WS-U-E-A-0000000205.p.11.s.1</td>
<td>WS</td>
<td>Nederland exporteert in het verleden veel kennis over drooglegging, maar sinds kort is ons land koploper als het gaat om kennis over het gecontroleerd teruggeven van land aan het water.</td>
</tr>
<tr>
<td>8</td>
<td>WS-U-E-A-0000000237.p.44.s.1</td>
<td>WS</td>
<td>Aan het eind van deze uitzending kijken we nog even hoe het gaat bij de Nederlandse hockeyerinnen in Athene, die spelen sinds half acht vanavond in de finale tegen Australië.</td>
</tr>
</tbody>
</table>
Dutch *sinds* - Present vs. Past

GrETEL now allows to directly analyze these results, for example with respect to tense of the main verb:

**Analysis Table**

<table>
<thead>
<tr>
<th>$\text{Table}$</th>
<th>$\text{$\text{node1.pt}$}$</th>
<th>$\text{$\text{node1.lemma}$}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Count</strong></td>
<td>$\Downarrow$</td>
<td>$\Uparrow$</td>
</tr>
<tr>
<td>$\text{$\text{node1.pvtijd}$}$</td>
<td>$\text{Totals}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\text{tgw}$</td>
<td>181</td>
</tr>
<tr>
<td></td>
<td>$\text{verl}$</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td><strong>Totals</strong></td>
<td><strong>209</strong></td>
</tr>
</tbody>
</table>

Diagram: Node structures and relationships.
Dutch *sinds* - stative vs. dynamic

We can then further focus on the lemmata and see stative verbs (esp. *zijn* ‘to be’ and *hebben* ‘to have’) feature as main verb.

### Analysis Table

Click on a cell in the table to view those documents.

<table>
<thead>
<tr>
<th>$node1.pvtfijd</th>
<th>$node1.lemma</th>
<th>hebben</th>
<th>zijn</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$node1.pvtfijd</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tgw</td>
<td></td>
<td>13</td>
<td>62</td>
<td>75</td>
</tr>
<tr>
<td>verl</td>
<td></td>
<td>1</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>14</td>
<td>74</td>
<td>88</td>
</tr>
</tbody>
</table>
Dutch *sinds* - tense

Conclusions:

- Dutch allows *sinds* with simple tenses in the main clause as well
- **Present** seems more frequent than **Past**
- Stative verbs are most frequent (more on that later)
Dutch *sinds* - tense

Conclusions:

- Dutch allows *sinds* with simple tenses in the main clause as well
- Present seems more frequent than Past
- Stative verbs are most frequent (more on that later)

Let’s now turn to durations as complement.
Dutch *sinds* - durations

A simple search for *sinds* in OpenSoNaR shows that durations can actually function as complement:
Dutch *sinds* - durations allowed!

Closer analysis (*note: outside of the CLARIAH infrastructure*) reveals that with states, like in (5), Dutch *sinds* allows a duration. With activities, like in (6), this seems only allowed if we arrive at a **habitual** reading (i.e. non-episodic, compare (4)).

(5) **Sinds** een dag of twee / vlinders in mijn hoofd.¹ since a day or two / butterflies in my head

‘For a day or two now, I have butterflies in my head.’

(6) Ik volg **sinds** drie maanden Nederlandse les.² I follow since three months Dutch lesson

‘I have been following Dutch lessons for three months.’

---

¹Opening lyrics to Doe Maar’s song *32 jaar.*

²From Corpus Hedendaags Nederlands.
Dutch *sinds* - no durations in the *Perfect*?

From the examples above, in Dutch, *SINCE*-duration-adverbials seem limited to the *Present*, and require a stative or habitual reading. Recall however that German allows *seit* in the *Perfect* as well. Dutch in general does not like that, see (7) and (8).

(7) * Ik heb *sinds* een dag of twee vlinders in mijn hoofd gehad.

(8) * Ik heb *sinds* drie maanden Nederlandse les gevolgd.
Dutch *sinds* - durations in the **Perfect**!

However, if there is strong focus on the **consequent state**, as in (9), Dutch allows **sinds + duration** with a **Perfect**:

(9) Zij hebben **sinds** twee weken alle werkzaamheden gestaakt.
they have since two weeks all activities ceased

‘They ceased all activities two weeks ago.’
Dutch *sinds* - durations

Conclusions:

- Dutch allows *sinds* with durations
- ...but only with stative or habitual predicates in the Present
- ...and only with focus on the consequent state in the Perfect
Mapping variation with parallel corpora

Problem: monolingual research does not allow us to directly compare Dutch *sinds* with German *seit* and English *since*. 
Mapping variation with parallel corpora

Problem: monolingual research does not allow us to directly compare Dutch *sinds* with German *seit* and English *since*.

Solution: parallel corpora (e.g. translations of novels). Parallel corpora provide us with *form variation* while *meaning stays stable*. 
Mapping variation with parallel corpora

Problem: monolingual research does not allow us to directly compare Dutch *sinds* with German *seit* and English *since*.

Solution: parallel corpora (e.g. translations of novels). Parallel corpora provide us with *form variation* while *meaning stays stable*.

In the remainder of this talk, we shift our focus to the complete tense-aspect system, rather than just *SINCE*. 
Mapping variation with parallel corpora

What we would want, is to be able to map the variation between languages. For example, let’s suppose the French tense-aspect system looks a little bit like this...
French tense/aspect system

Perfect

Imperfective

PluPerfect

Present
English tense/aspect system

Perfect

Simple Past
Spanish tense/aspect system

Perfect

Simple Past
Dutch tense/aspect system
German tense/aspect system
Visualization with MDS

What sounds like a nice dream, has actually become reality!
Visualization with MDS

What sounds like a nice dream, has actually become reality!

We annotated verb phrases used in translations of Camus’ *L’étranger*, chapter 1 (OCR’d and cleaned up using PICCL).

We use **multidimensional scaling** (MDS) to visualize differences in translation. Our MDS visualizations allow to see the bigger picture, but also to drill down to the raw data.
French tense/aspect system (on repeat)
Visualization with MDS - French
## Drilling down

### Source
**French**

<table>
<thead>
<tr>
<th>French</th>
<th>passé composé</th>
<th>1.xml</th>
</tr>
</thead>
<tbody>
<tr>
<td>Je crois que j’ai somnolé un peu.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Translations

<table>
<thead>
<tr>
<th><strong>German</strong></th>
<th><strong>English</strong></th>
<th><strong>Simple past</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ich glaube, ich habe ein bißchen gedöse</td>
<td>I think I dozed off for a while.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Spanish</strong></th>
<th><strong>Italian</strong></th>
<th><strong>Infinito</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Creo que dormí un poco.</td>
<td>Credo di aver sonnecchiato un po’.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Dutch</strong></th>
<th><strong>Polish</strong></th>
<th><strong>No annotation available.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ik geloof dat ik een beetje heb gedozaad.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
English tense/aspect system (on repeat)
Visualization with MDS - English
Taking stock

CLARIAH tools allowed us to accelerate our research:

- Searching in *OpenSoNaR / Corpus Hedendaags Nederlands*
- Querying through dependency parses and pivoting in *GrETEL*
- Cleaning up OCRed text in *PICCL*
Taking stock

CLARIAH tools allowed us to accelerate our research:

▶ Searching in OpenSoNaR / Corpus Hedendaags Nederlands
▶ Querying through dependency parses and pivoting in GrETEL
▶ Cleaning up OCRed text in PICCL

CLARIAH could (should?) improve:

▶ Tooling for parallel corpora
▶ Annotation inside tools (rather than in Excel...)
Taking stock

CLARIAH tools allowed us to accelerate our research:

▶ Searching in OpenSoNaR / Corpus Hedendaags Nederlands
▶ Querying through dependency parses and pivoting in GrETEL
▶ Cleaning up OCRed text in PICCL

CLARIAH could (should?) improve:

▶ Tooling for parallel corpora
▶ Annotation inside tools (rather than in Excel...)

Hop on our *Time in Translation* wagon?
Thanks for your attention!

Visit us at https://time-in-translation.hum.uu.nl/