# 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

Universiteit Utrecht

# 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.

Universiteit Utrecht

# 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.

(1) a. John and Mary **have used** CLARIAH tools.

b. Jan en Marie **hebben** CLARIAH tools **gebruikt**.

Universiteit Utrecht

# 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.

(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**.

Universiteit Utrecht

# 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.

(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.

Universiteit Utrecht

# Intro - why SINCE-adverbials?

In this talk, we show that CLARIAH WP3-tools allow to accelerate **monolingual** linguistic research. Nevertheless, they currently fall short for **multilingual** analysis.

Universiteit Utrecht

# Intro - why SINCE-adverbials?

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.

Universiteit Utrecht

# 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)).

Universiteit Utrecht

# 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.

Universiteit Utrecht

# 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.

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

Universiteit Utrecht

# 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

Example / Parse / Matrix / Treebanks / Results / Analysis

XPath

```
//node[@cat="smain" and
    node[@rel="hd" and @pt="ww"] and
    node[@cat="pp" and @rel="mod" and
        node[@lemma="sinds" and @pt="vz" and @rel="hd"]]]
```

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

Components

| Name | Hits | All Sentences |
|---|---|---|
| DPC | 48 | 11,716 |
| WIKI | 35 | 7,341 |
| WRPE | 28 | 14,420 |
| WRPP | 49 | 17,691 |
| WSU | 49 | 14,032 |
| Total | 209 | 65,200 |

## 209 Results for *"lassy"*

Previous Next

| # | ID | Component | Sentence |
|---|---|---|---|
| 1 | WS-U-E-A-0000000241.p.30.s.2 | WS | **Hij is sinds 1992 directeur van het Kunstmuseum in het Duitse Wolfsburg** . |
| 2 | WS-U-E-A-0000000036.p.20.s.5 | WS | **Sinds 2002 is hij weer Kamerlid** . |
| 3 | WS-U-E-A-0000000013.p.41.s.3 | WS | **Israel houdt de Gazastrook bezet sinds de zesdaagse oorlog in 1967** . |
| 4 | WS-U-E-A-0000000032.p.48.s.6 | WS | Kwoot over buit **Antwerpen telt dus een en dertig overvallen sinds begin dit jaar** . |
| 5 | WS-U-E-A-0000000005.p.37.s.7 | WS | **In Amsterdam zijn sinds kort 3 eurocafé's** . |
| 6 | WS-U-E-A-0000000205.p.11.s.1 | WS | Nederland exporteerde 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** . |
| 7 | WS-U-E-A-0000000228.p.32.s.5 | WS | **Sinds 1996 vechten de rebellen tegen de monarchie** . |
| 8 | WS-U-E-A-0000000237.p.44.s.1 | WS | Aan het eind van deze uitzending kijken we nog even hoe het gaat bij de Nederlandse hockeyheren in Athene , **die spelen sinds half acht vanavond in de finale tegen Australie** . |

Universiteit Utrecht

# 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

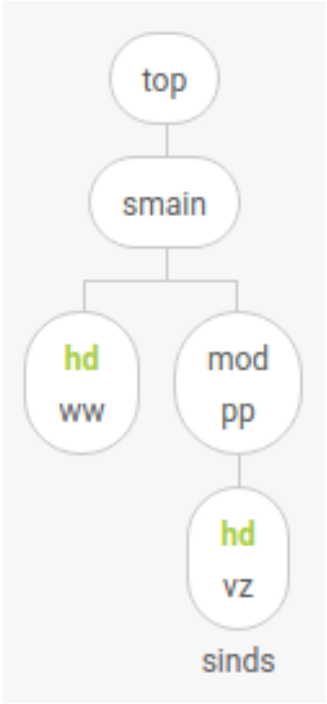
Click on a cell in the table to view those documents.

Table ▾ | $node1.pt ▾ | $node1.lemma ▾

Count ▾ ↕ ↔

$node1.pvtijd ▾

| $node1.pvtijd | Totals |
|---|---|
| tgw | 181 |
| verl | 28 |
| Totals | 209 |



# 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 ▾ | $node1.pt ▾

Count ▾ ↕ ↔ | $node1.lemma ▾

$node1.pvtijd ▾

| $node1.pvtijd | $node1.lemma | hebben | zijn | Totals |
|---|---|---|---|---|
| tgw | | 13 | 62 | 75 |
| verl | | 1 | 12 | 13 |
| | Totals | 14 | 74 | 88 |

Universiteit Utrecht

# 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)

Universiteit Utrecht

# 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.

Universiteit Utrecht

# Dutch *sinds* - durations

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

| # | Query | within | Metadata filters | Grouping | Status | Hits | Documents | |
|---|---|---|---|---|---|---|---|---|
| 1 | [word="sinds"] | document | | wordright:pos | FINISHED | 175666 | 129643 | EDIT X |

Query results:

Hits | Documents | Grouped hits | Grouped documents

p.o.s. right

Show 50 per page

<< < 1 2 3 > >>

Page 1 of 3 Go

Export

| Group ▲ ▼ | Hits ▲ ▼ |
|---|---|
| TW(hoofd,vrij) | 44531 |
| LID(bep,stan,rest) | 27146 |
| SPEC(deeleigen) | 18790 |
| LID(bep,stan,evon) | 9498 |
| ADJ(vrij,basis,zonder) | 8196 |
| BW() | 7798 |
| N(soort,ev,basis,onz,stan) | 7723 |
| TW(hoofd,prenom,stan) | 5654 |

| | | |
|---|---|---|
| ... ? De gastarbeiders die hier | **sinds dertig** | jaar voor de laagste loontjes ... |
| ... en 15 jaar , is | **sinds twee** | weken bij de apotheek verkrijgbaar ... |
| ... Vormingscentrum Wies Moens en is | **sinds twee** | jaar voorzitter van het ANZ ... |
| ... , onze gitarist , schrijft | **sinds twee** | maanden inderdaad opnieuw voor Humo ... |
| ... Athene speelt mee in Parijs | **Sinds tien** | jaar is het China wat ... |
| ... Rosas . Voor het eerst | **sinds twintig** | jaar danst choreografe Anne Teresa ... |
| ... miljoen pixels , in Nederland | **sinds twee** | weken te koop . Het ... |
| ... de financiering . Huisartsen kunnen | **sinds vier** | jaar bij een euthanasieverzoek een ... |
| ... heeft donderdagavond voor het eerst | **sinds 25** | jaar opnieuw een avondmarkt kunnen ... |
| ... . Er is zelfs , | **sinds 83** | jaar , een heuse onderscheiding ... |
| ... de cel . Vermeulen is | **sinds vijf** | jaar de uitbaatster van Vlaanderens ... |
| ... . Intussen is de vogelwaanzin | **sinds twee** | jaar compleet : mensen verlaten ... |
| ... zijn vrouw en zes kinderen | **sinds vier** | dagen de trotse bewoners van ... |
| ... Hans Pelsch , zelf al | **sinds vijf** | jaar voltijds schaakleerkracht in dertien ... |
| ... Internationale Groene Kruis waarvan hij | **sinds tien** | jaar voorzitter is . De ... |
| ... overgevoelig , glimlacht Margarita . | **Sinds zeven** | jaar weduwe bovendien . " ... |
| ... ook helemaal niet levensmoe : | **sinds twee** | weken was hij zelfs bijzonder ... |
| ... mensen . Den Haag richt | **sinds twee** | jaar de aandacht op de ... |
| ... , emeritus hoogleraar kunstgeschiedenis en | **sinds 1968** | lid van het Rembrandt Research ... |
| ... en ook België is daar | **sinds 1911** | lid van " , vertelt ... |

Universiteit Utrecht

# 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.'

[1] Opening lyrics to *Doe Maar*'s song *32 jaar*.
[2] From *Corpus Hedendaags Nederlands*.

Universiteit Utrecht

# 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.

Universiteit Utrecht

# 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.'


Universiteit Utrecht

# 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

Universiteit Utrecht

# Mapping variation with parallel corpora

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

Universiteit Utrecht

# 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*.

Universiteit Utrecht

# 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.

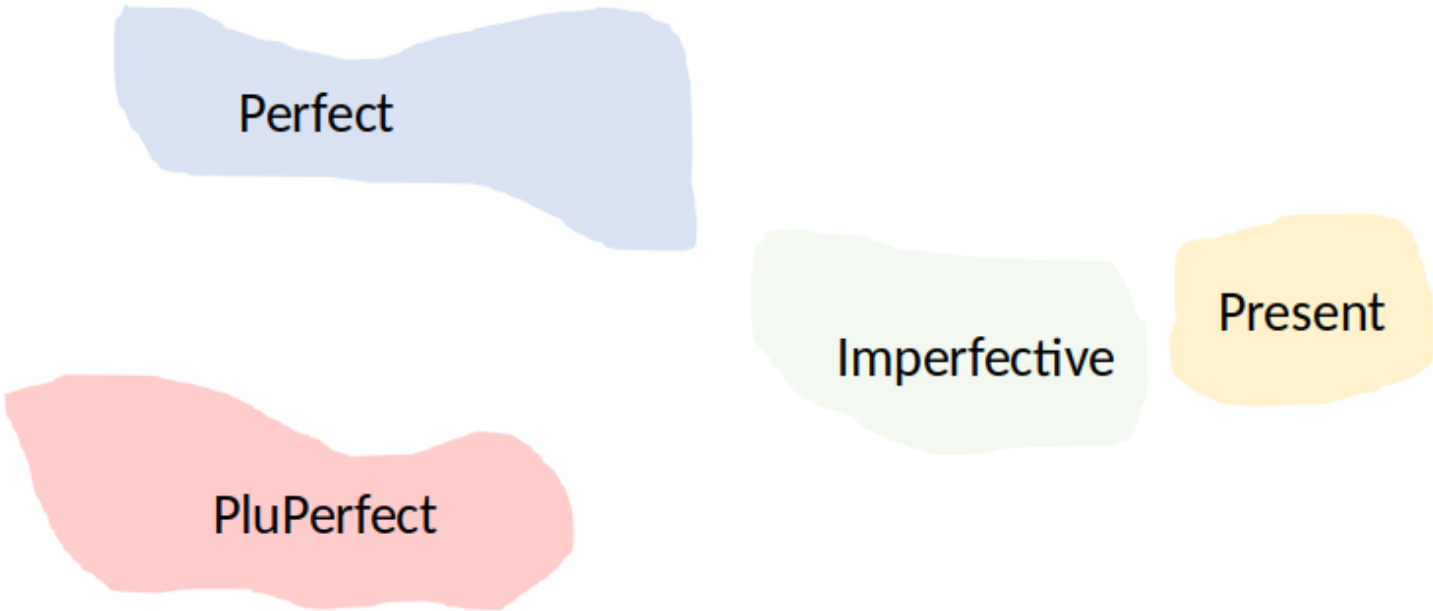
Universiteit Utrecht

# Mapping variation with parallel corpora

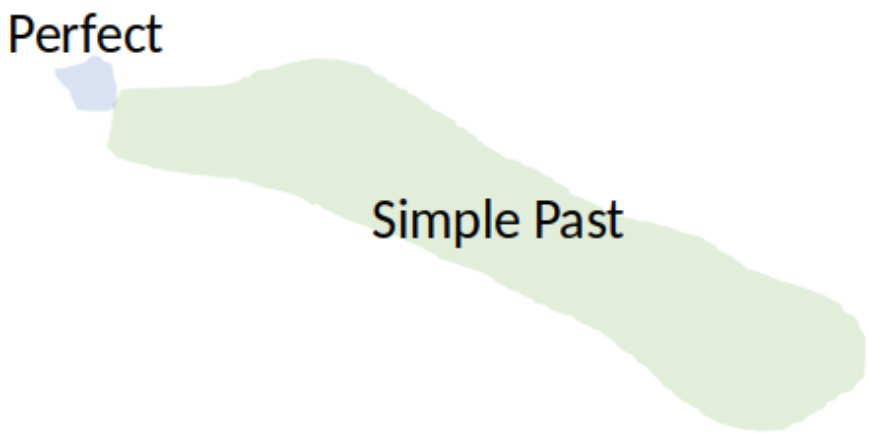
What we would want, is to be able to map the variation between languages. For example, let's suppose the French tense-aspect system looks a little bit like this...

Universiteit Utrecht

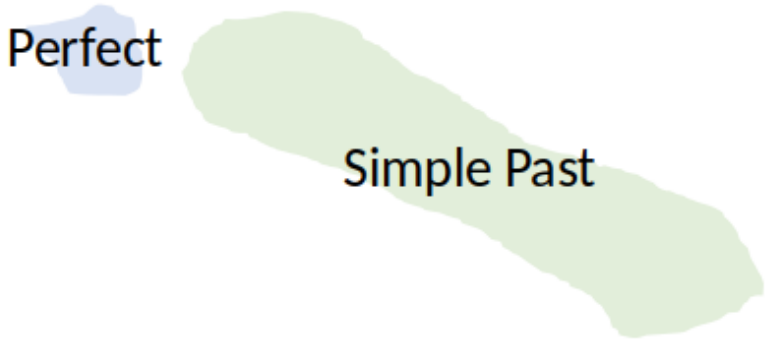
# French tense/aspect system



# English tense/aspect system

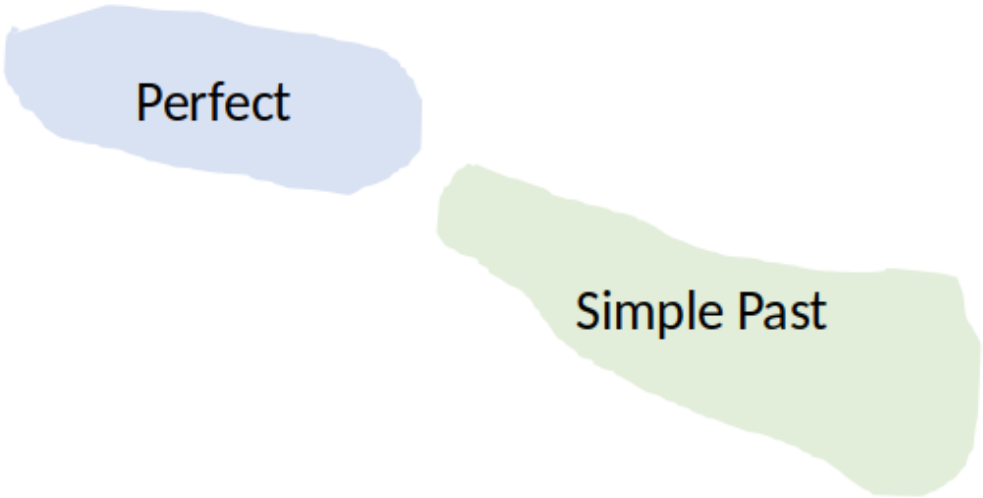


# Spanish tense/aspect system



# Dutch tense/aspect system

# German tense/aspect system




Universiteit Utrecht

# Visualization with MDS

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

Universiteit Utrecht

# Visualization with MDS

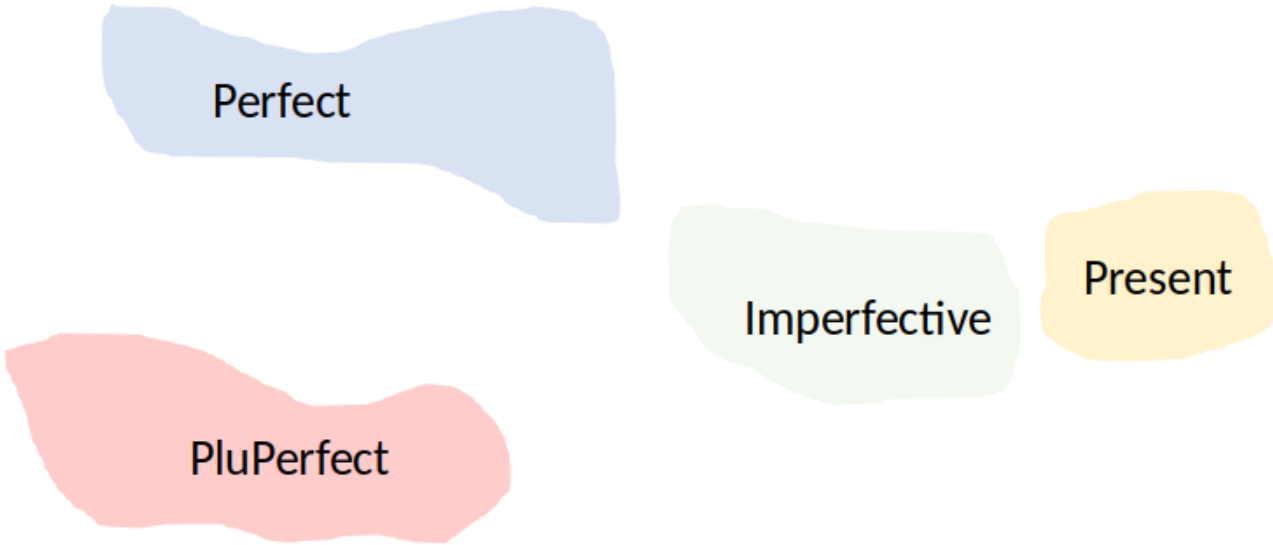
What sounds like a nice dream, has actually become reality!

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

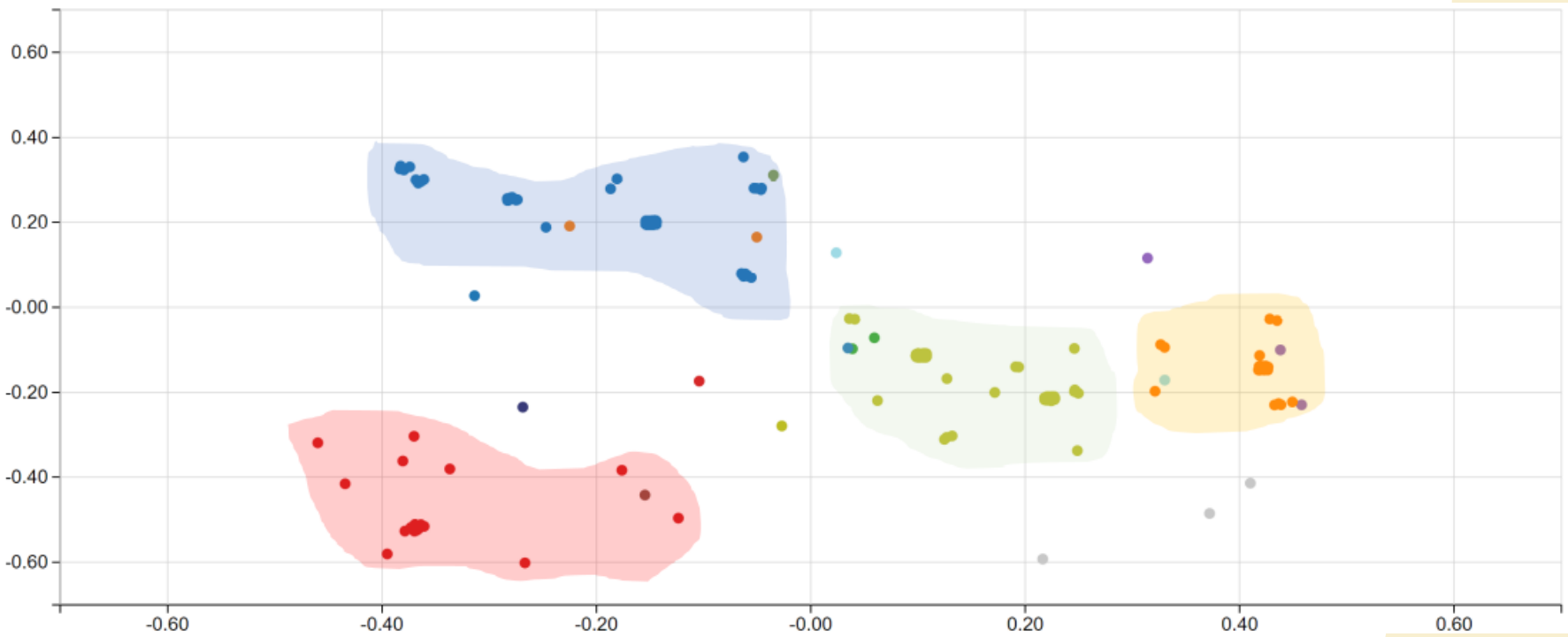
We use **multidimensional scaling** (MDS) to visualize differences in translation. Our MDS visualizations allow to see the bigger picture, but also to drill down to the raw data.

Universiteit Utrecht

# French tense/aspect system (on repeat)

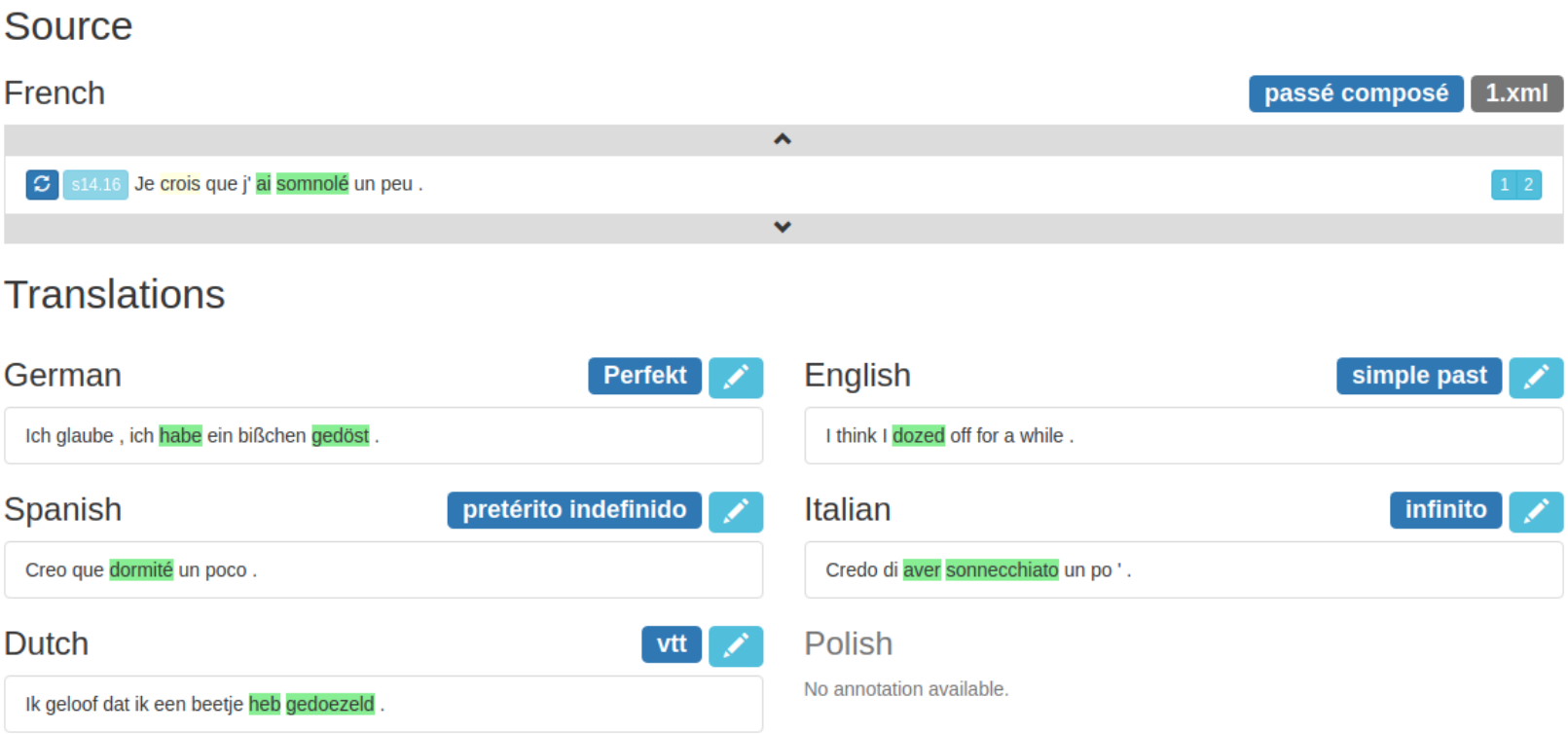


# Visualization with MDS - French



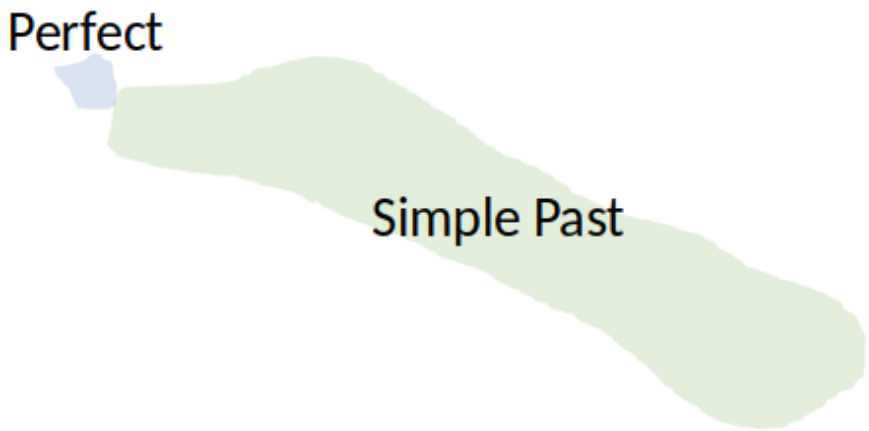

Universiteit Utrecht

# Drilling down

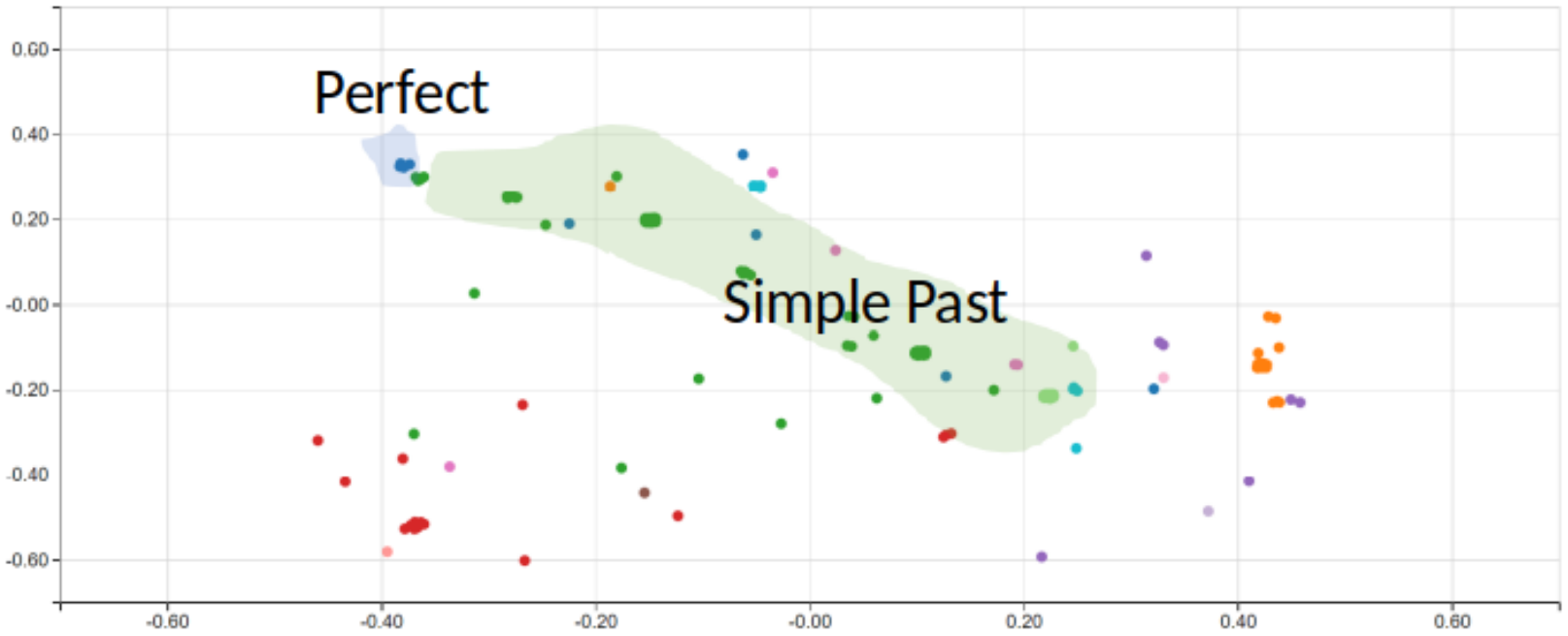



Universiteit Utrecht

# English tense/aspect system (on repeat)



# Visualization with MDS - English




Universiteit Utrecht

# 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**

Universiteit Utrecht

# 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...)

Universiteit Utrecht

# 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?

Universiteit Utrecht

# Thanks!

## Thanks for your attention!

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

Universiteit Utrecht