Using parallel and non-parallel corpora side by side

Maarten Bogaards | Leiden University Centre for Linguistics (LUCL)

Universiteit Leiden
The Netherlands

Heuristic extension of Translation Mining

(van der Klis et al. 2017; Lu & Verhagen 2016)

- Operationalizing crosslinguistic differences in grammaticalization/systematic expression by means of parallel corpora (translations)
- For example: viewpoint aspect in Mandarin and Russian vs. in Dutch and English

→ *Which questions can(not) be answered by means of parallel corpora? How can they be complemented by non-parallel corpora?*

Outline

1. The project
   - Aims
   - 4 components
2. Heuristic component (parallel)
3. Distributional component (non-parallel)
4. Integrative & Theoretical components
5. Trial run
   - Overview
   - Heuristic element: Mandarin *zhe* 着
   - Distributional analysis: Dutch [V_{POS} + ptc]
6. Outlook

The project: Aims

Aspect in Languages without Aspect

Funded by NWO (*PGW/PhDs in the Humanities*)

1 September 2020 – 31 August 2025

Supervisors: Sjef Barbiers & Ronny Boogaart

Main aims:

- Establish how a 'non-aspect language' expresses aspect and whether this differs from 'aspect languages'
  (in terms of aspect-prominence; cf. Bhat 1999)
- Specifically, along two lines:
  - Expressivity (Hypothesis of Equal Expressivity—e.g. Searle 1969)
  - Hierarchy (Universal Base Hypothesis—e.g. Cinque 1999)

The project: Aims

Main aims:

- Establish how a 'non-aspect language' expresses aspect and whether this differs from 'aspect languages'
  (in terms of aspect-prominence; cf. Bhat 1999)
- Specifically, along two lines:
  - Expressivity (Hypothesis of Equal Expressivity—e.g. Searle 1969)
  - Hierarchy (Universal Base Hypothesis—e.g. Cinque 1999)

Case study: Dutch vs. Mandarin/Russian
Viewpoint aspect: **systematicity** and **obligatoriness**

Empirical problem: how to identify aspectual expressions 'scattered throughout' the system?

→ *Heuristic Translation Mining* (1 of 4 components)

The project: 4 components

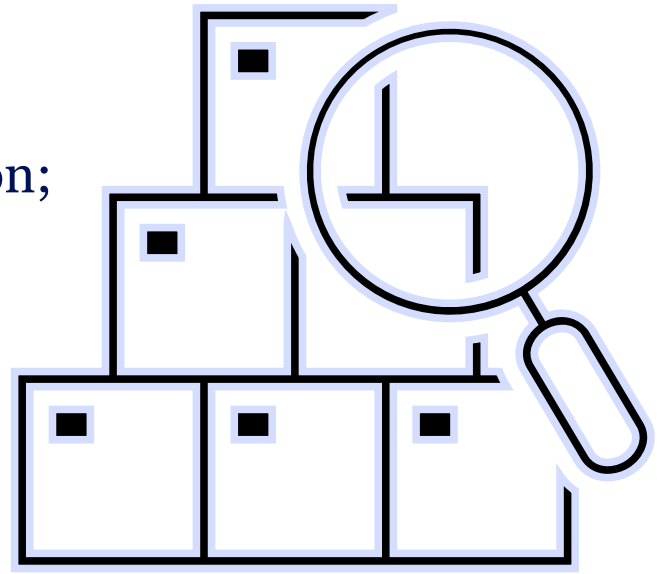
3. Integrative component
   - ? *How do these expressions carve up the aspectual domain in Dutch?*

4. Theoretical component
   - ? *Do the expressive potential and structural positions of aspectual expressions differ between Dutch and Mandarin/Russian?*

Heuristic component (parallel)

Heuristic Translation Mining (HTM): combining 2 methods

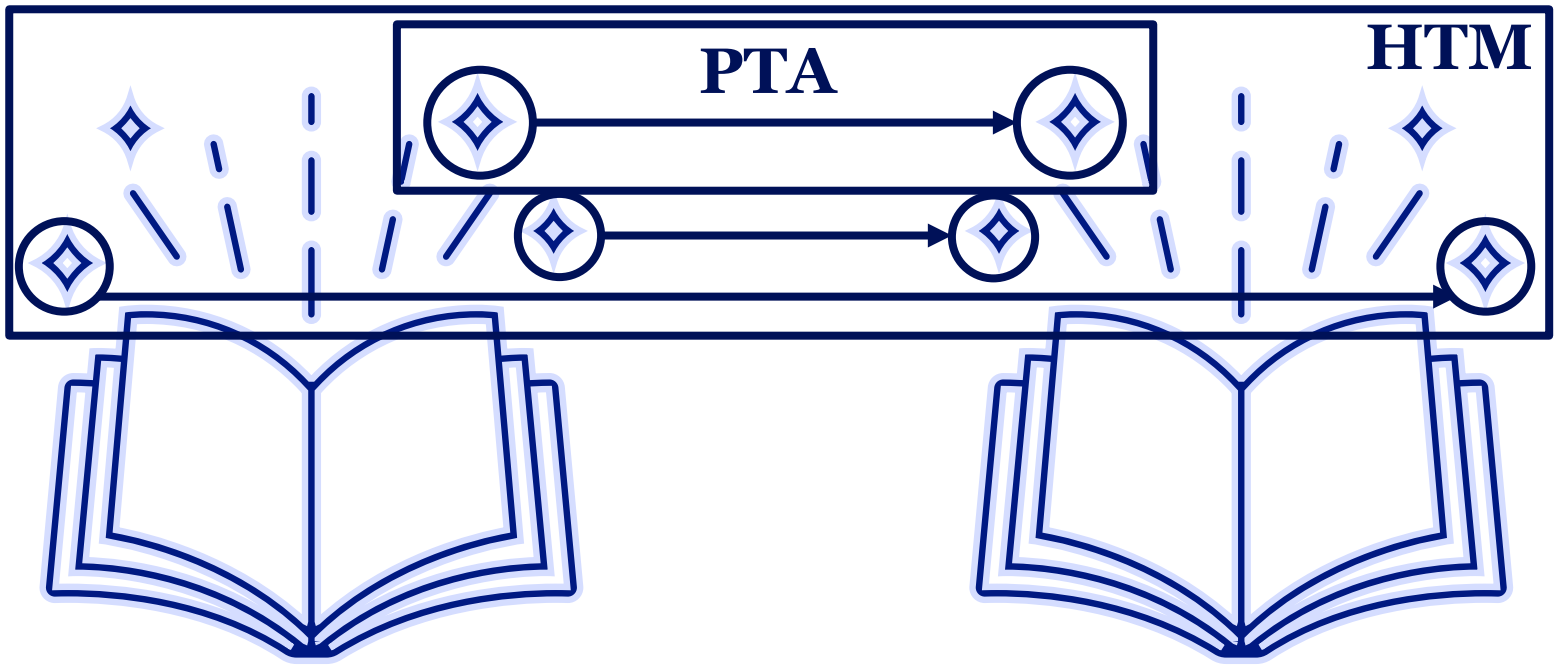
- **Parallel Text Analysis (PTA):** qualitative, fine-grained, 'one-way' comparison; multiple translations → translation strategies
  - Viewpoint (Lu & Verhagen 2016; Lu et al. 2018)
  - Tense (Lu 2019)
- **Translation Mining:** quantitative, large-scale, 'all-ways' comparison; comparing how different languages 'carve up' one conceptual space
  - Aspect (Dahl 2007, 2014; van der Klis et al. 2017, 2020)
  - Indefinite pronouns (Beekhuizen et al. 2017)
  - Conditionals (Tellings 2020)



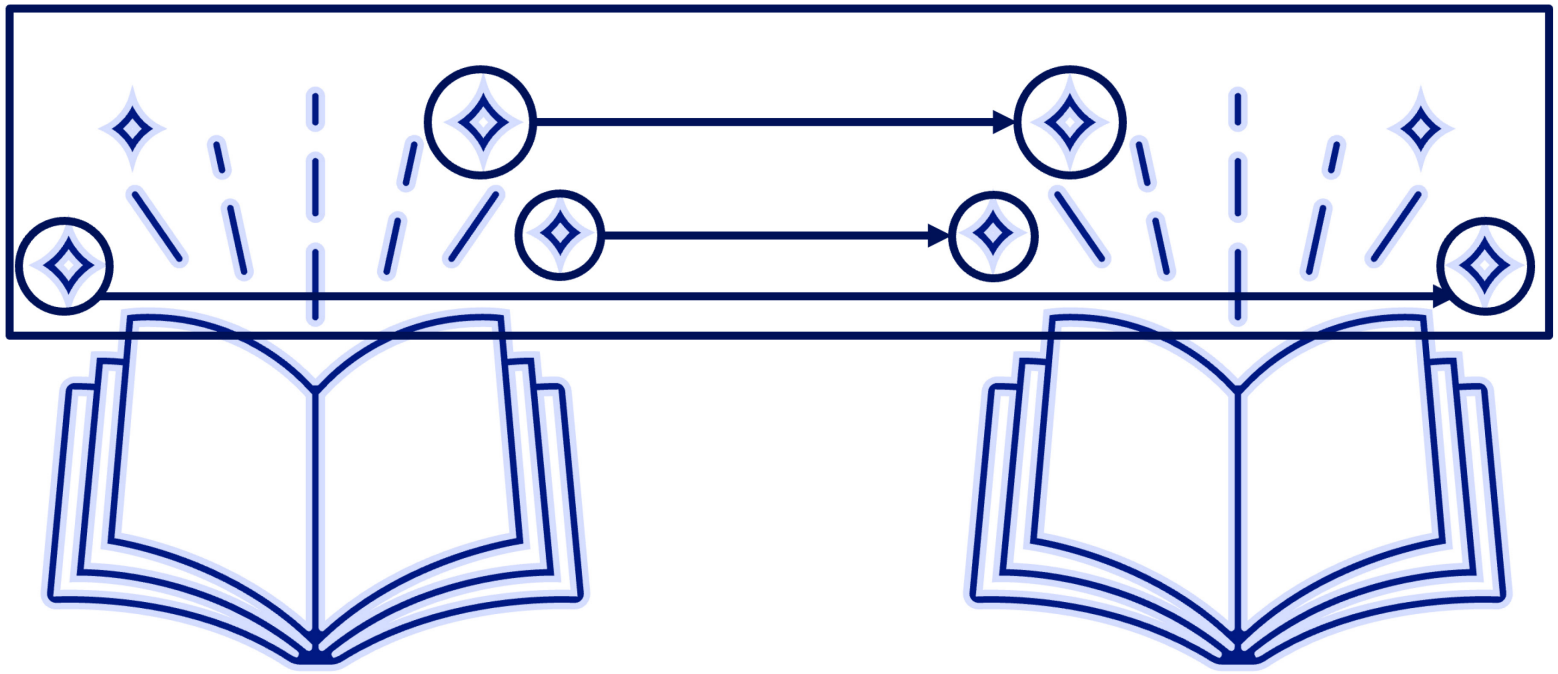
Core features of **HTM** (Bogaards 2019*a*):

1. Heuristic goal (prelude to comparison)
2. 'One-way street': instrument (source) → target language
3. Define conceptual content in terms of formal manifestation(s) in instrument language—avoiding the 'ontology problem'
4. Combination of qualitative (=PTA) and quantitative methods
5. Repeating HTM leads to TM

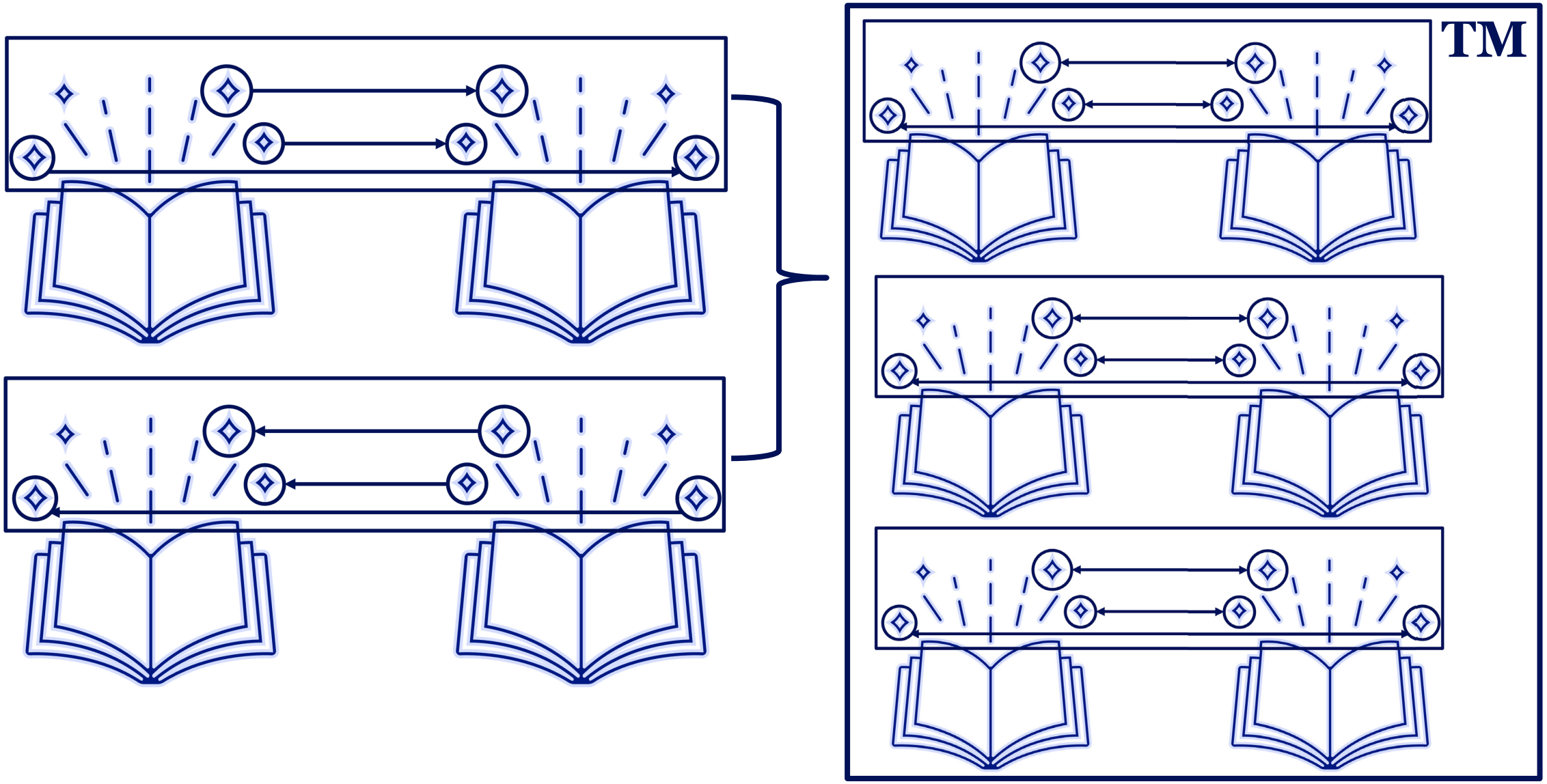
Heuristic component (parallel)



Heuristic component (parallel)



Heuristic component (parallel)



Heuristic component (parallel)

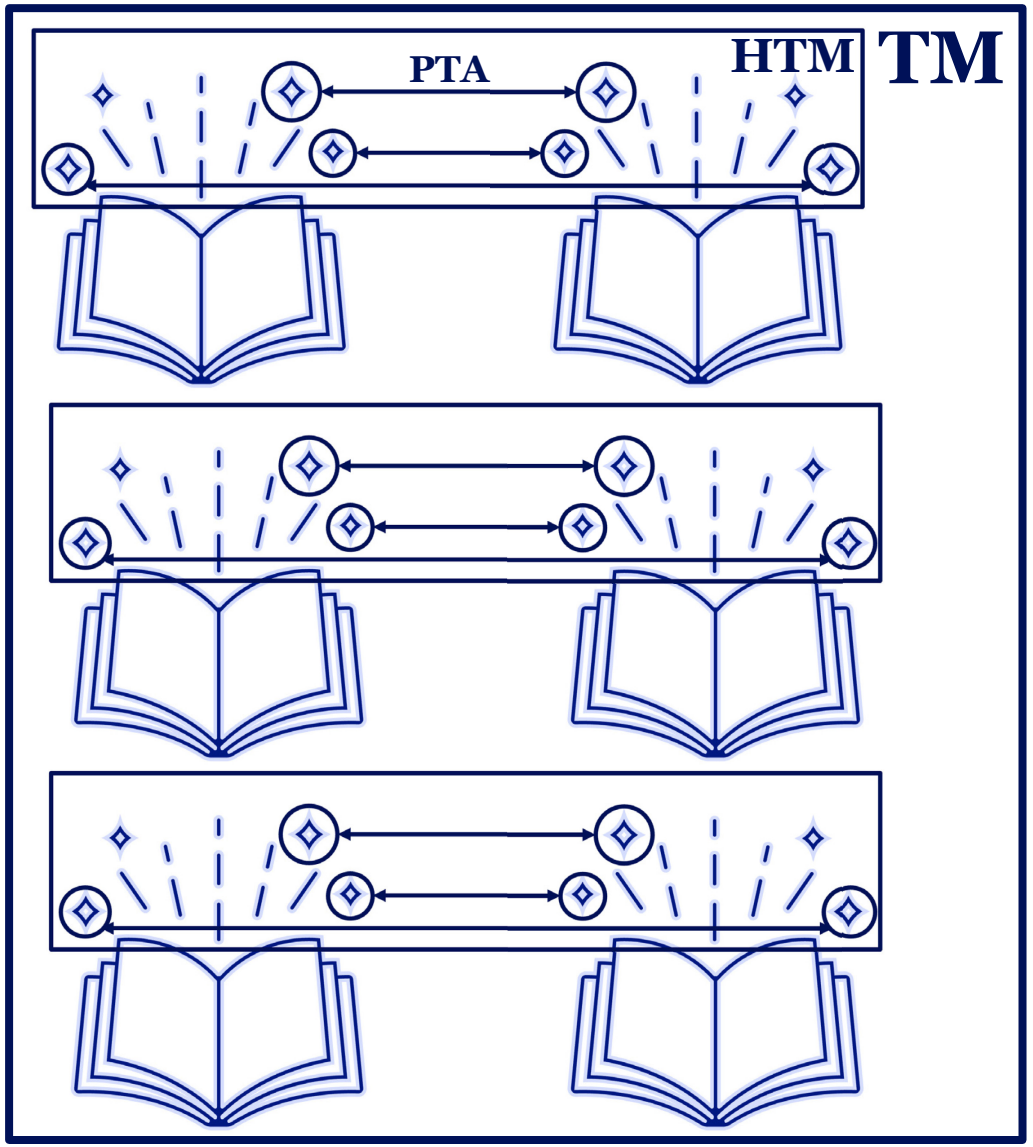
PTA → HTM → TM

Each have distinctly different goals

PTA establish with maximal specificity how translator(s) (re)produced conceptual content from instrument language in target language (focus on what is 'lost/gained in translation')

HTM link (part of) conceptual content from instrument language with (set of) structure(s) in target language, with quantitative basis (focus on 'where in linguistic inventory')

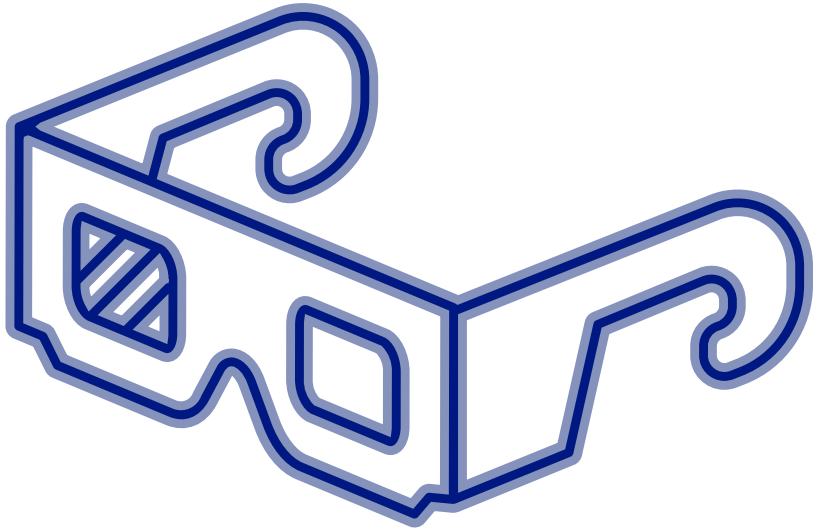
TM assume (partial) conceptual equivalence of (set of) structure(s) in examined languages, establish how they differ w.r.t. occupying part(s) of a conceptual space



Distributional component (non-parallel)

More traditional corpus analysis

But: ‘armed with’ knowledge gained from heuristic



Parallel corpora have their limits

1. Limited size
   → difficult to gain insight into restrictions
   → inadequate for general claims
2. Translated discourse ≠ ‘Untranslated’ discourse
   (cf. Vandevoorde et al. 2016 for Dutch)
3. Uneven quality of translations

So: non-parallel corpora/corpus are complementary

Heuristic findings as ‘hypotheses’;
develop and test hypotheses by means of non-parallel corpus

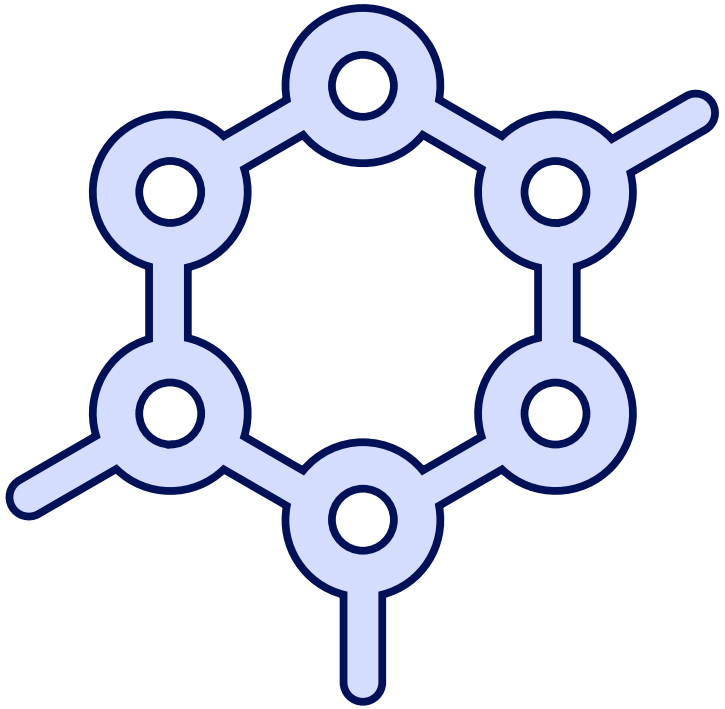
Integrative & Theoretical components

Aspectual expressions in Dutch: syntax and semantics characterized in distributional component based on **large** amount of **original** Dutch material

Integration: do (semi-)aspectual notions overlap?
(e.g. *resultativity* and *locativity*)
are all parts of the conceptual space 'taken care of'?

Theory: is the expressive potential of the 'integrated' set of aspectual expressions equal across Dutch, Mandarin, Russian?

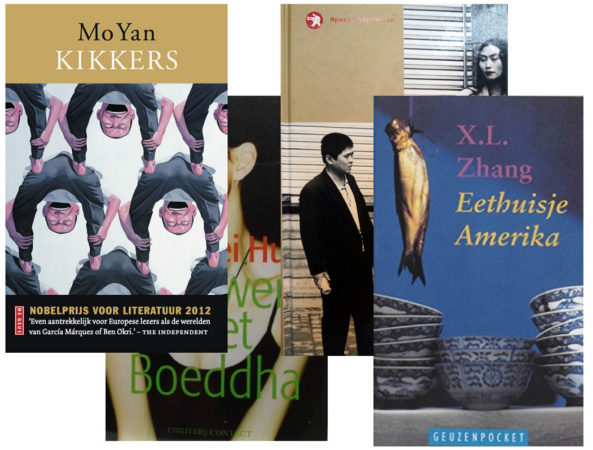
do equivalent (semi-)aspectual notions correspond to equivalent relative structural positions?



Trial run: Overview

HTM → Methodological concepts:

- **Heuristic element:** formal manifestation in terms of which conceptual content is defined
- **Theoretical benchmark:** basic template for target language in order to start interpreting results
- **Back-and-forth procedure:** way to get from individual observations (of the type of PTA) to overview (HTM)



Source materials: 5 Mandarin novels, 5 Dutch translations; first 100 instances of HE in each novel (→ 500 items)

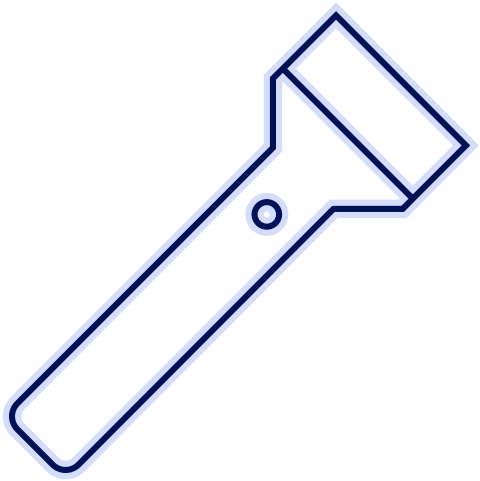
Heuristic element: Mandarin *zhe* 着 → durative/resultative

Trial run: Heuristic element (*zhe* 着)

V_{MAIN}-*zhe*: durative aspect (Chao 1968; Li & Thompson 1981; Dai 1997; Wiedenhof 2015, i.a.)

→ Focus on medial phase (defocusing boundaries)
→ Selects States, Activities, Accomplishments, Semelfactives (not Achievements)

(1) 老尤把守着大门。 (Xiao & McEnery 2004:192)
*Lǎo Yóu bǎshǒu-**zhe** dàmén.*
old You guard-**ZHE** gate
'Old You was guarding the gate.'



V_{MAIN}-*zhe*: may "present a state that follows the final point of a telic event" (Smith 1997:76) → resultative

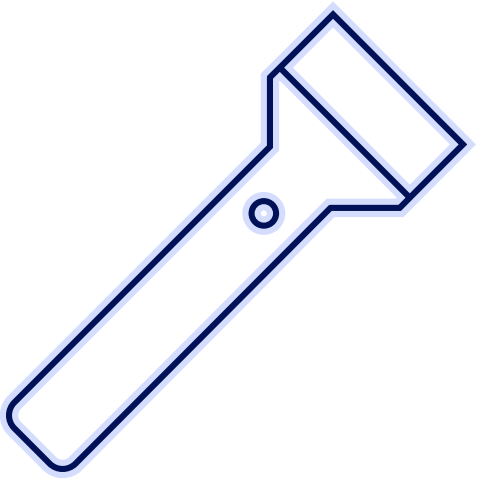
(2) 诗人穿着时兴的夹克。 (Dai 1997:90)
*Shīrén chuān-**zhe** shíxīng de jiākè*
poet put.on-**ZHE** fashionable SUB jacket
'The poet is wearing a fashionable jacket.' (Or: '…is putting on…')

Trial run: Heuristic element (*zhe* 着)

→ **Heuristic element:** Durativity (both stative and dynamic) and 'Resultative durativity' (only stative, following endpoint of telic event)

Main aspectual expressions found:

1. Unmarked **70.8%**
2. Posture verb (*zitten/staan/liggen* 'sit/stand/lie') **9%**
3. Posture verb + long infinitive **3%**



→ [V_{POS} + ptc]

(3) 堆**着**盛满空啤酒瓶和空可乐瓶的箱子。 [Bogaards 2018:B097]
*Duī-**zhe** chéng mǎn kōng píjiǔ píng hé kōng kělè píng de xiāngzi.*
stack-**ZHE** hold full empty beer bottle and empty cola bottle SUB box
(a) 'Er **stonden** dozen met lege bierflesjes en colablikjes **opgestapeld**.'
(b) 'There were stacks of boxes with empty beer bottles and coke cans.

Trial run: Distributional analysis ([V_{POS}+ptc])

→ **Hypothesis produced by heuristic:** [V_{POS}+ptc] expresses durativity and possibly resultativity

Non-parallel corpus: SoNaR (Oostdijk et al. 2013):

1. Extract sequences of V_{POS}+ptc and ptc+V_{POS}
   → [lemma="zitten"&pos="WW.pv.*"] [pos="WW.vd.vrij.*"]
   → [pos="WW.vd.vrij.*"] [lemma="zitten"&pos="WW.pv.*"]
2. Remove noise: 8,044 → 5,893 items
3. Annotate for various features (e.g. *main/sub; locative adjunct; resultative prefix*)

Features motivated by previous research *and* heuristic:

- *Main/sub*: "Verbal or adjectival complement?" (Haeseryn et al. 1997:963; Broekhuis & Corver 2015:993)
- *Locativity*: "Posture as location verbs" (Lemmens 2002)
- *Resultativity*: Heuristic element *zhe* 着

Trial run: Distributional analysis ([V_{POS}+ptc])

- *Main/sub*: "Verbal or adjectival complement?" (Haeseryn et al. 1997:963; Broekhuis & Corver 2015:993)
- *Locativity*: "Posture as location verbs" (Lemmens 2002)
- *Resultativity*: Heuristic element *zhe* 着

Locativity

Basically all instances implicitly or explicitly locative

(4) *[…]Hirsi Ali, die sinds de moord op Van Gogh* ***zit ondergedoken****.*
'Hirsi Ali, who since the murder of Van Gogh is [sits] in hiding.'

Resultativity

Semantic constraint on productivity: "Location = Result"

(5) *…dat het horloge in het kistje* ***zit*** <***opgeborgen***> <***verstopt***> <****gerepareerd***> <****vergeten***>.
'that the watch is [sits] <put away> <hidden> <*repaired> <*forgotten> in the case.'

→ Direct causal + spatial link between participle and posture verb

→ "Locative-resultative meaning" (Bogaards 2019*b*)

Trial run: Distributional analysis ([V_{POS}+ptc])

- *Main/sub*: "Verbal or adjectival complement?" (Haeseryn et al. 1997:963; Broekhuis & Corver 2015:993)

Main/sub and verb cluster order

Word order (V_{POS}-ptc OR ptc-V_{POS}) indicates status of complement: V_{POS}-ptc unacceptable with adjectival participle, cf. V-Adj:

(6) *…dat het broodje <lekker> is <*lekker>.*
'that the sandwich is tasty.'

Broekhuis & Corver (2015:963) claim: ptc in [V_{POS}+ptc] probably nominal
→ i.e. V_{POS}-ptc should not be attested

(7) *…dat mijn oma daar <**begraven**> **ligt** <??**begraven**>.*
'that my grandmother is [lies] buried there.'

Trial run: Distributional analysis ([V_{POS}+ptc])

Broekhuis & Corver (2015:963) claim: ptc in [V_{POS}+ptc] probably nominal
→ i.e. V_{POS}-ptc should not be attested

Corpus paints a more nuanced picture—major differences between V_{POS}- and ptc-type (cf. Bogaards 2019c)

percentage V_{POS}-ptc order (the closer to zero, the more 'nominal')

zit gevangen 'is imprisoned' (0/333 tokens)	**0%**
ligt verborgen 'is hidden'	**6.38%**
zit opgesloten 'is locked up'	**35.1%**
ligt opgebaard 'is laid out'	**69.6%**
staat opgetekend 'is written up'	**90.3%**
staat afgebeeld 'is portrayed'	**92.9%**

V_{POS}-ptc: only verbal
ptc-V_{POS}: verbal/adjectival

Verbal/adjectival status seems tied to resultativity:
more nominal = less resultative
→ Relevance of *resultativity* gained from parallel corpus
→ Relation to syntax only observable in non-parallel corpus

Outlook

Aspect in Languages without Aspect

Currently: Heuristic component—more 'heuristic elements' from more languages

Expectations:

- Light Verbs and Light Nouns (*slaan-slag*, *gaan-gang*)
- 'Situational PPs' (*aan de drugs*, *op gang*)
- Resultative particles (*lezen-uitlezen*, *schrijven-afschrijven*

...and hopefully: surprising forms we hadn't thought of in relation to aspect...!

Thank you for your attention!

Maarten Bogaards | Leiden University Centre for Linguistics (LUCL)
m.p.m.bogaards@hum.leidenuniv.nl

Universiteit Leiden
The Netherlands