University of Zurich UZH

English Department

prepositional English

patterns verbs

PEAS Project, funded by SNF
grant 100015_175987

PPs in the history of English:
methodological & theoretical challenges

Eva Zehentner

Workshop *Understanding grammar change: Digital resources and evolutionary modelling*
University of Edinburgh, 04 June 2024

University of Zurich UZH

English Department

How AI can solve subjectivity in (historical) data classification

Eva Zehentner

Workshop *Understanding grammar change: Digital resources and evolutionary modelling*
University of Edinburgh, 04 June 2024

University of Zurich UZH

English Department

prepositional
English

PEAS Project, funded by SNF
grant 100015_175987

PPs in the history of English: methodological & theoretical challenges

Eva Zehentner

Workshop *Understanding grammar change: Digital resources and evolutionary modelling*
University of Edinburgh, 04 June 2024

workshop aims

- prospects for creation of new, high-quality resources, including scope for addressing the current overrepresentation of English in historical corpora

- limitations of existing large corpora, and strategies for addressing them

- getting the most out of existing small, high-quality corpora

- lessons from sociophonetics and sociophonology, where larger, higher-quality resources are more common

- characterising key properties of language and its users in mathematical models

- accounting for the fluctuations arising from processes that are unobservable or not directly modelled

- using these as the basis for statistical inference

workshop aims

- prospects for creation of new, high-quality resources, including scope for addressing the current overrepresentation of English in historical corpora

- limitations of existing large corpora, and strategies for addressing them

- **getting the most out of existing small, high-quality corpora**

- lessons from sociophonetics and sociophonology, where larger, higher-quality resources are more common

- **characterising key properties of language and its users in mathematical models**

- accounting for the fluctuations arising from processes that are unobservable or not directly modelled

- using these as the basis for statistical inference

talk overview

- **overarching question:**
changes in strategies for argument disambiguation in history of English (synthetic/ morphological → analytic/ syntactic): prepositional patterns

- **issue A:**
challenges relating to historical data (especially Middle English)

- **issue B:**
challenges relating to theoretically and methodologically problematic phenomena (prepositions)

- **main point:**
tools and methods ↔ careful (historical) linguistic consideration

talk overview

- **overarching question:**
  changes in strategies for argument disambiguation in history of English (synthetic/ morphological → analytic/ syntactic): prepositional patterns

- **issue A:**
  challenges relating to historical data (especially Middle English)

- **issue B:**
  challenges relating to theoretically and methodologically problematic phenomena (prepositions)

- **main point:**
  tools and methods ↔ careful (historical) linguistic consideration

argument disambiguation strategies

- information/ question:
  - o **who** does **what** to **whom**? (agent vs patient/theme vs experiencer, etc.)

- strategies:
  - o semantic-pragmatic asymmetries (e.g. animacy biases)
  - o noun morphology (e.g. case)
  - o verb morphology (e.g. subject/verb agreement)
  - o constituent order (e.g. strict SVO)
  - o adpositions (e.g. prepositional marking)

argument disambiguation strategies

- information/ question:
  - **who** does **what** to **whom**? (agent vs patient/theme vs experiencer, etc.)

- strategies: **Old English**
  - semantic-pragmatic asymmetries (e.g. animacy biases)
  - noun morphology: extensive nominal inflection
  - verb morphology: extensive verbal inflection
  - constituent order: flexible
  - adpositions: some prepositional marking

argument disambiguation strategies

- information/ question:
  - **who** does **what** to **whom**? (agent vs patient/theme vs experiencer, etc.)

- strategies: **PDE**
  - semantic-pragmatic asymmetries (e.g. animacy biases)
  - noun morphology: little/ no nominal inflection
  - verb morphology: little verbal inflection
  - constituent order: fixed SVO
  - adpositions: extensive prepositional marking

Đone cyning	hi	*brohton*	cucene	**to** Iosue.
The king$_{ACC.SG}$	they$_{NOM.PL}$	brought$_{PL}$	alive	to Joshua

'They brought the king alive to Joshua'

(Ælfric, Libellus; Baker 2012:117)

argument disambiguation strategies

- information/ question:
  - **who** does **what** to **whom**? (agent vs patient/theme vs experiencer, etc.)

- strategies:
  - semantic-pragmatic asymmetries (e.g. animacy biases)
  - noun morphology (e.g. case)
  - verb morphology (e.g. subject/verb agreement)
  - constituent order (e.g. strict SVO)
  - adpositions (e.g. prepositional marking)

- *PEAS*-project (2018-2023): prepositions in English argument structure (PI: M. Hundt)

- *SArDEEn*-project (?2025-2029): strategies of argument disambiguation in early English

argument disambiguation strategies

- information/ question:
  - **who** does **what** to **whom**? (agent vs patient/theme vs experiencer, etc.)

- strategies:
  - semantic-pragmatic asymmetries (e.g. animacy biases)
  - noun morphology (e.g. case)
  - verb morphology (e.g. subject/verb agreement)
  - constituent order (e.g. strict SVO)
  - adpositions (e.g. prepositional marking)

- *PEAS*-**project (2018-2023):** prepositions in English argument structure (PI: M. Hundt)

- *SArDEEn*-**project (?2025-2029):** strategies of argument disambiguation in early English

argument disambiguation strategies

- information/ question:
  - **who** does **what** to **whom**? (agent vs patient/theme vs experiencer, etc.)

- strategies:
  - semantic-pragmatic asymmetries (e.g. animacy biases)
  - noun morphology (e.g. case)
  - verb morphology (e.g. subject/verb agreement)
  - constituent order (e.g. strict SVO)
  - adpositions (e.g. prepositional marking)

- *PEAS*-**project (2018-2023):** prepositions in English argument structure (PI: M. Hundt)

- *SArDEEn*-**project (?2025-2029):** strategies of argument disambiguation in early English

prepositional patterns

- development of verb-attached PPs from Middle to Late Modern English

- changes in types/ functions of PPs (adjunct vs complements)

- changes in relation between nominal and prepositional patterns

PP
- many *believed* **on his name** (LModE; NEWCOME-NEW-1796-1,2,20J.151)

- so *mounted* **uppon his hors** 'so mounted upon his horse' (ME; CMMALORY,181.2448)

- **in þis ȝere** […] þe sete *was voyde* 'in this year, the seat was voided' (ME; CMCAPCHR,141.3277)

NP
- you may *beleive* **mee** 'you may believe me' (EModE; HOXINDEN-1660-E3-H,276.91)

- and sir philip *mounted* **his horse** (LModE; BOSWELL-1776-1,47.393)

- **the same ȝere** cam þe kyng to seynt albones (ME; CMCAPCHR,139.3242)

(e.g. Mustanoja 1960; Strang 1970; De la Cruz 1973; Denison 1981, 1985, 2014; Fischer 1992; Traugott 1972, 1992; Lundskær-Nielsen 1993; Allen 1995, 2005; Claridge 2000; Baugh & Cable 2002; Fischer & van der Wurff 2006; De Cuypere 2015a, 2015b; Zehentner 2019)

data & methods

- **PPCME2**: 1150-1500, 1.2 million words
- **PPCEME**: 1500-1710, 1.8 million words
- **PPCMBE2**: 1700-1914, 2.8 million words

- V $ PP/ V $ NP-…
- N = 406,490

period	N (pmw)
ME	96,802 (83,741)
EME	121,573 (67,766)
LME	188,115 (67,424)

variant	N
NP	200,321
PP	206,169

talk overview

- **overarching question:**
changes in strategies for argument disambiguation in history of English (synthetic/ morphological → analytic/ syntactic): prepositional patterns

- **issue A:**
challenges relating to historical data (especially Middle English)

- **issue B:**
challenges relating to theoretically and methodologically problematic phenomena (prepositions)

- **main point:**
tools and methods ↔ careful (historical) linguistic consideration

talk overview

- **overarching question:**
  changes in strategies for argument disambiguation in history of English (synthetic/ morphological → analytic/ syntactic): prepositional patterns

- **issue A:**
  challenges relating to historical data (especially Middle English)

- **issue B:**
  challenges relating to theoretically and methodologically problematic phenomena (prepositions)

- **main point:**
  tools and methods ↔ careful (historical) linguistic consideration

talk overview

- **overarching question:**
  changes in strategies for argument disambiguation in history of English (trade-off synthetic/ morphological → analytic/ syntactic): new approaches, new tools

- **issue A:**
  challenges relating to historical data (especially Middle English)

- **issue B:**
  challenges relating to theoretically and methodologically problematic phenomena

- **main point:**
  tools and methods ↔ careful (historical) linguistic consideration

prepositional patterns

- development of verb-attached PPs from Middle to Late Modern English

- changes in types/ functions of PPs (adjunct vs complements)

- changes in relation between nominal and prepositional patterns

PP
- o   many *believed* **on his name** (LModE; NEWCOME-NEW-1796-1,2,20J.151)
- o   so *mounted* **uppon his hors** 'so mounted upon his horse' (ME; CMMALORY,181.2448)
- o   **in þis ȝere** [...] þe sete *was voyde* 'in this year, the seat was voided' (ME; CMCAPCHR,141.3277)

NP
- o   you may *beleive* **mee** 'you may believe me' (EModE; HOXINDEN-1660-E3-H,276.91)
- o   and sir philip *mounted* **his horse** (LModE; BOSWELL-1776-1,47.393)
- o   **the same ȝere** cam þe kyng to seynt albones (ME; CMCAPCHR,139.3242)

(e.g. Mustanoja 1960; Strang 1970; De la Cruz 1973; Denison 1981, 1985, 2014; Fischer 1992; Traugott 1972, 1992; Lundskær-Nielsen 1993; Allen 1995, 2005; Claridge 2000; Baugh & Cable 2002; Fischer & van der Wurff 2006; De Cuypere 2015a, 2015b; Zehentner 2019)

adjuncts vs complements

(on) that day   *they searched*   (for) us

'adjunct' (non-core, peripheral)   **'complement' (core)**

- optionality

  They searched [for us].

  They searched for us [on that day].

- prepositional passive

  They were searched for.

  That day was searched for us on.

- *do*-paraphrase-ability

  They searched for us, and Laurie did so for you.

  They searched for us on that day, and Laurie did so on the next day.

- *happen*-paraphrase-ability

  They searched. This happened for us.

  They searched for us. This happened on that day.

adjuncts vs complements

((on) that day) *they searched* ((for) us)

↓ ↓

'adjunct' (non-core, peripheral)

- optionality

  They searched [for us].

  They searched for us [on that day].

- prepositional passive

  They were searched for.

  That day was searched for us on.

'complement' (core)

- position/ ordering

  ?[For us] they searched.

  They searched for us [on that day].

  They searched [on that day] [for us].

- iterativity

  ?They searched for us [through the house]…

  They searched for us [on that day] [in the morning] [at home] [with us]…

adjuncts vs complements

"a terminological and conceptual mess" (Bergs 2021: 145)

"In practice it is hard to make an absolute distinction"
(Biber et al. 1999: 403; cf. also Quirk et al. 1985: 501-511, 1162-1167; Huddleston & Pullum 2000; Hoffmann 2007)
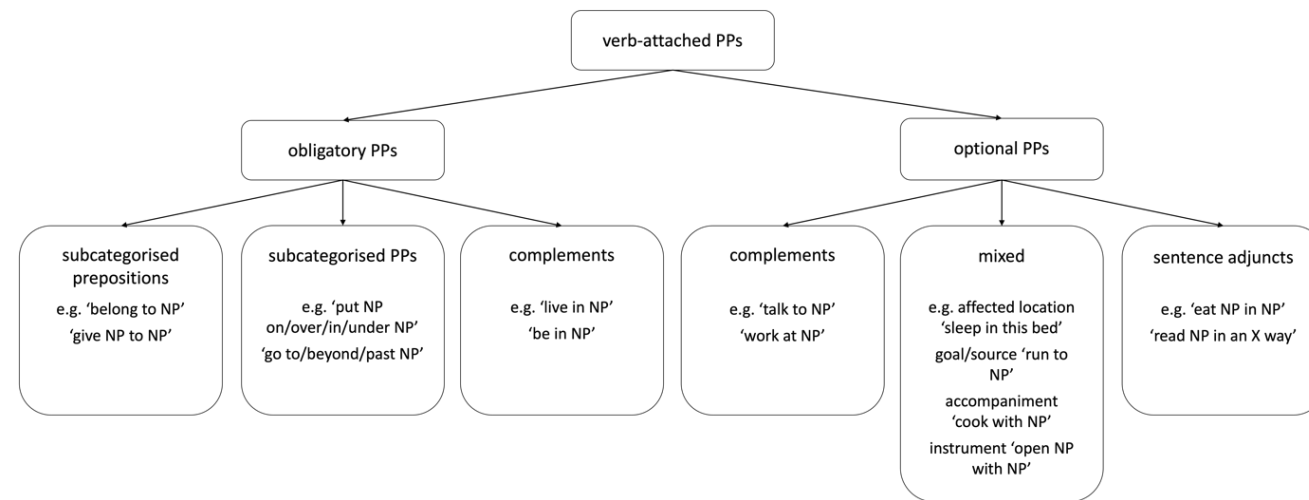
network of verb-attached PPs in PDE (adapted from Hoffmann 2007 and Bergs 2021)

adjuncts vs complements

"a terminological and conceptual mess" (Bergs 2021: 145)

"In practice it is hard to make an absolute distinction"
(Biber et al. 1999: 403; cf. also Quirk et al. 1985: 501-511, 1162-1167; Huddleston & Pullum 2000; Hoffmann 2007)

"native speakers' judgments on the argument and adjunct status of PPs are very unstable"

"the tests of argumenthood are often difficult to judge or even contradictory with each other"
(Merlo & Esteve Ferrer 2006: 31)

"few attempts have been made to perform this distinction automatically"

"The usual expectation has been that this discrimination is not amenable to a corpus-based treatment"
(Merlo & Esteve Ferrer 2006: 3)

data & methods

- **PPCME2**: 1150-1500, 1.2 million words
- **PPCEME**: 1500-1710, 1.8 million words
- **PPCMBE2**: 1700-1914, 2.8 million words

○ V $ PP/ V $ NP-…

○ N = 406,490

(IP-MAT (**PP** (P From) (NP (D the) (NUM 22.) (N day) (CP-REL (WNP-1 0) (C that) (IP-SUB (IP-SUB (NP-TMP *T*-1) (NP-SBJ (PRO wee)) (VBD espied) (**NP-OB1** (D the) (N Iland) (CODE) (NP-PRN (NPR Gomora)))) (CODE) (, ,) (CONJP (CONJ and) (IP-SUB (NP-TMP *T*-1) (NP-SBJ *con*) (VBD came) (**PP** (P amongst) (NP (D the) (NS Ilands)))))))) …
(NP-SBJ (PRO we)) (VBD came) (**PP** (P to) (NP (D an) (N ancor))) (**NP-TMP** (D that) (N night)

From the 22. day that wee *espied* **the Iland Gomora**, and *came* **amongst the Ilands** […] we *came* **to an ancor** **that night**
'From the 22nd day that we espied the island Gomora, and came amongst the islands […] we came to an anchor that night'
(EModE; COVERTE-E2-P1,7.85)

"This goal requires us to avoid subjective judgments since they are extremely error-prone. So, for example, we do not distinguish adjectival from verbal passive participles, nor do we attempt to implement the **argument-adjunct distinction**"

(Santorini 2016)

data & methods

- **PPCME2**: 1150-1500, 1.2 million words
- **PPCEME**: 1500-1710, 1.8 million words
- **PPCMBE2**: 1700-1914, 2.8 million words

○ V $ PP/ V $ NP-…

○ N = 406,490

(IP-MAT (**PP** (P From) (NP (D the) (NUM 22.) (N day) (CP-REL (WNP-1 0) (C that) (IP-SUB (IP-SUB (NP-TMP *T*-1) (NP-SBJ (PRO wee)) (VBD espied) (**NP-OB1** (D the) (N Iland) (CODE) (NP-PRN (NPR Gomora)))) (CODE) (, ,) (CONJP (CONJ and) (IP-SUB (NP-TMP *T*-1) (NP-SBJ *con*) (VBD came) (**PP** (P amongst) (NP (D the) (NS Ilands)))))))) …
(NP-SBJ (PRO we)) (VBD came) (**PP** (P to) (NP (D an) (N ancor))) (**NP-TMP** (D that) (N night)

From the 22. day that wee *espied* **the Iland Gomora**, and *came* **amongst the Ilands** […] we *came* **to an ancor that night**
'From the 22nd day that we espied the island Gomora, and came amongst the islands […] we came to an anchor that night'
(EModE; COVERTE-E2-P1,7.85)

semantic role annotation:

- NPs: e.g. temporal, locative, directional, measure, object(s), adjunct
- PPs: no distinction

approaches

- manual sub-setting based on specific alternations/ prepositions/ verb types/ semantic roles (Zehentner 2022a; 2023a,b; 2024; in prep. a)

- manual role classification (Zehentner & Hundt 2021a; Zehentner 2022b)

- *FrameNet*-based role/ type classification (Zehentner & Hundt 2021b, in prep.)

- manual feature-based type clustering (Zehentner 2023c)

- automatised feature-based (NLP-derived) type clustering (Zehentner 2021c, d, in prep. b)

- automatised (NLP-derived) sense/role/type disambiguation (SArDEEn-project)

approaches

- manual sub-setting based on specific alternations/ prepositions/ verb types/ semantic roles (Zehentner 2022a; 2023a,b; 2024; in prep. a)

- manual role classification (Zehentner & Hundt 2021a; Zehentner 2022b)

- *FrameNet*-based role/ type classification (Zehentner & Hundt 2021b, in prep.)

- manual feature-based type clustering (Zehentner 2023c)

- automatised feature-based (NLP-derived) type clustering (Zehentner 2021c, d, in prep. b)

- automatised (NLP-derived) sense/role/type disambiguation (SArDEEn-project)

sub-setting: time expressions

NP

- o **That day** I saw them for three hours (COCA, 2012; tinyarticle.com)
- o She had made no prior purchases **that week**. (COCA, 2012; ebcitizen.com)
- o **That year**, Missouri and West Virginia were the upset victims (COCA, 2012; ocregister.com)

PP

- o I told the police where I was **at that day** (COCA, 2011; Paradise Lost 3: Purgatory)
- o What are they going to be talking about **in that week?** (COCA, 2019; NPR_Morning)
- o **In that year**, I did not go with them (COCA, 2012; Demand EUPHORIA)

- **14 time- head nouns** (attested in all periods and both NP- and PP-patterns): *afternoon* [200], *day* [2,462], *hour* [176], *month* [153], *morning* [736], *morrow* [103], *night* [812], *noon* [69], *season* [110], *summer* [75], *time* [3,040], *week* [154], *winter* [75], *year* [1,147]

- **N=9,312** (NPs: 4,779 vs PPs: 4,885)

sub-setting: time expressions
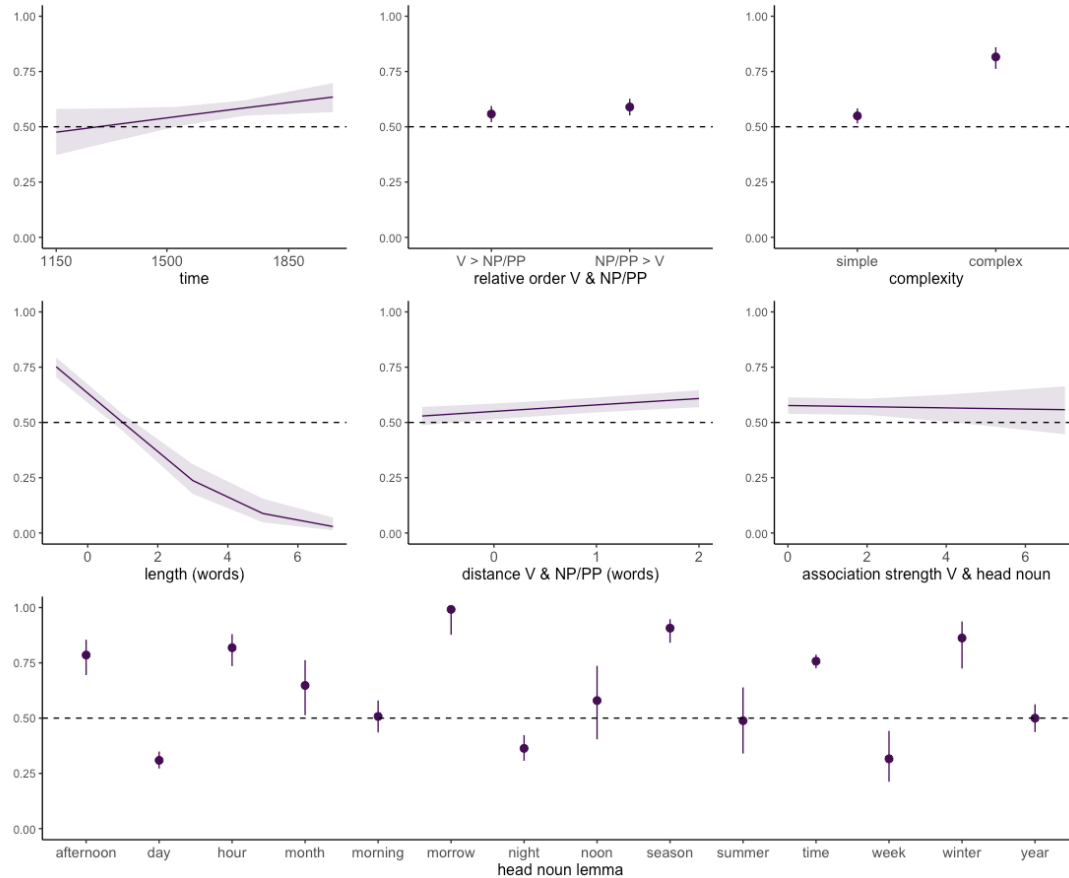
- mixed-effects logistic regression
- random effects: verb lemma, text ID
- significant **positive** impact of later time, greater morphosyntactic complexity, greater distance, and head nouns such as *hour* on PP-use
- significant **negative** impact of greater length and nouns such as *week* on PP-use
- non-significant impact of relative order and association strength
- significant interaction between time and relative order (greater likelihood of pre-verbal PPs in later texts)
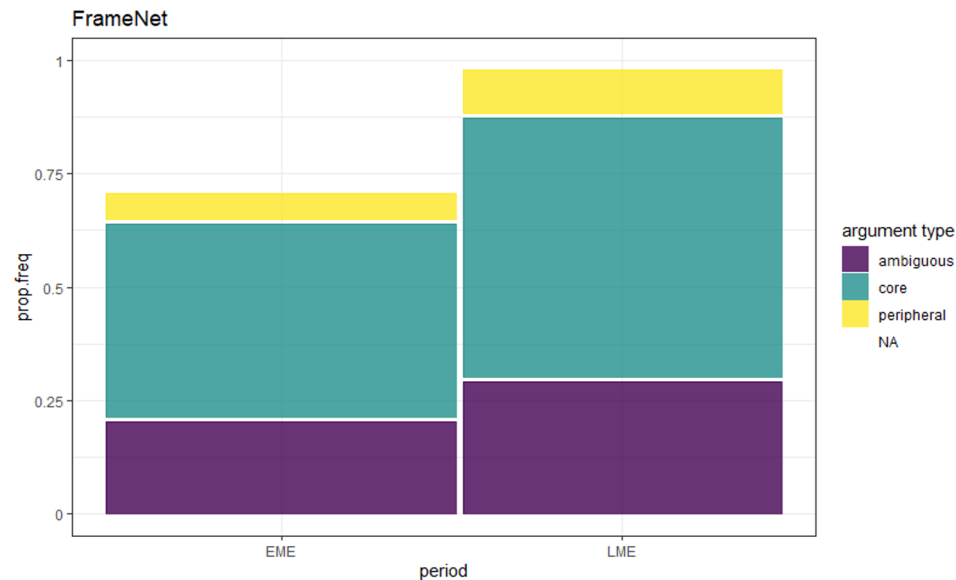
approaches

- manual sub-setting based on specific alternations/ prepositions/ verb types/ semantic roles (Zehentner 2022a; 2023a,b; 2024; in prep. a)

- manual role classification (Zehentner & Hundt 2021a; Zehentner 2022b)

- *FrameNet*-based role/ type classification (Zehentner & Hundt 2021b, in prep.)

- manual feature-based type clustering (Zehentner 2023c)

- automatised feature-based (NLP-derived) type clustering (Zehentner 2021c, d, in prep. b)

- automatised (NLP-derived) sense/role/type disambiguation (SArDEEn-project)

approaches

- manual sub-setting based on specific alternations/ prepositions/ verb types/ semantic roles (Zehentner 2022a; 2023a,b; 2024; in prep. a)

- **manual role classification** (Zehentner & Hundt 2021a; Zehentner 2022b)

- ***FrameNet*-based role/ type classification** (Zehentner & Hundt 2021b, in prep.)

- manual feature-based type clustering (Zehentner 2023c)

- automatised feature-based (NLP-derived) type clustering
(Zehentner 2021c, d, in prep. b)

- automatised (NLP-derived) sense/role/type disambiguation
(SArDEEn-project)

manual/ *FrameNet*: communication verbs

- semantic approach to argument structure constructions (ASCs)
- **coreness** distinction of elements in semantic frames (FEs) *(Ruppenhofer et al.* 2010)

peripheral: "Frame elements that do not introduce additional, independent or distinct events from the main reported event" *(Ruppenhofer et al.* 2010: 20)

- distribution of automatically assigned *FrameNet* labels
- much higher amount of non-captured/ missing instances in EModE

manual/ *FrameNet*: communication verbs

- semantic approach to argument structure constructions (ASCs)
- **coreness** distinction of elements in semantic frames (FEs) *(Ruppenhofer et al. 2010)*

peripheral: "Frame elements that do not introduce additional, independent or distinct events from the main reported event" *(Ruppenhofer et al. 2010: 20)*
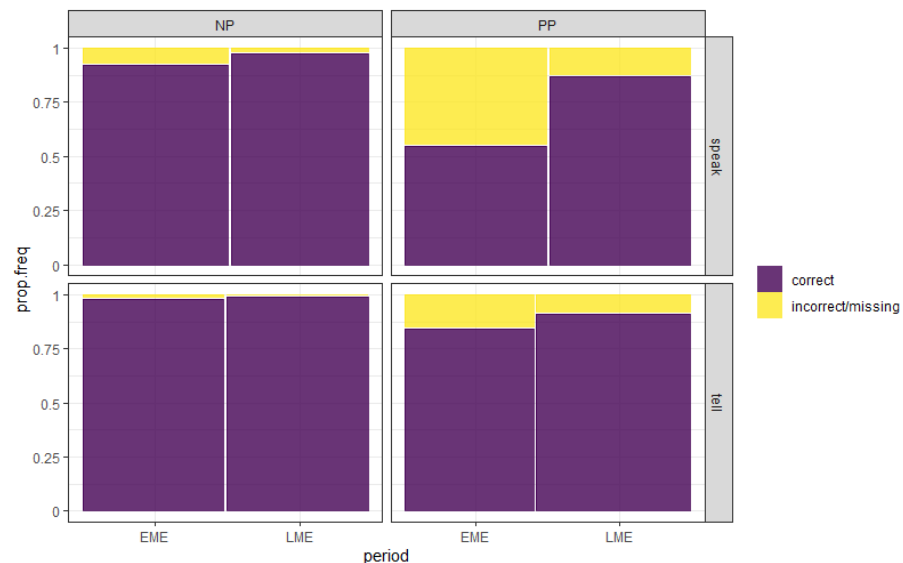
- manual evaluation of *FrameNet* analysis with random sample of 200 instances per verb (*speak/tell*) per period (total N=800)

- 'better' results with NPs/ for LME

- 'worse' results with PPs/ for EME

manual/ *FrameNet*: NP/PP-alternation

- random sample of 1,500 instances (500 per period/ corpus)
- type, semantic role, animacy, definiteness, complexity, etc.

group
- syntactic/semantic
- lexical
- semantic/pragmatic
- (processing) complexity
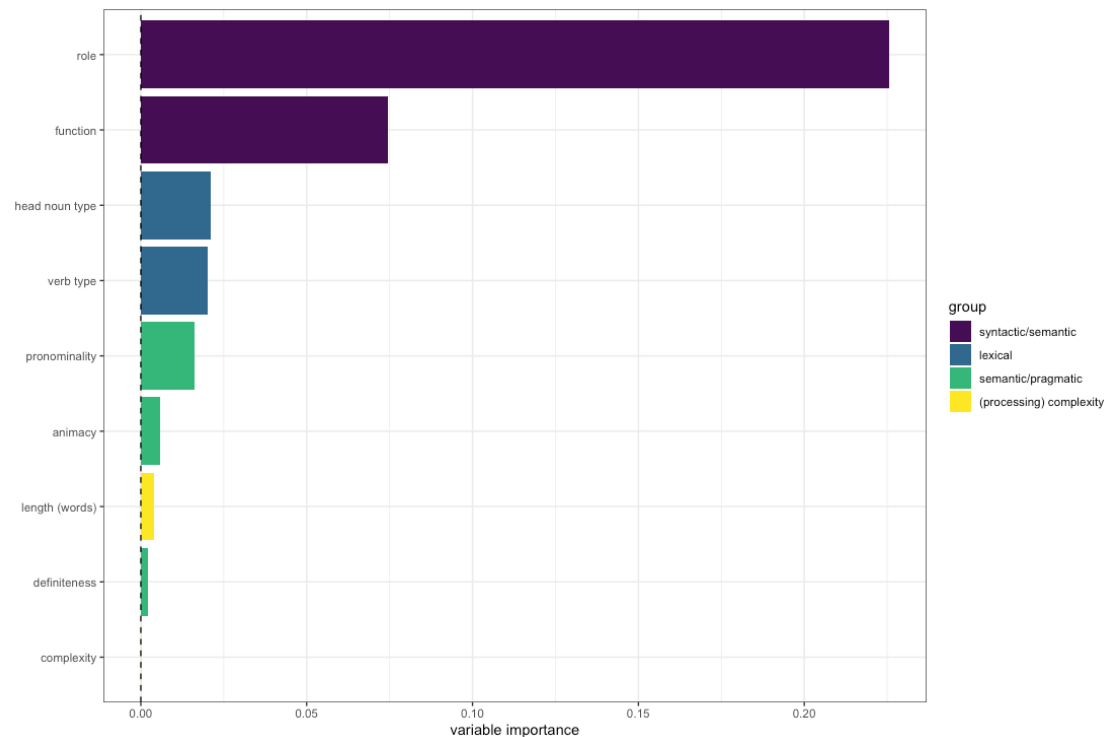
○ conditional random forest analysis (variable importance)

○ function and role as clearly most impactful variables, followed by lexical preferences

approaches

- manual sub-setting based on specific alternations/ prepositions/ verb types/ semantic roles (Zehentner 2022a; 2023a,b; 2024; in prep. a)

- manual role classification (Zehentner & Hundt 2021a; Zehentner 2022b)

- *FrameNet*-based role/ type classification (Zehentner & Hundt 2021b, in prep.)

- manual feature-based type clustering (Zehentner 2023c)

- automatised feature-based (NLP-derived) type clustering (Zehentner 2021c, d, in prep. b)

- automatised (NLP-derived) sense/role/type disambiguation (SArDEEn-project)
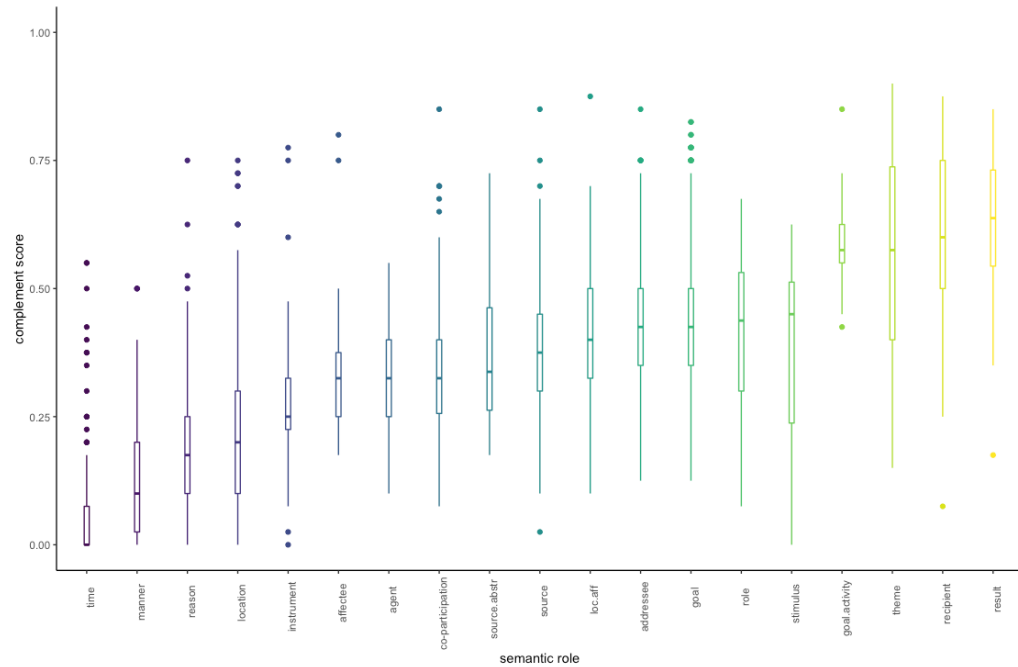
approaches

- manual sub-setting based on specific alternations/ prepositions/ verb types/ semantic roles (Zehentner 2022a; 2023a,b; 2024; in prep. a)

- manual role classification (Zehentner & Hundt 2021a; Zehentner 2022b)

- *FrameNet*-based role/ type classification (Zehentner & Hundt 2021b, in prep.)

- **manual feature-based type clustering** (Zehentner 2023c)

- automatised feature-based (NLP-derived) type clustering (Zehentner 2021c, d, in prep. b)

- automatised (NLP-derived) sense/role/type disambiguation (SArDEEn-project)

manual feature-based type clustering

- random sample of 1,500 instances (500 per period/ corpus)

- 2 annotators: 'optionality', *happen*-paraphrase-ability, *do*-paraphrase-ability, prepositional passive (yes/no/maybe) → translated into joint (average) numerical values for each variable (0-1) and an overall 'complement/ adjunct-hood' value (0-1)

- manual/*FrameNet*-based semantic role annotation

I must and will correspond **with you**.
(LModE; POPE-172X-2,159.77)

- obligatory (1/1)
- *do*-paraphrase: disagreement (0/1)
- *happen*-paraphrase: disagreement (1/1)
- prepositional passive: acceptable? (0.5/0.5)
- **overall complement score: 0.688**
- addressee/ interlocutor
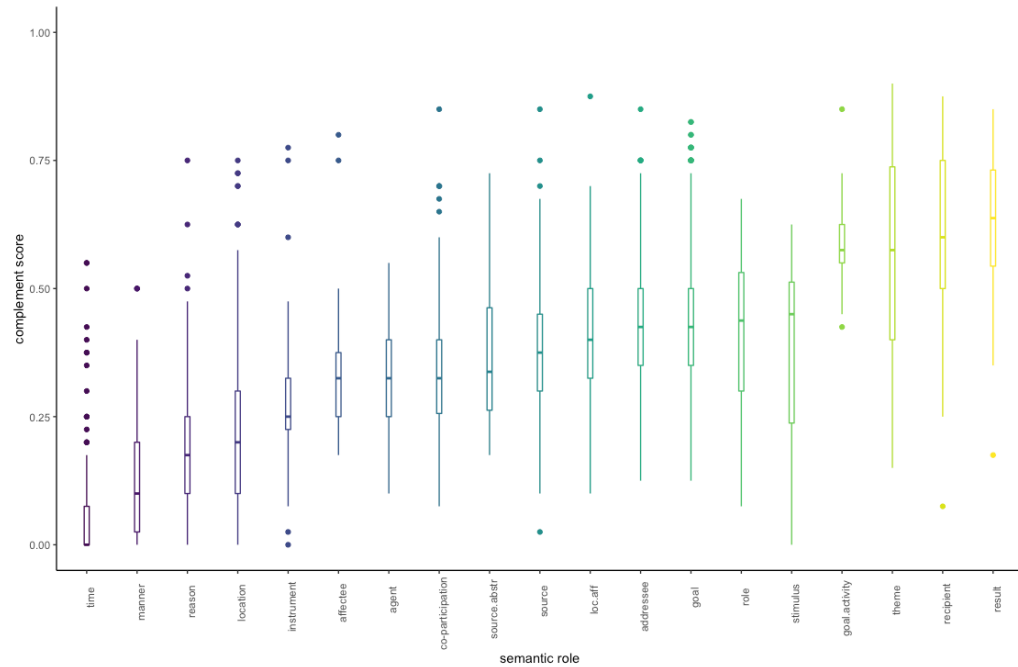- post-verbal position, directly adjacent, single PP

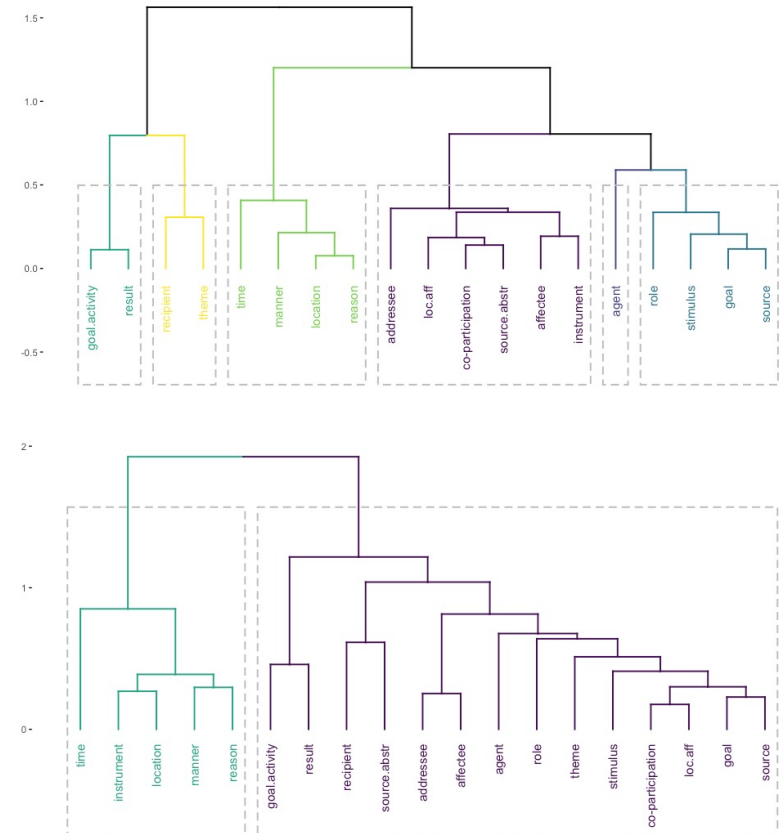# manual feature-based type clustering

- cline in complement scores depending on semantic role
- clusters based on variables
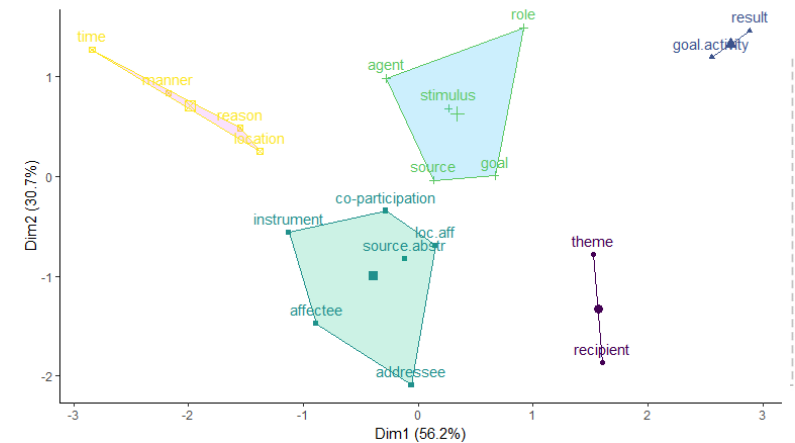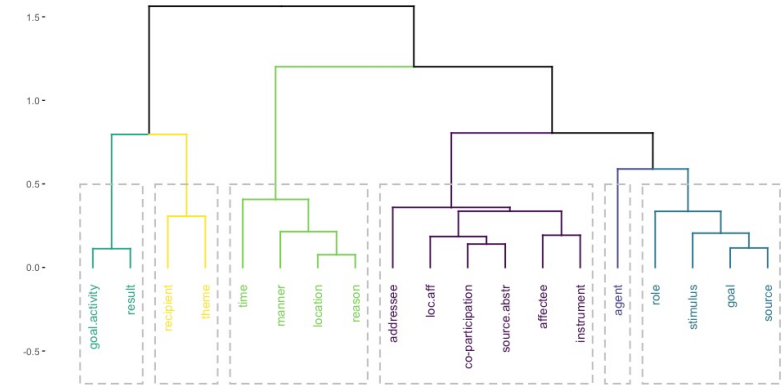
manual feature-based type clustering

- cline in complement scores depending on semantic role
- clusters based on variables

approaches

- manual sub-setting based on specific alternations/ prepositions/ verb types/ semantic roles (Zehentner 2022a; 2023a,b; 2024; in prep. a)

- manual role classification (Zehentner & Hundt 2021a; Zehentner 2022b)

- *FrameNet*-based role/ type classification (Zehentner & Hundt 2021b, in prep.)

- manual feature-based type clustering (Zehentner 2023c)

- automatised feature-based (NLP-derived) type clustering (Zehentner 2021c, d, in prep. b)

- automatised (NLP-derived) sense/role/type disambiguation (SArDEEn-project)

approaches

- manual sub-setting based on specific alternations/ prepositions/ verb types/ semantic roles (Zehentner 2022a; 2023a,b; 2024; in prep. a)

- manual role classification (Zehentner & Hundt 2021a; Zehentner 2022b)

- *FrameNet*-based role/ type classification (Zehentner & Hundt 2021b, in prep.)

- manual feature type clustering (Zehentner 2023c)

- **automatised feature-based (NLP-derived) type clustering**
  (Zehentner 2021c, d, in prep. b)

- automatised (NLP-derived) sense/role/type disambiguation
  (SArDEEn-project)

automatised type clustering

Merlo & Esteve Ferrer (2006)

- lexical information/ lexical classes
- type frequency, transitional probabilities
- entropy
- iterativity

complements

- lower number of verb types (more restrictive semantics)
- stronger association (higher transitional probabilities) between V and NP/PP
- lower entropy (higher predictability)

adjuncts

- higher number of verb types (more open semantics)
- weaker association (lower transitional probabilities) between V and NP/PP
- higher entropy (lower predictability)

automatised type clustering

- clustering based on dissimilarity matrix (gower-distances)
- optimal number of clusters: 2 (or 10+)

	cluster#1	cluster#2
verb lemma	*go*	*turn*
prep lemma	*to*	*in*
NP-head lemma	*him*	*holy writing*
trans.prob	0.014	0.0019
entropy	0.713	0.832
distance.char	1 (-0.66)	6 (0.761)

cluster k-medoids

cluster#1:
more complement-like (?)/ core

- *go to him*
- lower normalised entropy (higher predictability)
- stronger association
- shorter distance
- ...

automatised type clustering

- clustering based on dissimilarity matrix (gower-distances)
- optimal number of clusters: 2 (or 10+)

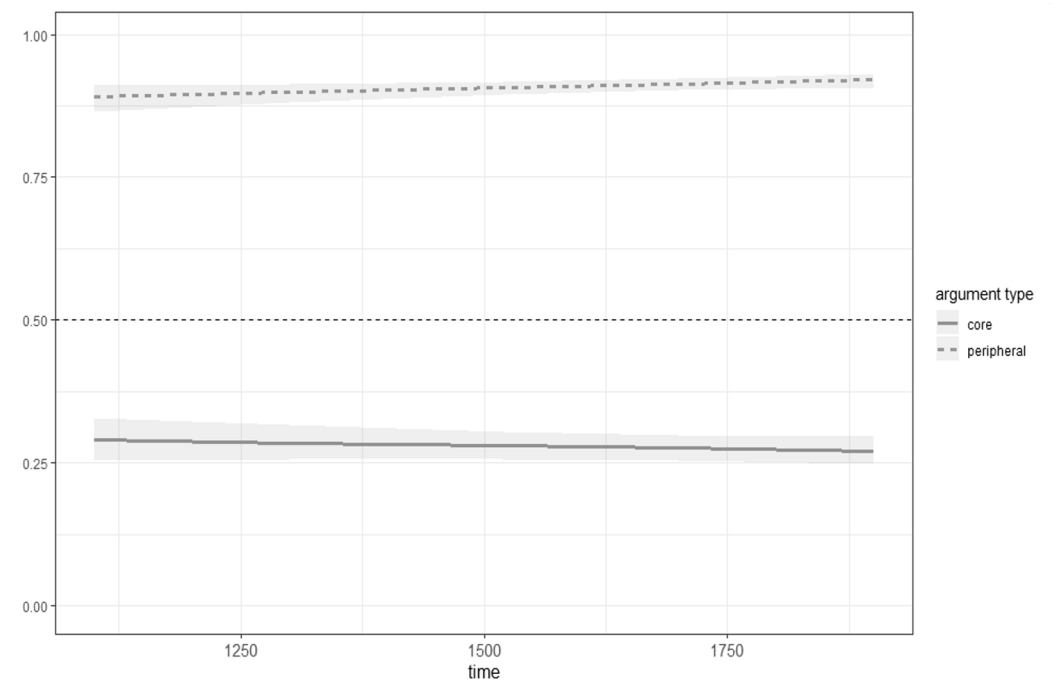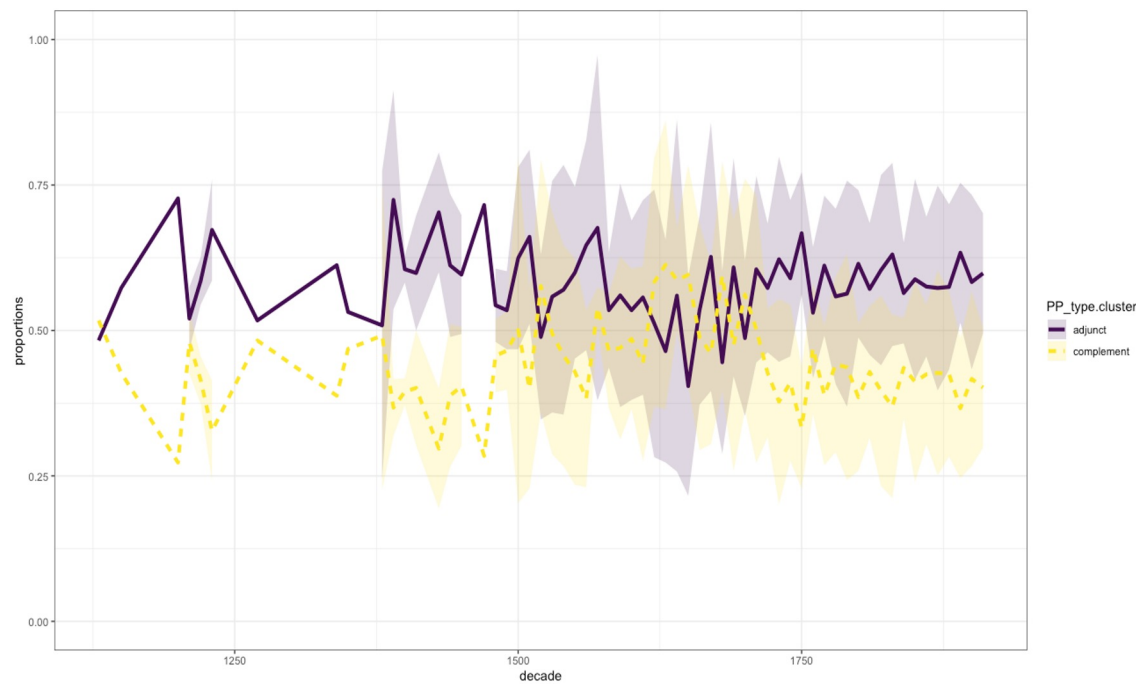|	cluster#1	cluster#2
verb lemma	*go*	*turn*
prep lemma	*to*	*in*
NP-head lemma	*him*	*holy writing*
trans.prob	0.014	0.0019
entropy	0.713	0.832
distance.char	1 (-0.66)	6 (0.761)

cluster k-medoids

cluster#2:
more adjunct-like (?)/ peripheral

- *turn [to sth.] in holy writing*
  - higher normalised entropy (greater variability, lower predictability)
  - weaker association
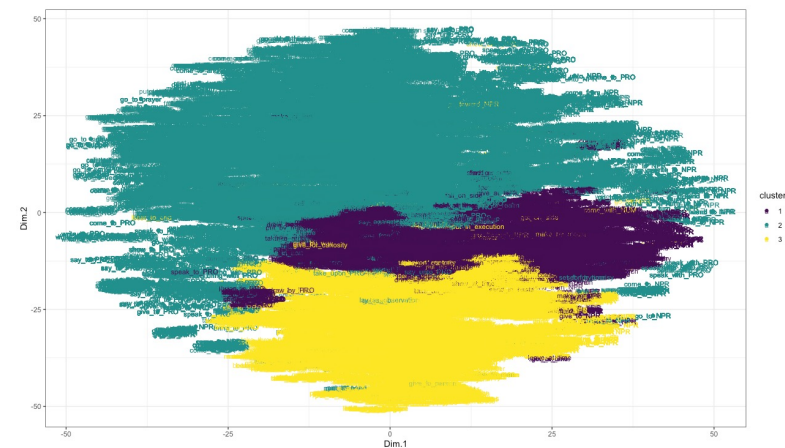 - greater distance
  - ...

automatised type clustering

- clustering based on dissimilarity matrix (gower-distances)
- optimal number of clusters: 2 (or 10+)

automatised type clustering

- clustering based on dissimilarity matrix (gower-distances)
- optimal number of clusters: 2 (or 10+)

o comparison between classification based on manual binary classification (decision tree-classification of sample of 35,000 tokens) and clustering → **overlap: 65.46%**

o main reason for divergences: e.g. goal-PPs classified as 'adjuncts' in classification trees vs 'complement' (cluster#1) in clustering approach

cluster/ function	adjunct	complement
adjunct	**15,451**	3,933
complement	8,126	**7,490**

approaches

- manual sub-setting based on specific alternations/ prepositions/ verb types/ semantic roles (Zehentner 2022a; 2023a,b; 2024; in prep. a)

- manual role classification (Zehentner & Hundt 2021a; Zehentner 2022b)

- *FrameNet*-based role/ type classification (Zehentner & Hundt 2021b, in prep.)

- manual feature-based type clustering (Zehentner 2023c)

- automatised feature-based (NLP-derived) type clustering (Zehentner 2021c, d, in prep. b)

- automatised (NLP-derived) sense/role/type disambiguation (SArDEEN-project)
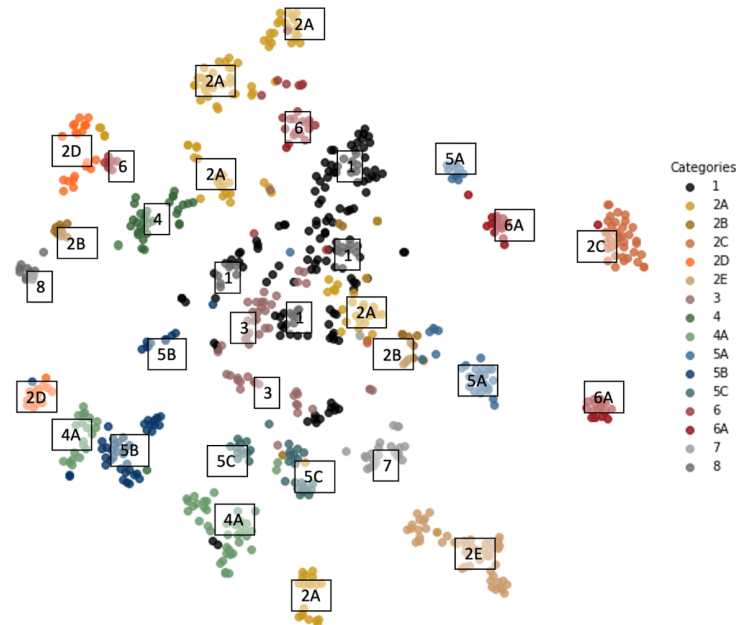
approaches

- manual sub-setting based on specific alternations/ prepositions/ verb types/ semantic roles (Zehentner 2022a; 2023a,b; 2024; in prep. a)

- manual role classification (Zehentner & Hundt 2021a; Zehentner 2022b)

- *FrameNet*-based role/ type classification (Zehentner & Hundt 2021b, in prep.)

- manual feature-based type clustering (Zehentner 2023c)

- automatised feature-based (NLP-derived) type clustering (Zehentner 2021c, d, in prep. b)

- **automatised (NLP-derived) sense/role/type disambiguation** (SArDEEN-project)

automatised sense/role/type disambiguation

- automatic sense disambiguation in NLP

(e.g. Gildea & Jurafsky 2002; Litkowski & Hargraves 2007; Baldwin et al. 2009; Hovy et al. 2010; Palmer et al. 2013, 2022; Hermann et al. 2014; Gong et al. 2018; Schneider et al. 2018; Bhagat et al. 2019; Devlin et al. 2019; Huang 2020; Fonteyn 2021; Lebani & Lenci 2021; Menini et al. 2022; Papadimitriou et al. 2022; Proietti et al. 2022; Chronis et al. 2023; Kauf et al. 2023; Mahowald et al. 2023; Nikolaev & Padó 2023)

t-SNE embeddings of *over* (COHA data; Fonteyn 2021: 17)

MacBERTh

(Manjavacas & Fonteyn 2022)

- BERT-model pre-trained on data from 1450-1950

automatised sense/role/type disambiguation

- automatic sense disambiguation in NLP

    (e.g. Gildea & Jurafsky 2002; Litkowski & Hargraves 2007; Baldwin et al. 2009; Hovy et al. 2010; Palmer et al. 2013, 2022; Hermann et al. 2014; Gong et al. 2018; Schneider et al. 2018; Bhagat et al. 2019; Devlin et al. 2019; Huang 2020; Fonteyn 2021; Lebani & Lenci 2021; Menini et al. 2022; Papadimitriou et al. 2022; Proietti et al. 2022; Chronis et al. 2023; Kauf et al. 2023; Mahowald et al. 2023; Nikolaev & Padó 2023)

- **pilot study (SArDEEn project):**
    - fine-tuning MacBERTh on Middle English data
    - semantic role/ type classification

        (Manjavacas & Zehentner)

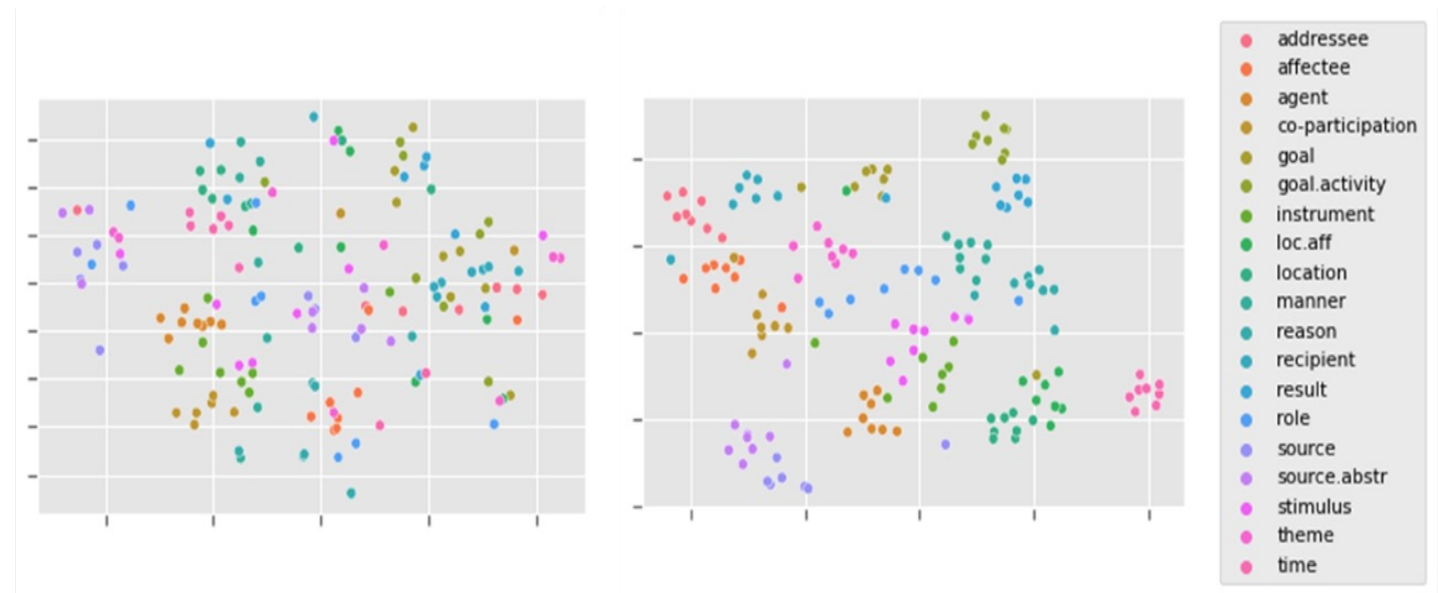# automatised sense/role/type disambiguation

- automatic sense disambiguation in NLP

    (e.g. Gildea & Jurafsky 2002; Litkowski & Hargraves 2007; Baldwin et al. 2009; Hovy et al. 2010; Palmer et al. 2013, 2022; Hermann et al. 2014; Gong et al. 2018; Schneider et al. 2018; Bhagat et al. 2019; Devlin et al. 2019; Huang 2020; Fonteyn 2021; Lebani & Lenci 2021; Menini et al. 2022; Papadimitriou et al. 2022; Proietti et al. 2022; Chronis et al. 2023; Kauf et al. 2023; Mahowald et al. 2023; Nikolaev & Padó 2023)

- **pilot study (SArDEEn project):**

    o fine-tuning MacBERTh on Middle English data

    o semantic role/ type classification

    (Manjavacas & Zehentner)

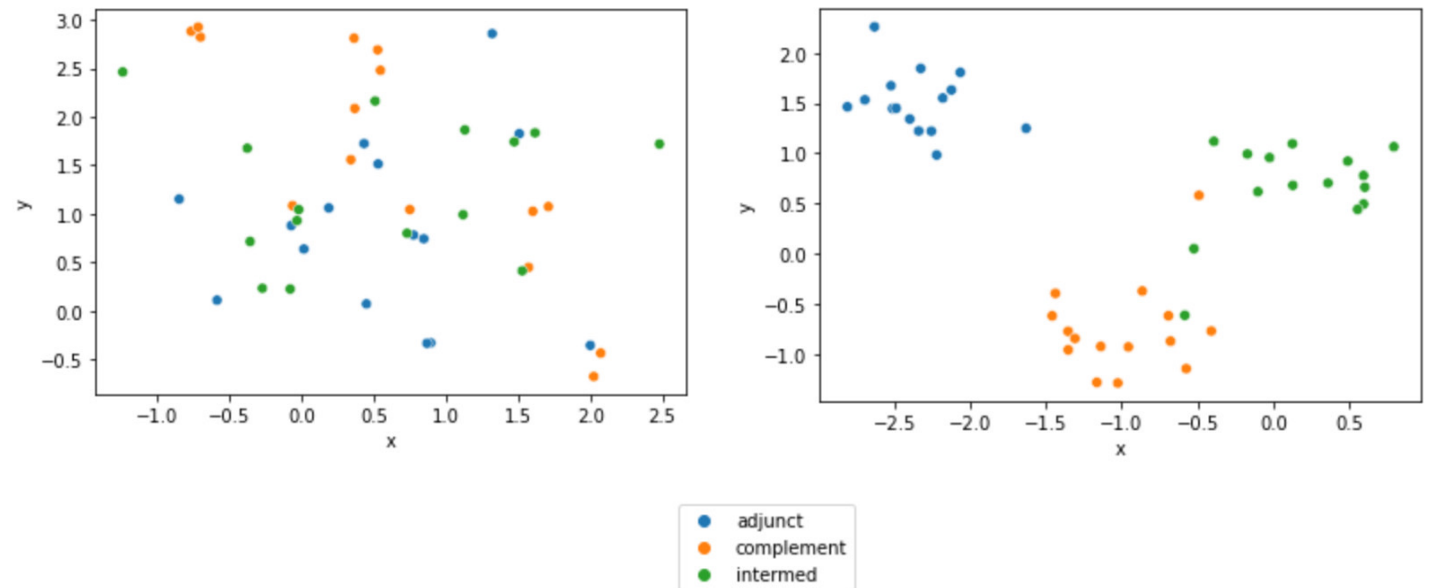# automatised sense/role/type disambiguation

- automatic sense disambiguation in NLP

  (e.g. Gildea & Jurafsky 2002; Litkowski & Hargraves 2007; Baldwin et al. 2009; Hovy et al. 2010; Palmer et al. 2013, 2022; Hermann et al. 2014; Gong et al. 2018; Schneider et al. 2018; Bhagat et al. 2019; Devlin et al. 2019; Huang 2020; Fonteyn 2021; Lebani & Lenci 2021; Menini et al. 2022; Papadimitriou et al. 2022; Proietti et al. 2022; Chronis et al. 2023; Kauf et al. 2023; Mahowald et al. 2023; Nikolaev & Padó 2023)

- **pilot study (SArDEEn project):**

  - fine-tuning MacBERTh on Middle English data

  - semantic role/ type classification

    (Manjavacas & Zehentner)

field	method	f1-micro	f1-macro	accuracy
PP_core_types	finetune	0.727528	0.367033	0.727528
	metric-finetune	0.761938	0.490866	0.761938
PP_role	finetune	0.745787	0.630716	0.745787
	metric-finetune	0.754213	0.712362	0.754213
PP_role_spec	finetune	0.790730	0.700776	0.790730
	metric-finetune	0.717697	0.639832	0.717697
PP_type	finetune	0.813202	0.811254	0.813202
	metric-finetune	0.887640	0.886621	0.887640

talk overview

- **overarching question:**
changes in strategies for argument disambiguation in history of English (synthetic/ morphological → analytic/ syntactic): prepositional patterns

- **issue A:**
challenges relating to historical data (especially Middle English)

- **issue B:**
challenges relating to theoretically and methodologically problematic phenomena (prepositions)

- **main point:**
use of new tools and methods ↔ careful (historical) linguistic consideration

talk overview

- **overarching question:**
  changes in strategies for argument disambiguation in history of English (trade-off synthetic/ morphological → analytic/ syntactic): new approaches, new tools

- **issue A:**
  challenges relating to historical data (especially Middle English)

- **issue B:**
  challenges relating to theoretically and methodologically problematic phenomena

- **main point:**
  use of new tools and methods ←→ careful (historical) linguistic consideration

conclusions

- small, high-quality historical corpora

- data scarcity, specific features of historical data

- linguistically challenging phenomena: theoretical questions, methodological approaches (e.g. prepositional patterns)

- semantic classification of historical data

- new tools and methods to remedy issues with manual annotation/ classification (time-consuming, subjective, error-prone)

- unclear/ varying correspondence between different methods of classification (especially manual annotation/ linguistic intuitions)

- no benefits in using new tools without clear linguistic questions

"accelerat[e] dialogue on how to integrate these models in theoretical [historical] linguistic research, and vice versa" (Fonteyn 2021: 24)

thank you!

eva.zehentner@es.uzh.ch

www.evazehentner.wordpress.com
www.prepcomp.uzh.ch

references

Aissen J. 2003. Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory* 21: 435-483.

Aldezabal, I., M. Aranzabe, K. Gojenola, K. Sarasola & A. Atutxa. 2002. Learning argument/adjunct dictinction for Basque. In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon on Unsupervised Lexical Acquisition*, 42-50. Philadelphia, PA.

Allen, C. 1995. *Case marking and reanalysis: Grammatical relations from Old to Early Modern English*. Oxford: OUP.

Allen, C. 2005. Changes in case marking in NP: From Old English to Middle English. In M. Amberber & H. de Hoop (eds). *Competition and variation in natural languages: The case for case*. Amsterdam: Elsevier, 223-249.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511801686.

Baldwin, T., V. Kordoni & A. Villavicencio. 2009. Prepositions in applications: A survey and introduction to the special issue. *Computational Linguistics* 35(2): 119-149.

Bates, Douglas, Martin Mächler, Ben Bolker & Steven Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1): 1-48. https://doi.org/10.18637/jss.v067.i01.

Baugh, A. & T. Cable. 2002. *A history of the English language*. 5th edn. London: Routledge.

Bergs, A. 2021. Complements and adjuncts". In B. Aarts, A. McMahon & L. Hinrichs (eds.), *The handbook of English linguistics*. Hoboken, NJ: Wiley-Blackwell.

Bhagat, P., A. Varde & A. Feldman. 2019. WordPrep: Word-based preposition prediction tool. In *2019 IEEE International Conference on Big Data (Big Data),* 2169-2176. https://doi.org/10.1109/BigData47090.2019.9005608.

Biber, D., B. Gray, S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education

Claridge, C. 2000. *Multi-word verbs in Early Modern English: A corpus-based study*. Amsterdam: Rodopi.

De Cuypere, L. 2015a. A multivariate analysis of the Old English ACC+DAT double object alternation. *Corpus Linguistics and Linguistic Theory* 11(2): 225-254. https://doi.org/10.1515/cllt20140011.

De Cuypere, L. 2015b. The Old English to dative construction. *English Language and Linguistics* 19(1): 1-26. https://doi.org/10.1017/S1360674314000276.

De Hoop, H. & P. de Swart. 2008. *Differential subject marking.* Dordrecht: Springer.

De la Cruz, J. 1973. A late 13th century change in English structure. *Orbis* 22: 161-167.

Denison, D. 1981. *Aspects of the history of English group-verbs. With particular attention to the syntax of the Ormulum*. Oxford: Oxford University PhD thesis.

Denison, D. 1985. Why Old English had no prepositional passive. *English Studies* 66, 189-204.

references

Denison, D. 2014. *English historical syntax*. London: Longman.

Fischer, O. 1992. Syntax. In N. Blake (ed.), *The Cambridge history of the English language, Vol. 2*, 207-408. Cambridge: CUP.

Fischer, O. & W. van der Wurff. 2006. Syntax. In Hogg, R. & D. Denison (eds). *A history of the English language,* 109-198. Cambridge: CUP.

Flach, S. 2021. Collostructions: An R implementation for the family of collostructional methods. Package version v.0.2.0, https://sfla.ch/collostructions/.

Fox, J. & S. Weisberg. 2019. *An R companion to applied regression*, 3rd edn. Thousand Oaks, CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Garnier, S., N. Ross, R. Rudis, P.Camargo, M. Sciaini & C. Scherer. 2021. viridis - Colorblind-friendly color maps for R. https://sjmgarnier.github.io/viridis/.

Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27(15): 2865-2873. https://doi.org/10.1002/sim.3107.

Gelman, A & J. Hill. 2007. *Data analysis using regression and multilevel/ hierarchical models*. Cambridge: CUP.

Gong, H., S. Bhat & P. Viswanath. 2018. Embedding syntax and semantics of prepositions via tensor decomposition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1*, 986-906.

Grafmiller, J. 2019. jasongraf1/JGmermod: Custom functions for mixed-effects regression models. R package.

Haspelmath, M. 2019. Differential place marking and differential object marking. *Language Typology and Universals* 72(3): 313-334.

Hoffmann, T. 2007. Complements versus adjuncts? A construction grammar account of English prepositional phrases. *Occasional Papers in Language and Linguistics (University of Nairobi)* 3, 92-119.

Hothorn, T., K. Hornik & A. Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3): 651-674.

Hovy, D., S. Tratz & E. Hovy. 2010. What's in a preposition? Dimensions of sense disambiguation for an interesting word class. In *Coling 2010: Poster Volume*, 4554-4562.

Huang, G., J. Wang, H. Tang & X. Ye. 2020. BERT-based contextual semantic analysis for English preposition error correction. *Journal of Physics: Conf. Ser.* 1693: 012115. https://doi.org/10.1088/1742-6596/1693/1/012115.

Huddleston Rodney & Geoffrey Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: CUP.

Iemmolo, G. 2011. *Towards a typological study of differential object marking and differential object indexation*. University of Pavia.

Levshina, N. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: Benjamins. https://doi.org/10.1075/z.195.

Levshina, N. 2018. Anybody (at) home? Communicative efficiency knocking on the Construction Grammar door. *Yearbook of the German Cognitive Linguistics Association* 6: 71-90. https://doi.org/10.1515/gcla-2018-0004.

references

Levshina, N. 2021. Communicative efficiency and differential case marking: A reverse engineering approach. *Linguistics Vanguard* 7(s3): 20190087.

Levy, R. & G. Andrew. 2006. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. *5th International Conference on Language Resources and Evaluation* (LREC 2006).

Litkowski, K. & O. Hargraves. 2007. Argument vs. adjunct. *CL Research Working Paper 07-02*. CL Research, Damascus, MD.

Lüdecke, D. 2018. ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software* 3(26): 77.

Lundskær-Nielsen, T. 1993. *Prepositions in Old and Middle English.* Odense: Odense UP.

Kroch, A., A. Taylor & B. Santorini. 2000. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania, second edition, release 4. www.ling.upenn.edu/hist–corpora/PPCME2–RELEASE–3/index.html.

Kroch, A., B. Santorini & L. Delfs. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania, first edition, release 3. https://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-3/index.html.

Kroch, A., B. Santorini & A. Diertani. 2016. *The Penn Parsed Corpus of Modern British English*. http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1.

Merlo, P. & E. Esteve Ferrer. 2006. The notion of argument in prepositional phrase attachment. *Computational Linguistics* 32(3): 341-378.

Mitchell, Bruce. 1985. *Old English syntax, Vol. 1*. Oxford: Clarendon.

Mustanoja, T. 1960. *A Middle English syntax, Vol. 1*. Helsinki: Société Néophilologique.

Percillier, M. 2018. A toolkit for lemmatising, analysing, and visualising Middle English data. In A. Frank, C. Ivanovic, F. Mambrini, M. Passarotti & C. Sporleder (eds.), *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities*, 153-160.

Pijpops, D., D. Speelman, S. Grondelaers & F. Van de Velde. 2018. Comparing explanations for the Complexity Principle. *Language and Cognition*, 514-543.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org.

Rohdenburg, G. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2): 149-182.

Schneider, N., J. Hwang, V. Srikumar, J. Prange, A. Blodgett, S. Moeller, A. Stern, A. Bitan & O. Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1*, 185-196.

Stefanowitsch, A. & S. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209-243.

University of Zurich UZH

English Department

prepositional
English

PEAS Project, funded by SNF
grant 100015_175987

references

Strang, B. 1970. *A history of English*. London: Methuen & Co.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin & A. Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9(307).

Szmrecsanyi, B. 2012. Analyticity and syntheticity in the history of English. In T. Nevalainen & E. C. Traugott (eds). *The Oxford handbook of the history of English,* 654-665. Oxford: OUP.

Szmrecsyani, B. 2016. An analytic-synthetic spiral in the history of English. In E. van Gelderen (ed.), *Cyclical change continued*, 93-112. Amsterdam: Benjamins. https://doi.org/10.1075/la.227.04szm.

Tagliamonte, S. & H. Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2): 135-178.

Tal, S., K. Smith, J. Culbertson, E. Grossman & I. Arnon. 2022. The impact of information structure on the emergence of differential object marking: An experimental study. *Cognitive Science* 46(3): e13119. https://doi.org/10.1111/cogs.13119.

Traugott, E. C. 1972. *A history of English syntax*. New York, NY: Holt, Rinehart, and Winston.

Traugott, E. C. 1992. Syntax. In R. Hogg (ed.), *The Cambridge history of the English language, Vol. 1*, 168-289. Cambridge: CUP.

Villavicencio, A. 2002. Learning to distinguish PP arguments from adjuncts. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, 84-90. Taipei.

Witzlack-Makarevich, A. & I. Seržant. 2018. Differential argument marking: Patterns of variation. In I. Seržant & A. Witzlack-Makarevich (eds.), *Diachronic typology of differential argument marking*, 1-40. Language Science Press.

Wickham, H. 2016. *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer. https://ggplot2-book.org/.

Winter, B. 2019. *Statistics for linguists: An introduction using R*. New York, NY: Routledge.

Zehentner, E. 2019. *Competition in language change: The rise of the English dative alternation*. Berlin: De Gruyter Mouton.

Zehentner, E. 2022. Ambiguity avoidance as a factor in the rise of the English dative alternation." *Cognitive Linguistics* 33 (1): 3-33. https://doi.org/10.1515/cog-2021-0018.

Zuur, A., E. Ieno, N. Walker, A. Saveliev & G. Smith. 2009. *Mixed effects models and extensions in ecology with R*. New York, NY: Springer.