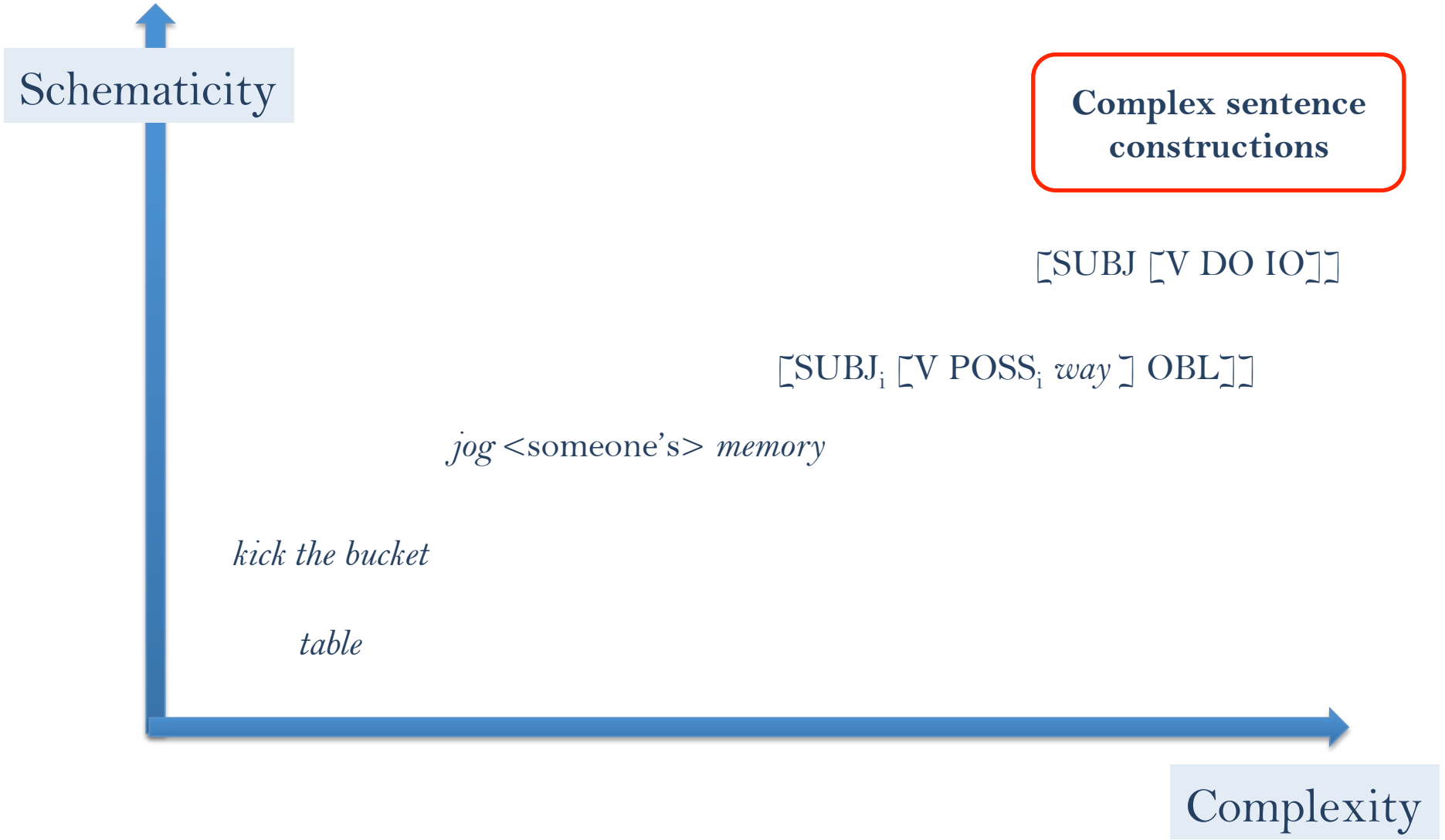
Tilburg University
Formulaic Language Research Network
28-30 March 2012
Tilburg University

RWTH AACHEN UNIVERSITY

Genre-specific properties of adverbial clauses in English academic writing and newspaper texts

Elma Kerz & Daniel Wiechmann

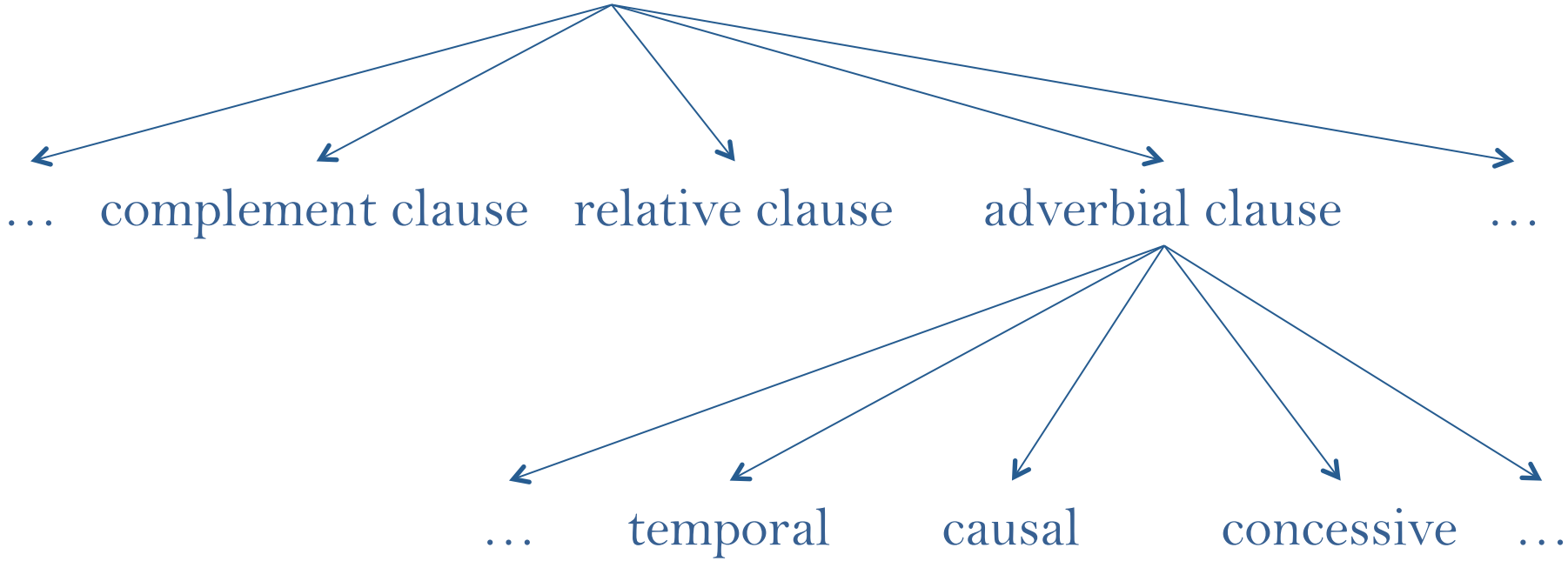
A constructionist perspective on patterns



Complex sentences

(minimally) **bi-clausal construction**

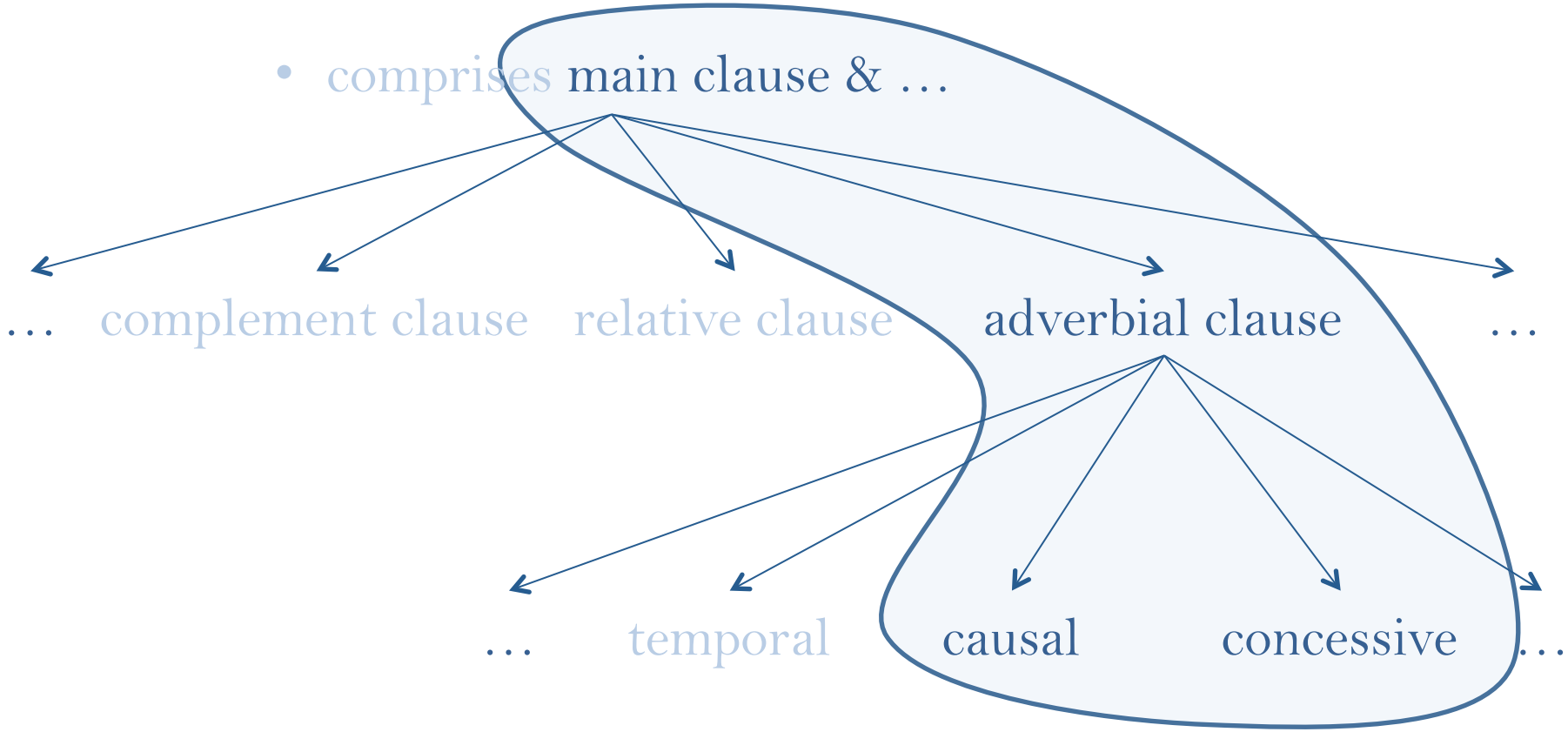
- comprises main clause & …



Complex sentences

(minimally) **bi-clausal construction**

- comprises main clause & …



Genre-specific patterns

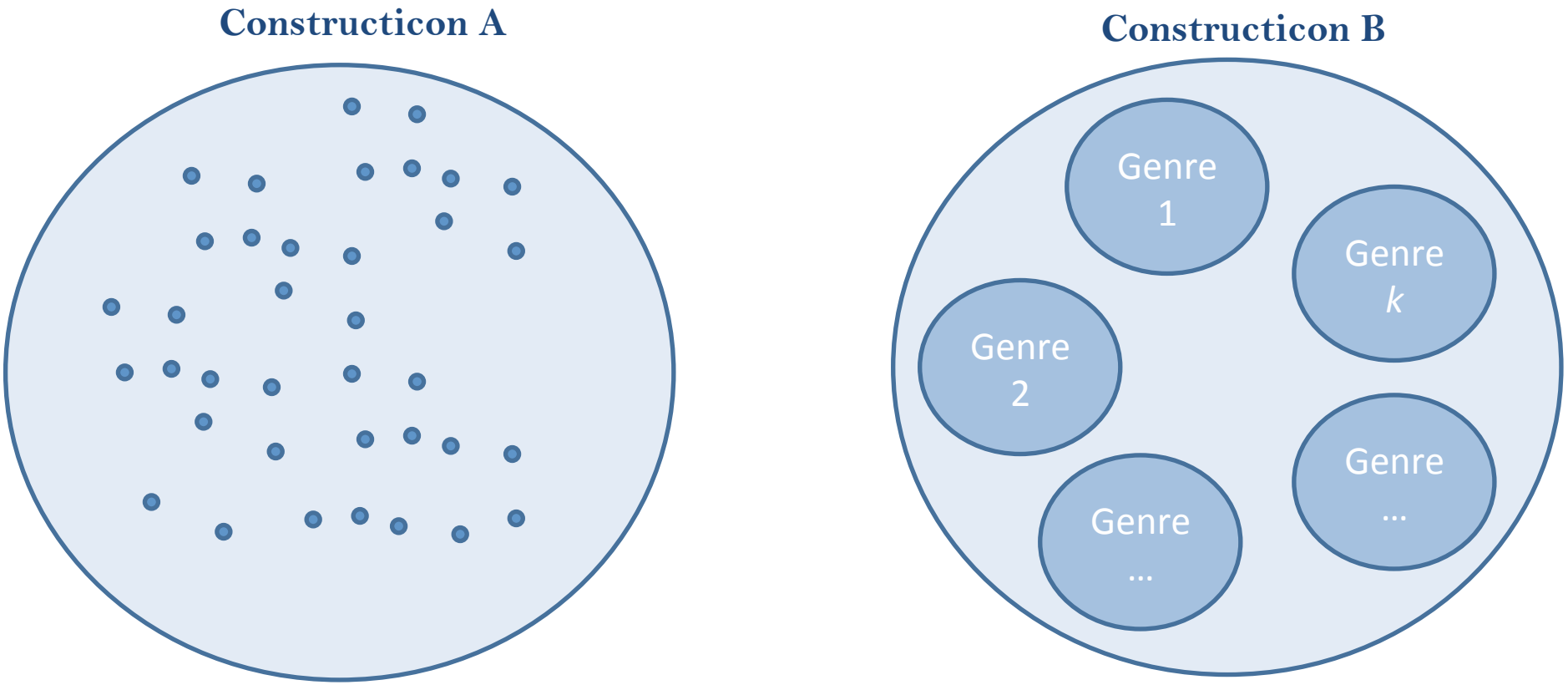
- Killgarriff (2003: 54):
  - "General English" is a theoretically difficult notion, and a language can be seen as a modest core of lexis and constructions, plus a wide array of different sublanguages, as used in each of a myriad of human activities."

- "any patterns generalized for all of English are not likely to be valid for any actual text or register – rather, **generalized patterns would merely level the important patterns of use found across registers**" (Biber et al. 1998: 83)

- Case in point: **Academic writing***

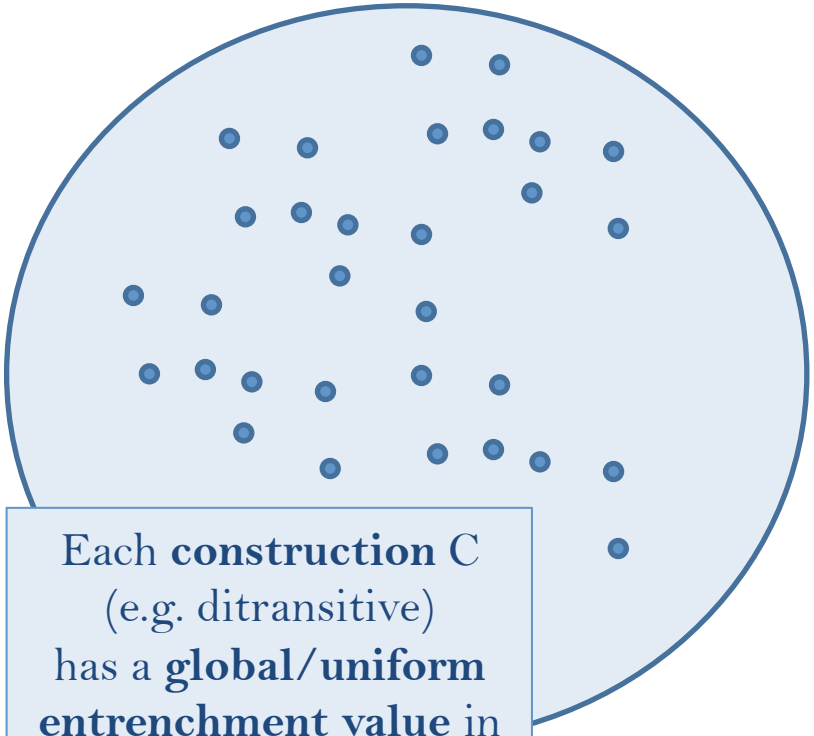
(A growing number of studies looking at recurring multiword expressions (e.g. Biber, Conrad & Cortes 2004, Biber 2006, Hyland 2008, Simpson-Vlach & Ellis 2010)

Two views of the constructicon



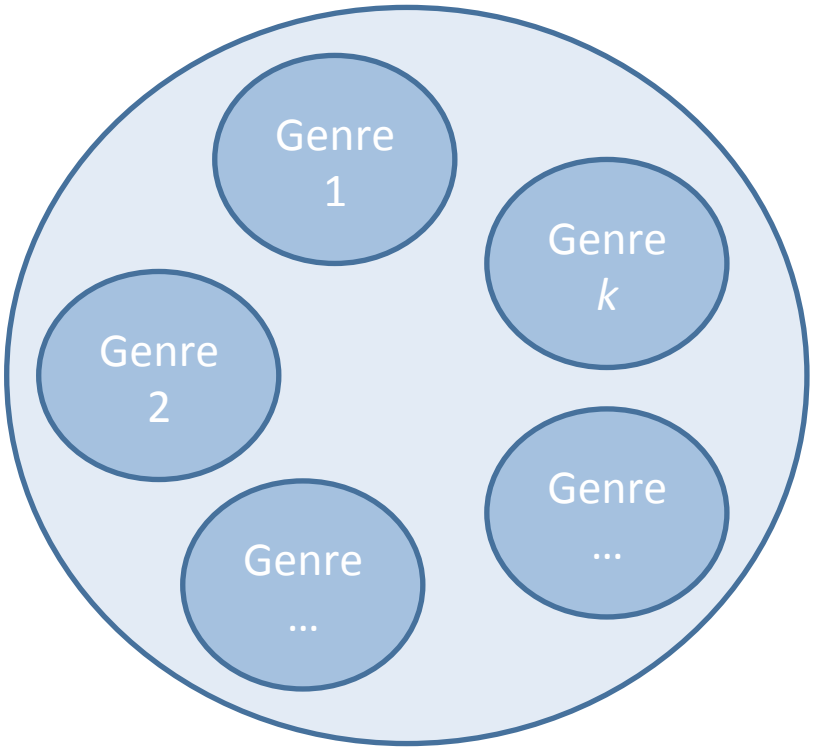
Two views of the constructicon

Constructicon A



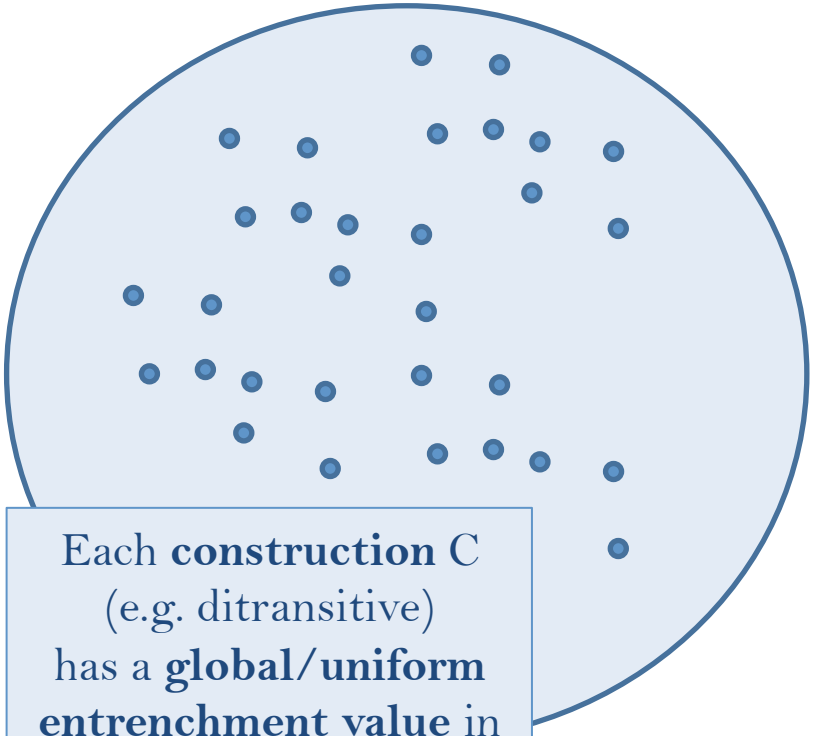
Each **construction** C (e.g. ditransitive) has a **global/uniform entrenchment value** in speaker's mind

Constructicon B



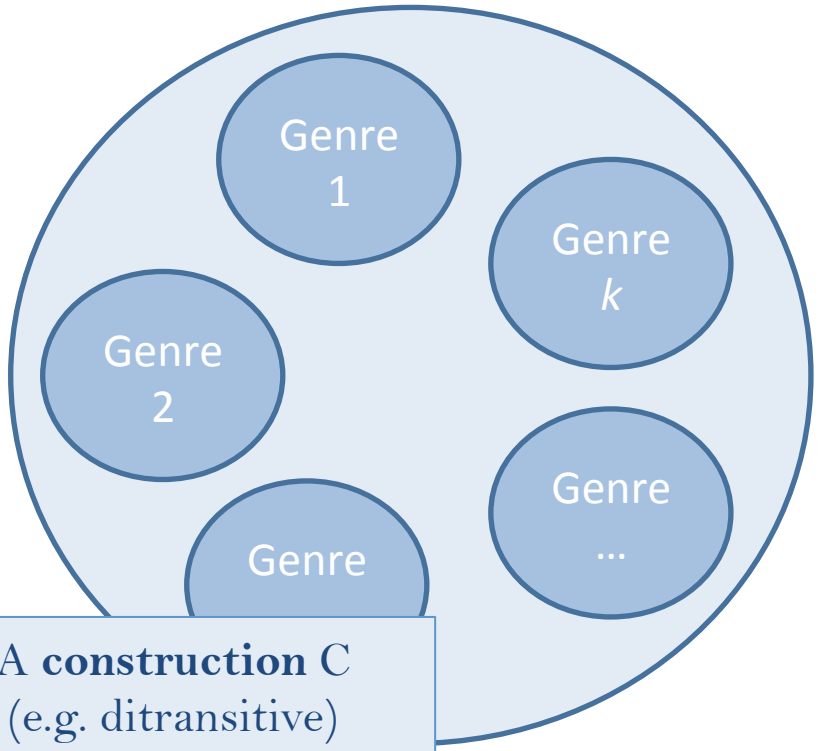
Two views of the constructicon

Constructicon A



Each **construction** C (e.g. ditransitive) has a **global/uniform entrenchment value** in speaker's mind

Constructicon B



A **construction** C (e.g. ditransitive) has a **local/genre-specific entrenchment value** in speaker's mind

Aim of the study

- **Assumptions:** Grammar can be conceived of as a set of sub-repositories encompassing genre-specific constructions

- **Goal** : Detect conventionalized & cognitively entrenched (~strongly represented) types of concessive and reason constructions in the domain of academic writing

Corpus & data used in the study

- **Academic writing**:
  - ~700,000 words extracted from 100 peer-reviewed research articles covering a wide range of academic disciplines
- **Newspaper**:
  - same-sized sample was extracted from the newspaper sub-component of the BNC
- **Query for subordinators**
  - reason clauses {*as*, *because*, *since*}
  - concessive clauses {*although*, *even though*, *though*, *while*, *even if*, *whereas*}

- A total of **1911 data points** coded in terms of **15 variables**
  - Academic writing ($N_{Concessive} = 575$; $N_{Reason} = 471$)
  - News ($N_{Concessive} = 517$; $N_{Reason} = 348$)

Overview: Variables investigated

no	variable	name	scale	values
dep	genre	GENRE	factor	2
1	**subordinator**	**SUB**	**factor**	**9**
1b	subordinator (simplified)	SUB.SIMPLE	factor	7
2	**semantic type**	**TYPE**	**factor**	**2**
3	**relative position of AdvCl**	**POS**	**factor**	**3**
4	**deranking**	**DERANKED**	**factor**	**2**
5	**cross-sentential anaphor**	**CROSS.ANAPH**	**factor**	**2**
6	**complexity of AdvCl**	**COMPLEX**	**factor**	**2**
7	**clause serialization**	**PATTERN**	**factor**	**19**
7b	clause serialization (simplified)	PATTERN.SIMPLER	factor	4
8	**presence matrix clause**	**MAT**	**factor**	**2**
9	**tense of AdvCl**	**TENSE**	**factor**	**4**
10	**aspect of AdvCl**	**ASPECT**	**factor**	**2**
11	**presence of modal verb**	**MODAL.VERB**	**factor**	**2**
12	**voice of AdvCl**	**VOICE**	**factor**	**2**
13	**sentence initial adverb**	**INITIAL.ADV**	**factor**	**14**
13b	sentence initial adverb (simplified)	INI.ADV.SIMPLER	factor	6
14	**within-sentence anaphor**	**WITHIN.ANAPH**	**factor**	**2**
15	**proportional size adverbial clause**	**LENGTHAC.BINNED**	**factor**	**3**
15b	length AdvCl	LENGTH.AC.TARGET	continuous	
15c	length sentence	LENGTH.ITEM	continuous	
15d	length reference clause	LENGTH.RC	continuous	
15e	combined length AdvCl & ref clause	SUM.RC.AC	continuous	
15f	proportional size adverbial clause	PROP.AC	continuous	
15g	proportional size adverbial clause	PROP.AC.NARROW	continuous	

Our **goal** is to **detect configurations** (factor level combinations) that occur w/ above chance frequency (**entrenched types**)

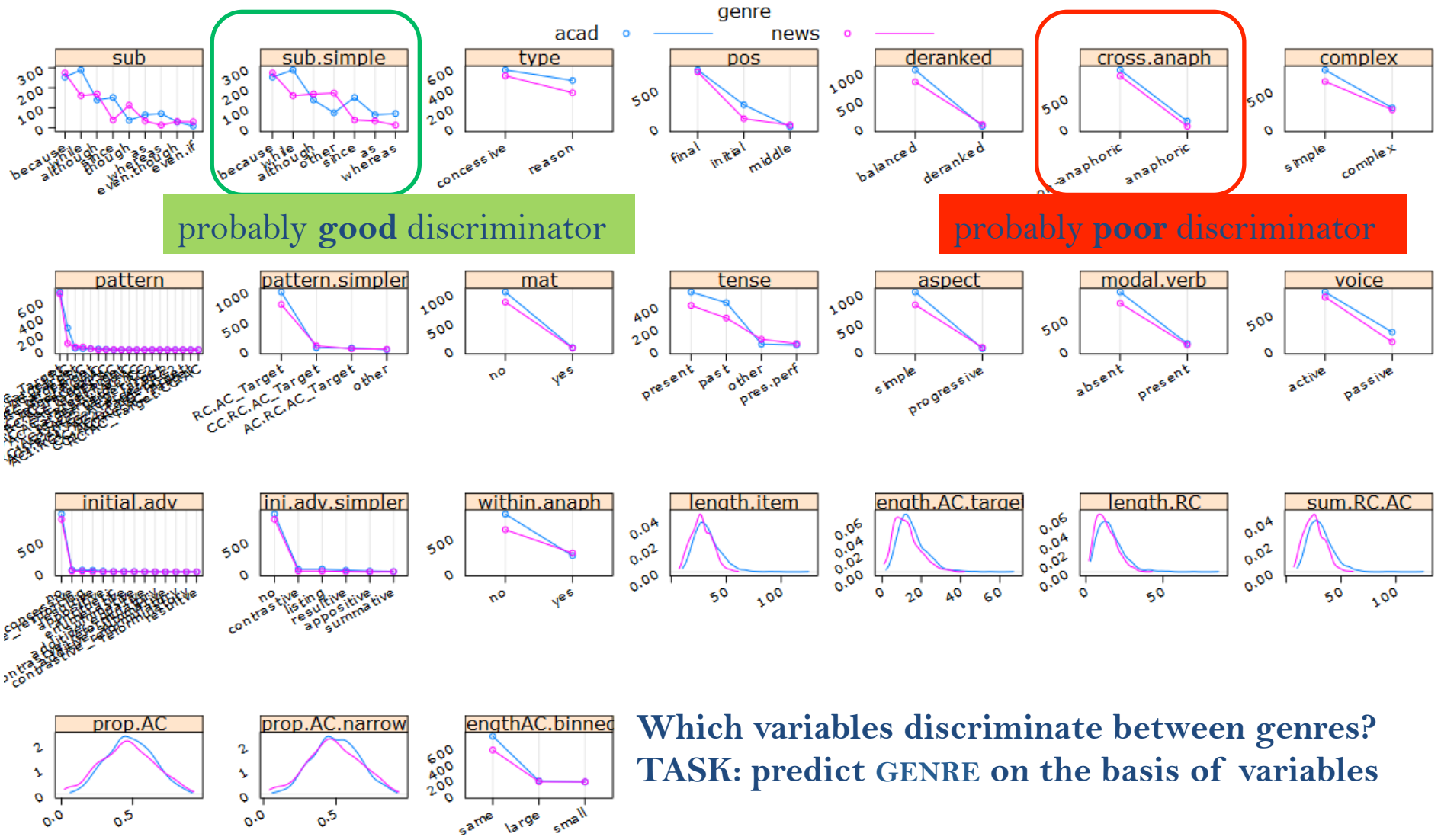
But data are too complex to evaluate frequencies of all possible configurations on the basis of the available data (~ 2k data points)

Reduce complexity → focus on those VARs that discriminate between genres

Reducing the search space

Identifying most relevant variables

Descriptive statistics (trellis plots)



Which variables discriminate between genres?
TASK: predict GENRE on the basis of variables

How to identify discriminating variables

Many conceivable solutions:

- many predictive models/classifiers could do the job (e.g. logistic regression, support vector machines, classification/regression trees, …)

Method of choice:

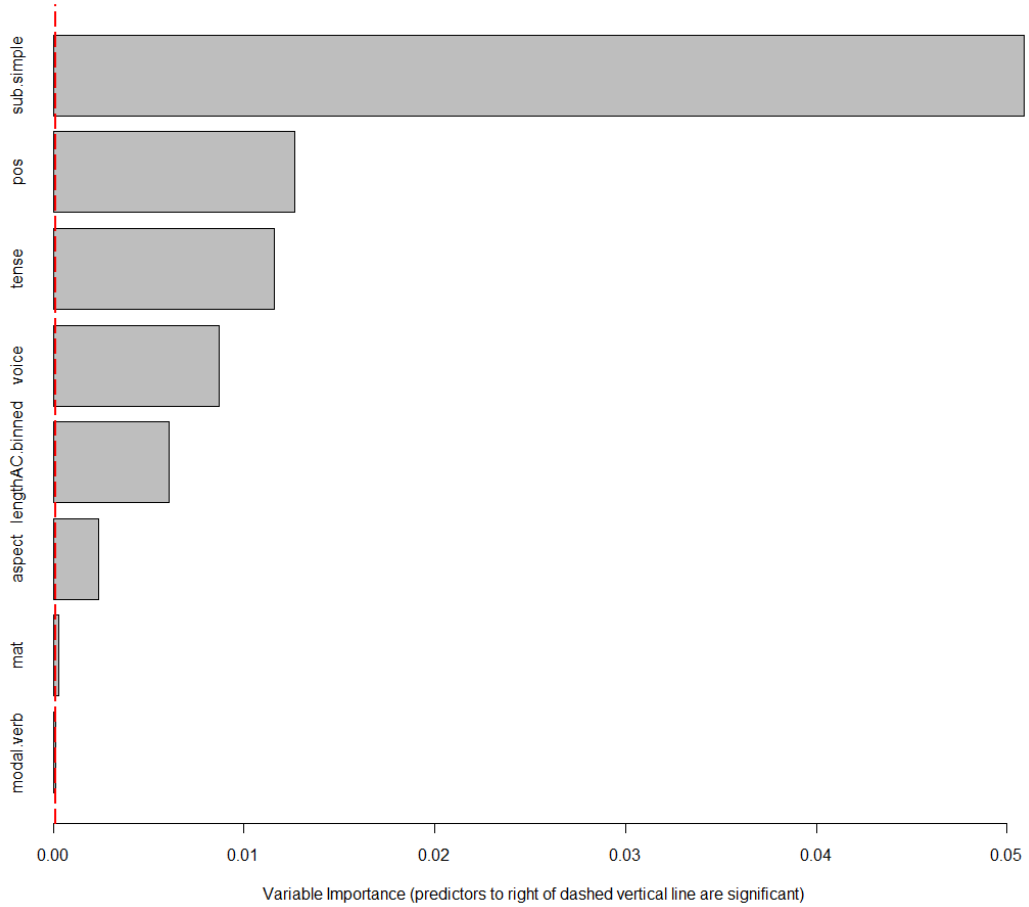
- **random forests of conditional inference trees** (Strobl et al. 2008, 2009)

Advantages of the employed method:

- Over traditional tree-based models:
  - conditional inference trees do not favor variables with many factor levels (or large ranges)
- Over regression:
  - they can handle correlating predictors
- output of random forests is more robust than that of single-tree models

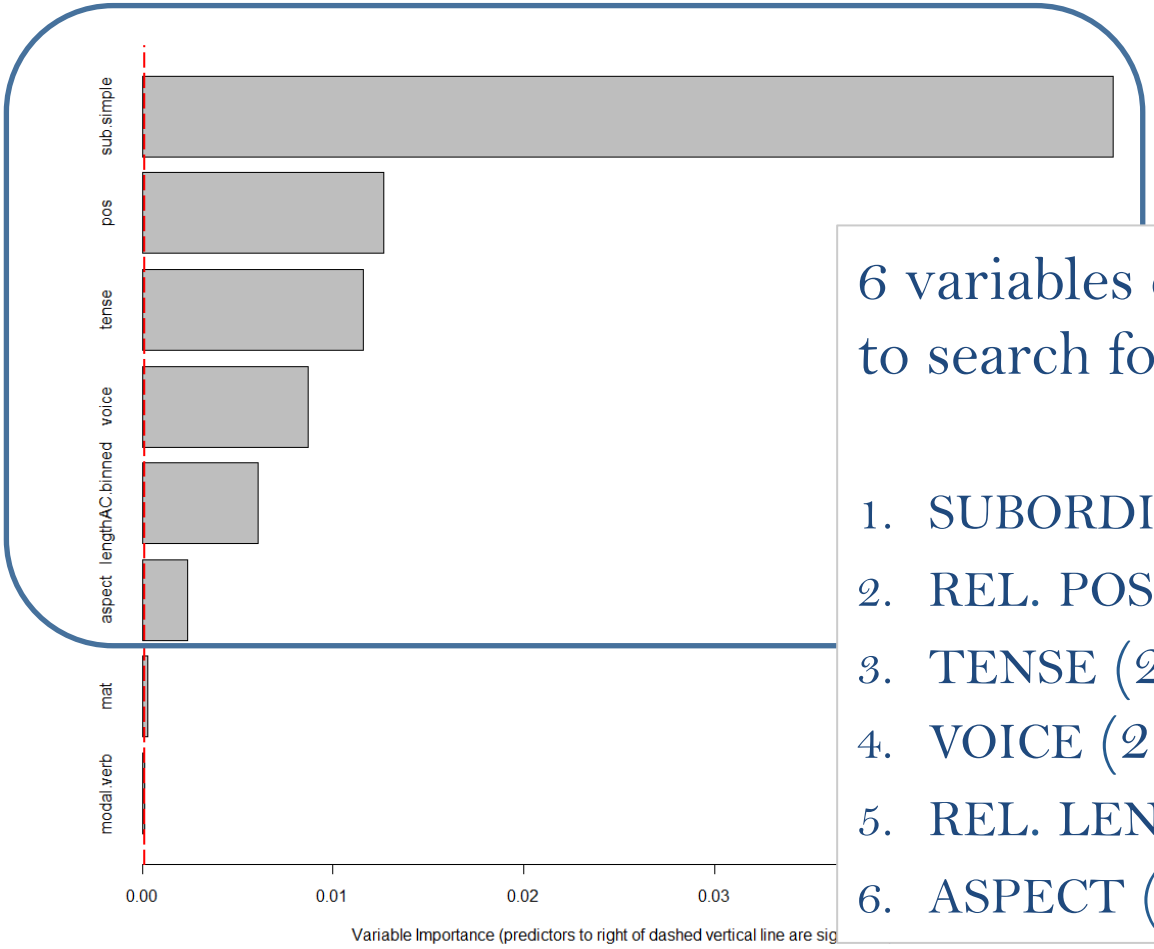
Results:

(Conditional permutation) **Variable Importance** (Strobl 2009)



Results:

(Conditional permutation) **Variable Importance** (Strobl 2009)



6 variables define the space in which to search for patterns

1. SUBORDINATOR (7 levels)
2. REL. POSITION (2 levels)
3. TENSE (2 levels)
4. VOICE (2 levels)
5. REL. LENGTH (binned; 3 levels)
6. ASPECT (2 levels)

Finding Patterns
(entrenched constructions)

How to find patterns (entrenched construction)

Method of choice:

- **Configural Frequency Analysis** (**CFA**; von Eye 1990)

- method for categorical data analysis
- CFA evaluates cell-frequencies in an n-dimensional contingency table
- It compares the observed frequency of a configuration with the expected frequency of that configuration
- configurations with above chance frequency are TYPES (= deeply entrenched)

How to find patterns (entrenched construction)

When we cross all (6+1=) 7 factors (SUBORDINATOR, TENSE, ASPECT, VOICE, RELATIVE LENGTH, RELATIVE POSITION and GENRE), we get…

- **7560** possible configurations
- **400** configurations are attested
- **43** configurations occur at least 10 times
- **21** configurations are statistically significantly more frequent than expected (=TYPES)
- **6** configurations are TYPES & expected frequency > 5

21 detected TYPES

	genre	subordinator	rel. position	tense	aspect	voice	rel. length	Freq	Exp	Cont.chisq	P.adj.bin	Q
1	acad	WHILE	INITIAL	PRESENT	SIMPLE	ACTIVE	SAME	46	12.41	90.96	0.00	0.018
2	acad	BECAUSE	FINAL	PRESENT	SIMPLE	ACTIVE	LARGE	33	12.97	30.91	0.02	0.011
3	acad	SINCE	INITIAL	PRESENT	SIMPLE	ACTIVE	SAME	27	5.25	90.00	0.00	0.011
4	acad	WHEREAS	FINAL	PAST	SIMPLE	ACTIVE	SAME	26	6.38	60.34	0.00	0.01
5	acad	WHILE	FINAL	PAST	SIMPLE	PASSIVE	SAME	25	7.92	36.88	0.01	0.009
6	acad	SINCE	INITIAL	PRESENT	SIMPLE	ACTIVE	SMALL	13	1.25	110.21	0.00	0.006
7	acad	SINCE	INITIAL	PAST	SIMPLE	PASSIVE	SAME	8	0.94	53.03	0.05	0.004
8	acad	OTHER	MIDDLE	NO	NO	NO	SMALL	4	0.00	Inf	0.00	0.002
9	acad	ALTHOUGH	FINAL	NO	NO	NO	SAME	3	0.00	3744.00	0.00	0.002
10	acad	WHILE	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
11	acad	ALTHOUGH	FINAL	NO	NO	NO	SMALL	2	0.00	6662.67	0.00	0.001
1	news	BECAUSE	FINAL	OTHER	SIMPLE	ACTIVE	SAME	22	6.81	33.92	0.02	0.008
2	news	OTHER	MIDDLE	NO	NO	NO	SMALL	9	0.00	Inf	0.00	0.005
3	news	OTHER	FINAL	NO	NO	NO	SAME	8	0.00	39984.00	0.00	0.004
4	news	ALTHOUGH	FINAL	NO	NO	NO	SMALL	4	0.00	31992.00	0.00	0.002
5	news	ALTHOUGH	FINAL	NO	NO	NO	SAME	4	0.00	7992.00	0.00	0.002
6	news	ALTHOUGH	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
7	news	ALTHOUGH	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
8	news	WHILE	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
9	news	OTHER	FINAL	NO	NO	NO	SMALL	2	0.00	9996.00	0.00	0.001
10	news	ALTHOUGH	FINAL	NO	NO	NO	LARGE	2	0.00	7996.00	0.00	0.001

21 detected TYPES

configurations (patterns)

	genre	subordinator	rel. position	tense	aspect	voice	rel. length	Freq	Exp	Cont.chisq	P.adj.bin	Q
1	acad	WHILE	INITIAL	PRESENT	SIMPLE	ACTIVE	SAME	46	12.41	90.96	0.00	0.018
2	acad	BECAUSE	FINAL	PRESENT	SIMPLE	ACTIVE	LARGE	33	12.97	30.91	0.02	0.011
3	acad	SINCE	INITIAL	PRESENT	SIMPLE	ACTIVE	SAME	27	5.25	90.00	0.00	0.011
4	acad	WHEREAS	FINAL	PAST	SIMPLE	ACTIVE	SAME	26	6.38	60.34	0.00	0.01
5	acad	WHILE	FINAL	PAST	SIMPLE	PASSIVE	SAME	25	7.92	36.88	0.01	0.009
6	acad	SINCE	INITIAL	PRESENT	SIMPLE	ACTIVE	SMALL	13	1.25	110.21	0.00	0.006
7	acad	SINCE	INITIAL	PAST	SIMPLE	PASSIVE	SAME	8	0.94	53.03	0.05	0.004
8	acad	OTHER	MIDDLE	NO	NO	NO	SMALL	4	0.00	Inf	0.00	0.002
9	acad	ALTHOUGH	FINAL	NO	NO	NO	SAME	3	0.00	3744.00	0.00	0.002
10	acad	WHILE	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
11	acad	ALTHOUGH	FINAL	NO	NO	NO	SMALL	2	0.00	6662.67	0.00	0.001
1	news	BECAUSE	FINAL	OTHER	SIMPLE	ACTIVE	SAME	22	6.81	33.92	0.02	0.008
2	news	OTHER	MIDDLE	NO	NO	NO	SMALL	9	0.00	Inf	0.00	0.005
3	news	OTHER	FINAL	NO	NO	NO	SAME	8	0.00	39984.00	0.00	0.004
4	news	ALTHOUGH	FINAL	NO	NO	NO	SMALL	4	0.00	31992.00	0.00	0.002
5	news	ALTHOUGH	FINAL	NO	NO	NO	SAME	4	0.00	7992.00	0.00	0.002
6	news	ALTHOUGH	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
7	news	ALTHOUGH	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
8	news	WHILE	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
9	news	OTHER	FINAL	NO	NO	NO	SMALL	2	0.00	9996.00	0.00	0.001
10	news	ALTHOUGH	FINAL	NO	NO	NO	LARGE	2	0.00	7996.00	0.00	0.001

21 detected TYPES

stats for configurations (patterns)

	genre	subordinator	rel. position	tense	aspect	voice	rel. length	Freq	Exp	Cont.chisq	P.adj.bin	Q
1	acad	WHILE	INITIAL	PRESENT	SIMPLE	ACTIVE	SAME	46	12.41	90.96	0.00	0.018
2	acad	BECAUSE	FINAL	PRESENT	SIMPLE	ACTIVE	LARGE	33	12.97	30.91	0.02	0.011
3	acad	SINCE	INITIAL	PRESENT	SIMPLE	ACTIVE	SAME	27	5.25	90.00	0.00	0.011
4	acad	WHEREAS	FINAL	PAST	SIMPLE	ACTIVE	SAME	26	6.38	60.34	0.00	0.01
5	acad	WHILE	FINAL	PAST	SIMPLE	PASSIVE	SAME	25	7.92	36.88	0.01	0.009
6	acad	SINCE	INITIAL	PRESENT	SIMPLE	ACTIVE	SMALL	13	1.25	110.21	0.00	0.006
7	acad	SINCE	INITIAL	PAST	SIMPLE	PASSIVE	SAME	8	0.94	53.03	0.05	0.004
8	acad	OTHER	MIDDLE	NO	NO	NO	SMALL	4	0.00	Inf	0.00	0.002
9	acad	ALTHOUGH	FINAL	NO	NO	NO	SAME	3	0.00	3744.00	0.00	0.002
10	acad	WHILE	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
11	acad	ALTHOUGH	FINAL	NO	NO	NO	SMALL	2	0.00	6662.67	0.00	0.001
1	news	BECAUSE	FINAL	OTHER	SIMPLE	ACTIVE	SAME	22	6.81	33.92	0.02	0.008
2	news	OTHER	MIDDLE	NO	NO	NO	SMALL	9	0.00	Inf	0.00	0.005
3	news	OTHER	FINAL	NO	NO	NO	SAME	8	0.00	39984.00	0.00	0.004
4	news	ALTHOUGH	FINAL	NO	NO	NO	SMALL	4	0.00	31992.00	0.00	0.002
5	news	ALTHOUGH	FINAL	NO	NO	NO	SAME	4	0.00	7992.00	0.00	0.002
6	news	ALTHOUGH	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
7	news	ALTHOUGH	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
8	news	WHILE	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
9	news	OTHER	FINAL	NO	NO	NO	SMALL	2	0.00	9996.00	0.00	0.001
10	news	ALTHOUGH	FINAL	NO	NO	NO	LARGE	2	0.00	7996.00	0.00	0.001

6 detected TYPES with exp. Frequency > 5

	genre	subordinator	rel. position	tense	aspect	voice	rel. length	Freq	Exp	Cont.chisq	P.adj.bin	Q
1	**acad**	**WHILE**	**INITIAL**	**PRESENT**	**SIMPLE**	**ACTIVE**	**SAME**	**46**	**12.41**	**90.96**	**0.00**	**0.018**
2	**acad**	**BECAUSE**	**FINAL**	**PRESENT**	**SIMPLE**	**ACTIVE**	**LARGE**	**33**	**12.97**	**30.91**	**0.02**	**0.011**
3	**acad**	**SINCE**	**INITIAL**	**PRESENT**	**SIMPLE**	**ACTIVE**	**SAME**	**27**	**5.25**	**90.00**	**0.00**	**0.011**
4	**acad**	**WHEREAS**	**FINAL**	**PAST**	**SIMPLE**	**ACTIVE**	**SAME**	**26**	**6.38**	**60.34**	**0.00**	**0.01**
5	**acad**	**WHILE**	**FINAL**	**PAST**	**SIMPLE**	**PASSIVE**	**SAME**	**25**	**7.92**	**36.88**	**0.01**	**0.009**
6	acad	SINCE	INITIAL	PRESENT	SIMPLE	ACTIVE	SMALL	13	1.25	110.21	0.00	0.006
7	acad	SINCE	INITIAL	PAST	SIMPLE	PASSIVE	SAME	8	0.94	53.03	0.05	0.004
8	acad	OTHER	MIDDLE	NO	NO	NO	SMALL	4	0.00	Inf	0.00	0.002
9	acad	ALTHOUGH	FINAL	NO	NO	NO	SAME	3	0.00	3744.00	0.00	0.002
10	acad	WHILE	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
11	acad	ALTHOUGH	FINAL	NO	NO	NO	SMALL	2	0.00	6662.67	0.00	0.001
1	**news**	**BECAUSE**	**FINAL**	**OTHER**	**SIMPLE**	**ACTIVE**	**SAME**	**22**	**6.81**	**33.92**	**0.02**	**0.008**
2	news	OTHER	MIDDLE	NO	NO	NO	SMALL	9	0.00	Inf	0.00	0.005
3	news	OTHER	FINAL	NO	NO	NO	SAME	8	0.00	39984.00	0.00	0.004
4	news	ALTHOUGH	FINAL	NO	NO	NO	SMALL	4	0.00	31992.00	0.00	0.002
5	news	ALTHOUGH	FINAL	NO	NO	NO	SAME	4	0.00	7992.00	0.00	0.002
6	news	ALTHOUGH	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
7	news	ALTHOUGH	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
8	news	WHILE	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
9	news	OTHER	FINAL	NO	NO	NO	SMALL	2	0.00	9996.00	0.00	0.001
10	news	ALTHOUGH	FINAL	NO	NO	NO	LARGE	2	0.00	7996.00	0.00	0.001

6 detected TYPES with exp. Frequency > 5

	genre	subordinator	rel. position	tense	aspect	voice	rel. length	Freq	Exp	Cont.chisq	P.adj.bin	Q
1	**acad**	**WHILE**	**INITIAL**	**PRESENT**	**SIMPLE**	**ACTIVE**	**SAME**	**46**	**12.41**	**90.96**	**0.00**	**0.018**
2	**acad**	**BECAUSE**	**FINAL**	**PRESENT**	**SIMPLE**	**ACTIVE**	**LARGE**	**33**	**12.97**	**30.91**	**0.02**	**0.011**
3	**acad**	**SINCE**	**INITIAL**	**PRESENT**	**SIMPLE**	**ACTIVE**	**SAME**	**27**	**5.25**	**90.00**	**0.00**	**0.011**
4	**acad**	**WHEREAS**	**FINAL**	**PAST**	**SIMPLE**	**ACTIVE**	**SAME**	**26**	**6.38**	**60.34**	**0.00**	**0.01**
5	**acad**	**WHILE**	**FINAL**	**PAST**	**SIMPLE**	**PASSIVE**	**SAME**	**25**	**7.92**	**36.88**	**0.01**	**0.009**
6	acad	SINCE	INITIAL	PRESENT	SIMPLE	ACTIVE	SMALL	13	1.25	110.21	0.00	0.006
11	acad	ALTHOUGH	FINAL	NO	NO	NO	SMALL	2	0.00	6662.67	0.00	0.001
1	**news**	**BECAUSE**	**FINAL**	**OTHER**	**SIMPLE**	**ACTIVE**	**SAME**	**22**	**6.81**	**33.92**	**0.02**	**0.008**
2	news	OTHER	MIDDLE	NO	NO	NO	SMALL	9	0.00	Inf	0.00	0.005
3	news	OTHER	FINAL	NO	NO	NO	SAME	8	0.00	39984.00	0.00	0.004
4	news	ALTHOUGH	FINAL	NO	NO	NO	SMALL	4	0.00	31992.00	0.00	0.002
5	news	ALTHOUGH	FINAL	NO	NO	NO	SAME	4	0.00	7992.00	0.00	0.002
6	news	ALTHOUGH	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
7	news	ALTHOUGH	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
8	news	WHILE	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
9	news	OTHER	FINAL	NO	NO	NO	SMALL	2	0.00	9996.00	0.00	0.001
10	news	ALTHOUGH	FINAL	NO	NO	NO	LARGE	2	0.00	7996.00	0.00	0.001

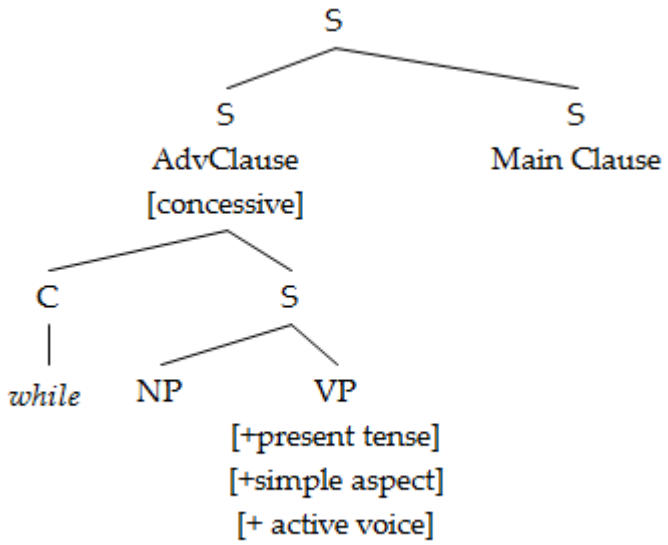
Academic writing is more formulaic

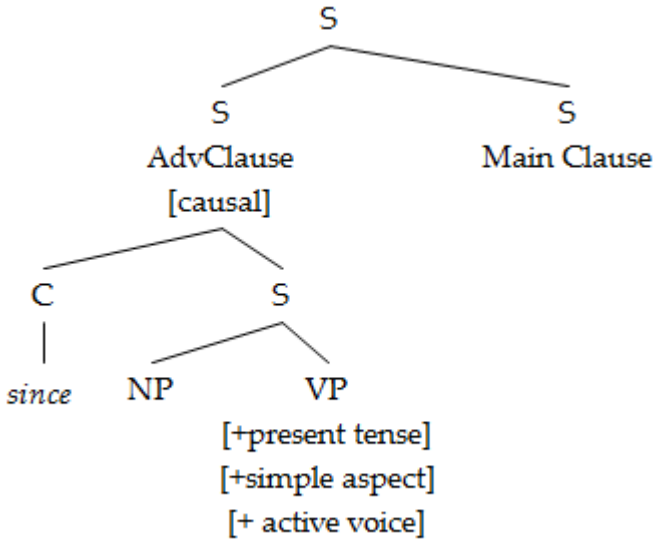
(157/1046=) 15% of the data are instances of the top 5 configurations

TYPES

	genre	subordinator	rel. position	tense	aspect	voice	rel. length	Freq	Exp	Cont.chisq	P.adj.bin	Q
1	**acad**	**WHILE**	**INITIAL**	**PRESENT**	**SIMPLE**	**ACTIVE**	**SAME**	**46**	**12.41**	**90.96**	**0.00**	**0.018**
2	acad	BECAUSE	FINAL	PRESENT	SIMPLE	ACTIVE	LARGE	33	12.97	30.91	0.02	0.011
3	**acad**	**SINCE**	**INITIAL**	**PRESENT**	**SIMPLE**	**ACTIVE**	**SAME**	**27**	**5.25**	**90.00**	**0.00**	**0.011**
4	acad	WHEREAS	FINAL	PAST	SIMPLE	ACTIVE	SAME	26	6.38	60.34	0.00	0.01
5	acad	WHILE	FINAL	PAST	SIMPLE	PASSIVE	SAME	25	7.92	36.88	0.01	0.009
6	acad	SINCE	INITIAL	PRESENT	SIMPLE	ACTIVE	SMALL	13	1.25	110.21	0.00	0.006
7	acad	SINCE	INITIAL	PAST	SIMPLE	PASSIVE	SAME	8	0.94	53.03	0.05	0.004
8	acad	OTHER	MIDDLE	NO	NO	NO	SMALL	4	0.00	Inf	0.00	0.002
9	acad	ALTHOUGH	FINAL	NO	NO	NO	SAME	3	0.00	3744.00	0.00	0.002
10	acad	WHILE	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
11	acad	ALTHOUGH	FINAL	NO	NO	NO	SMALL	2	0.00	6662.67	0.00	0.001
1	news	BECAUSE	FINAL	OTHER	SIMPLE	ACTIVE	SAME	22	6.81	33.92	0.02	0.008
2	news	OTHER	MIDDLE	NO	NO	NO	SMALL	9	0.00	Inf	0.00	0.005
3	news	OTHER	FINAL	NO	NO	NO	SAME	8	0.00	39984.00	0.00	0.004
4	news	ALTHOUGH	FINAL	NO	NO	NO	SMALL	4	0.00	31992.00	0.00	0.002
5	news	ALTHOUGH	FINAL	NO	NO	NO	SAME	4	0.00	7992.00	0.00	0.002
6	news	ALTHOUGH	MIDDLE	NO	NO	NO	SMALL	2	0.00	Inf	0.00	0.001
7	news	ALTHOUGH	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
8	news	WHILE	MIDDLE	NO	NO	NO	SAME	2	0.00	39996.00	0.00	0.001
9	news	OTHER	FINAL	NO	NO	NO	SMALL	2	0.00	9996.00	0.00	0.001
10	news	ALTHOUGH	FINAL	NO	NO	NO	LARGE	2	0.00	7996.00	0.00	0.001

Causal & concessive constructions characteristic of academic writing

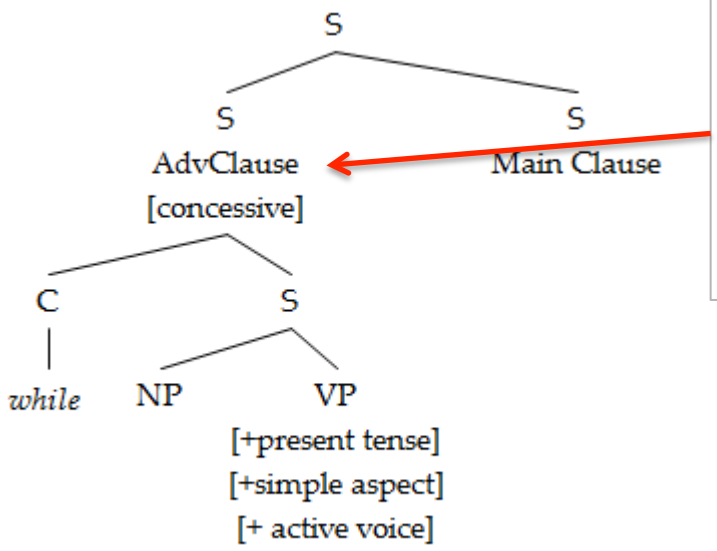




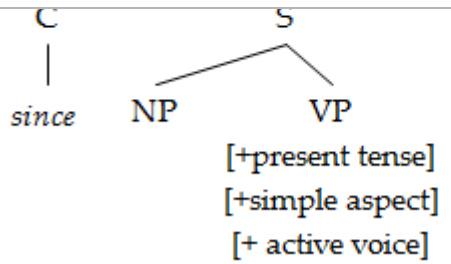
While this increases the time in the preprocess phase, it dramatically reduces the time it takes the end user to curate the data.

Since this is computationally expensive, a combination of both approaches would be better.

Causal & concessive constructions characteristic of academic writing



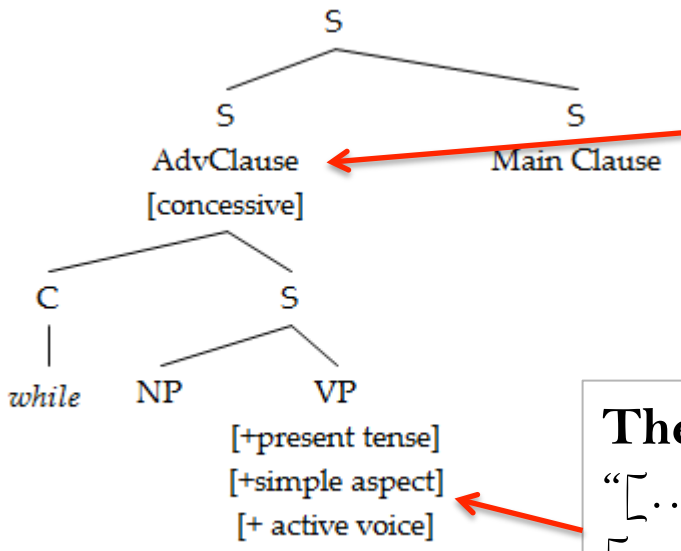
Discourse-pragmatic functions:
Bridging" function (Biber et al. 1999: 935)
"Setting-the-stage" (Thompson 1985; Ramsay 1987; Givón 1990; Ford 1993; Verstraete 2004)



While this increases the time in the preprocess phase, it dramatically reduces the time it takes the end user to curate the data.

Since this is computationally expensive, a combination of both approaches would be better.

Causal & concessive constructions characteristic of academic writing



Discourse-pragmatic functions:

Bridging" function (Biber et al. 1999: 935)

"Setting-the-stage" (Thompson 1985; Ramsay 1987; Givón 1990; Ford 1993; Verstraete 2004)

The use of present tense in academic writing:

"[…] using the present simple tense means that […] findings and deductions are strong enough to be considered as facts or truth." (Glasman-Deal 2010: 5)

While this increases the time in th *phase*, it dramatically reduces the time it takes the end user to curate the data.

combination of both approaches would be better.

Conclusion

- **Goal**:
  - Identify genre-specific patterns of complex, schematic constructions in academic writing
- **Case study**:
  - Comparison of concessive & reason adverbial clauses in academic writing and newspaper
- **Method:**
  - Theory-guided & data-driven identification of most relevant dimensions of contrast (via tree-based models)
  - Pattern detection within the resulting feature space (via Configural Frequency Analysis)
- **Result:**
  - Academic writing exhibits genre-specific patterns (constructions) that serve functions that are particularly "useful" in argumentative texts
  - more formulaic

References

- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. (1999). *Longman grammar of spoken and written English*. London: Longman.
- Diessel, Holger (2005). Competing motivations for the ordering of main and adverbial clauses. *Linguistics 43*: 449-470.
- Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis & Torsten Hothorn. 2008. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, 8-25.
- Strobl, Caroline, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4): 323-348.
- von Eye, A. (1990). *Introduction to Configural Frequency Analysis: The search for types and antitypes in cross-classifications*. Cambridge, UK: Cambridge University Press.

Thank you very much!

kerz@anglistik.rwth-aachen.de
wiechmann@anglistik.rwth-aachen.de

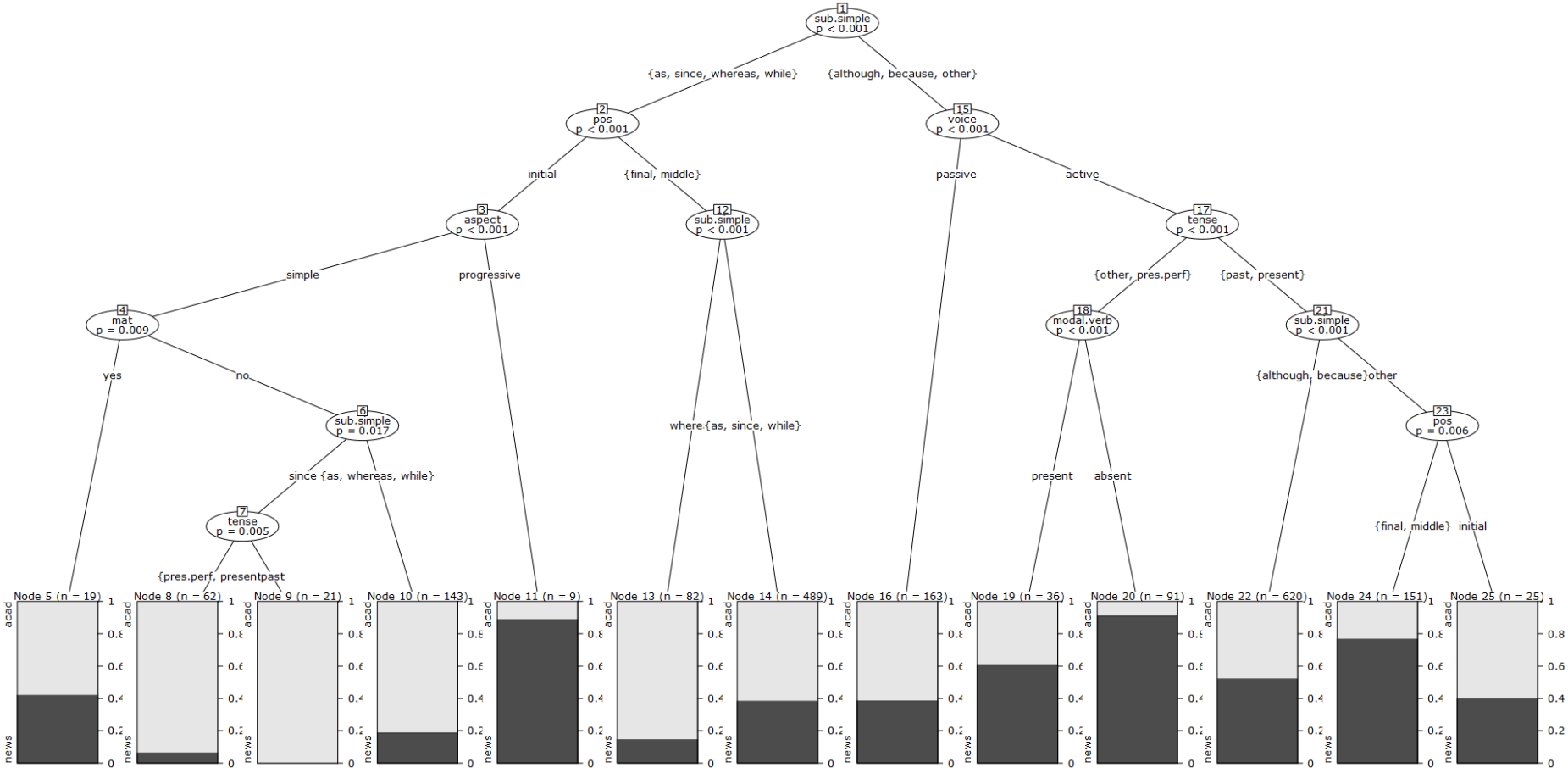
Outlook

- Implications for language processing
  - Expectation-based perspective:
    - Expectation-driven language processing account and its anticipatory character
      - Language users will have different expectations in different contexts
      - Genre as one of contextual cue: Different genres may cause language users to derive different sets of expectations towards constructional choices
    - Since the language system is 'experience-driven', the constant confrontation, particularly with domain-specific linguistic input has an enormous impact on the repository of constructions (constructicon) accessible to and used by language users of that particular language domain

Conditional Inference Tree (algorithm)

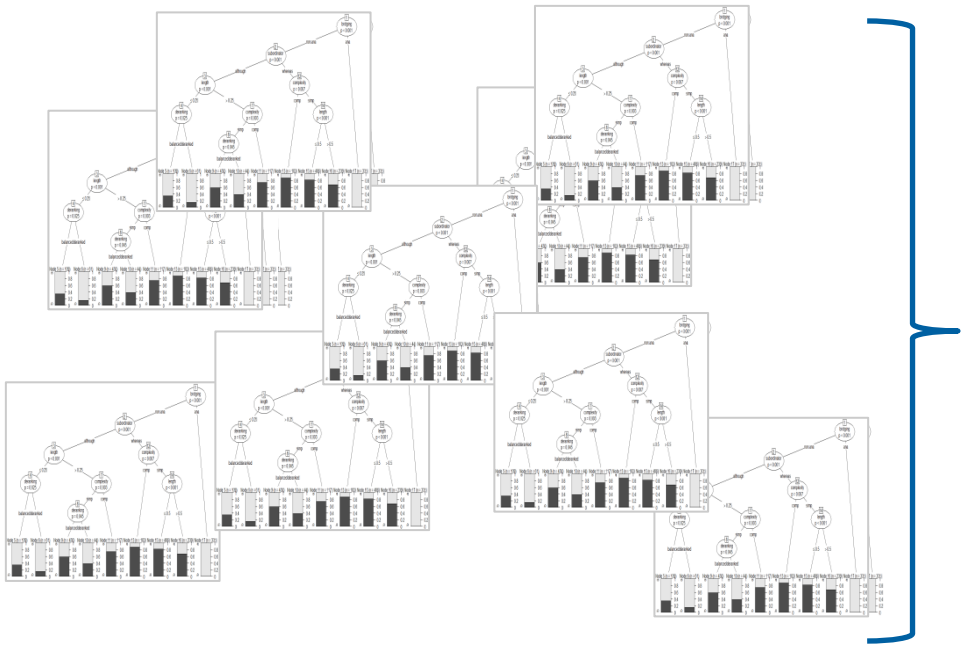
1) **Test** the global null hypothesis of **independence** between any of the **input variables** and the **response** (here: GENRE).
   - Stop if this hypothesis cannot be rejected.
   - Otherwise **select** the **input variable** with the **strongest association** to the response. This association is measured by a p-value corresponding to a test for the partial null hypothesis of a single input variable and the response.
2) Implement a binary split in the selected input variable.
3) Recursively repeat steps (1) and (2).

Conditional Inference Tree: Output



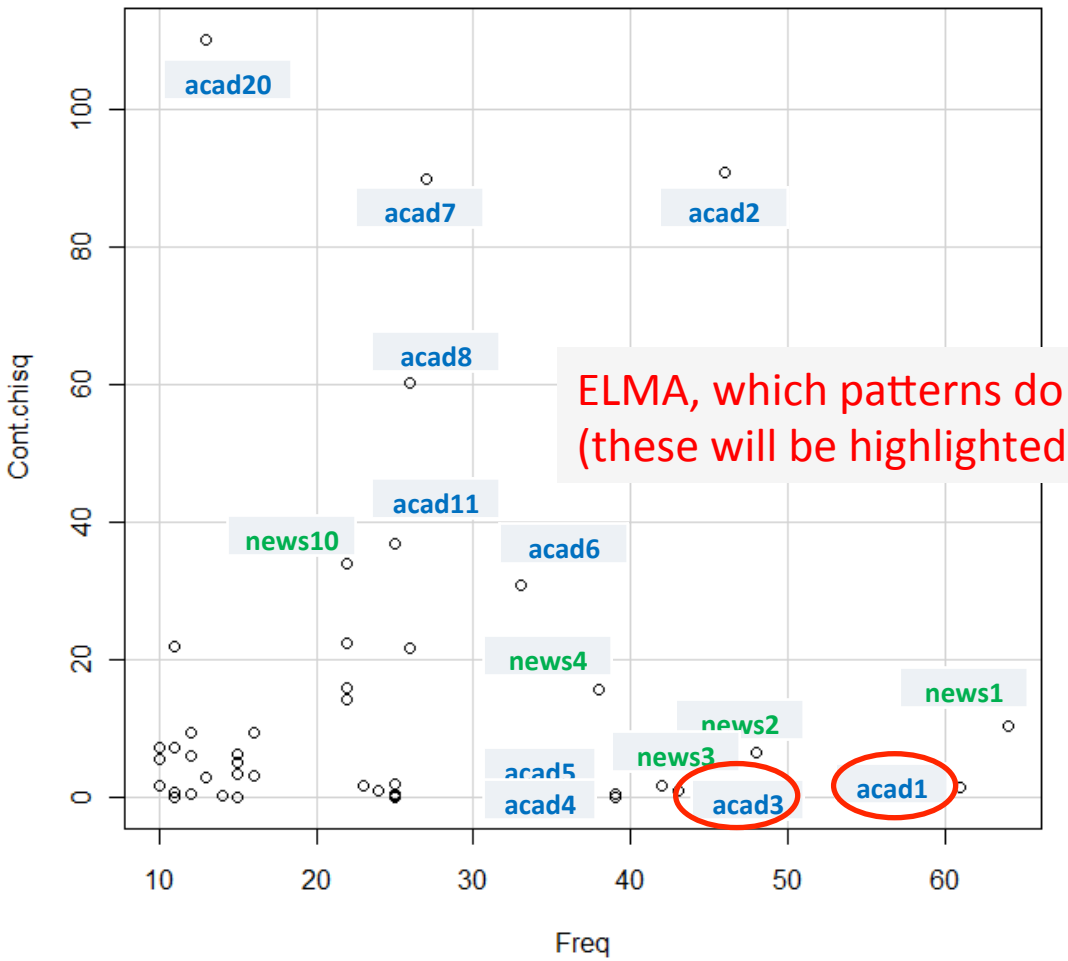
Why Random Forests?

- a single tree-based model can be vulnerable to small changes in the data set
- better grow a large number of trees (based on random samples of data)
- deciding a final predicted outcome by combining the results across all of the trees



Results are determined via **majority vote**

Frequency vs. Contribution to χ^2



ELMA, which patterns do you want to discuss?
(these will be highlighted here)

Detected TYPES: **Discussion**

E.g. TYPE for concessive clauses:

GENRE	SUB	POS	TENSE	ASPECT	VOICE	LENGTH
academic	**while**	**initial**	**present**	**simple**	**active**	**same**

- **Some instantiations of the pattern:**
  3. *While this increases the time in the preprocess phase*, it dramatically reduces the time it takes the end user to curate the data.

E.g. TYPE for casual clauses:

GENRE	SUB	POS	TENSE	ASPECT	VOICE	LENGTH
academic	**since**	**initial**	**present**	**simple**	**active**	**same**

- **Some instantiations of the pattern:**
  5. *Since this is computationally expensive*, a combination of both approaches would be better.

The use of present tense in academic writing:

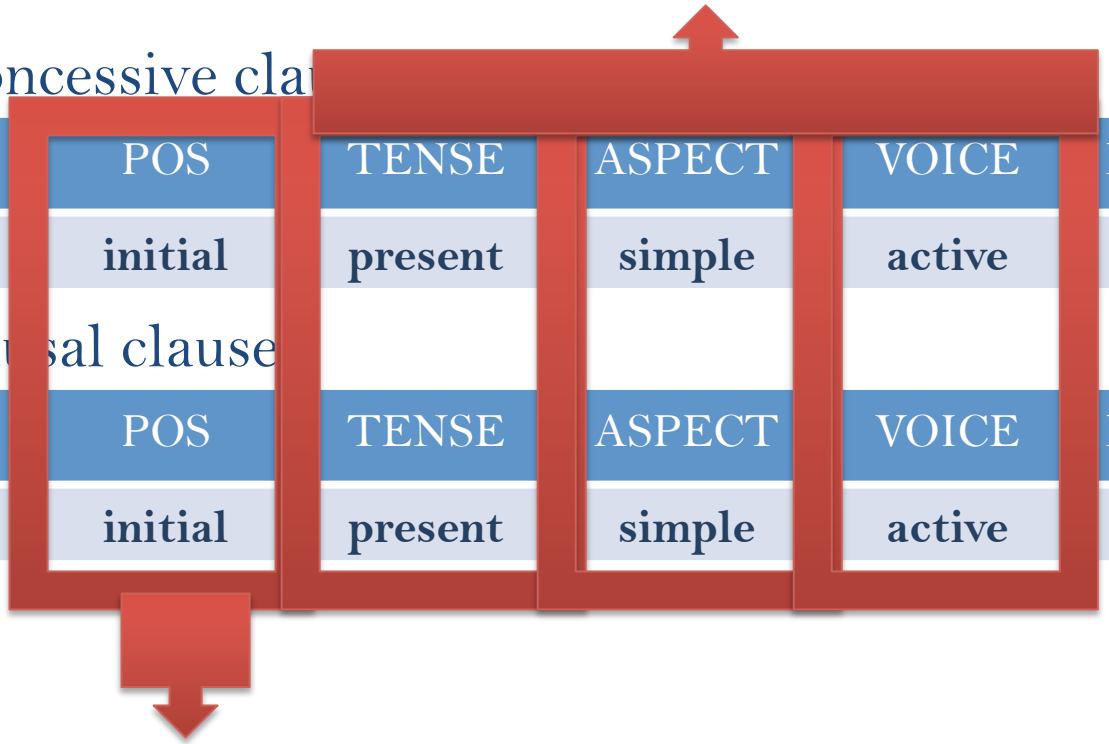
"[…] using the Present Simple tense means that […] findings and deductions are strong enough to be considered as facts or truth."

E.g. TYPE for concessive cla

GENRE	SUB	POS	TENSE	ASPECT	VOICE	LENGTH
academic	**while**	**initial**	**present**	**simple**	**active**	**same**

E.g. TYPE for ca sal clause

GENRE	SUB	POS	TENSE	ASPECT	VOICE	LENGTH
academic	**since**	**initial**	**present**	**simple**	**active**	**same**



Two discourse-pragmatic functions:

Bridging" function

(Biber et al. 1999: 935)

"Setting-the-stage"

(cf. Thompson 1985; Ramsay 1987; Givón 1990; Ford 1993; Verstraete 2004)