universität innsbruck

What corpora can (and can‘t) tell us about textual and lexical meaning?

Aleksandar Trklja

- Textual Meaning via Corpus Analysis: A Case Study
- Lexical Meaning via Corpus Analysis: A Case Study
- Function, Usage, Meaning

Textual (=discourse) meaning

- Language and Law at the ECJ
- EU case law corpus

The language EU case law

- EU case law is the body of judicial decisions delivered by the Court of Justice of the European Union (CJEU) and the General Court, which interprets and applies EU law and ensures its uniform implementation across all Member States.

Official languages

- The 24 official languages make a total of 552 possible combinations

Repetition degree: CJEU vs non-CJEU judgments

- Background:
  - repetitiveness often mentioned as one of key features of legal texts;
  - previous studies indicate that CJEU judgments are especially repetitive (McAuliffe, 2007)
- Objective:

  i) examine the claim empirically

  ii) investigate whether CJEU judgments are more formulaic than non-CJEU judgments

Data

- ECJ: A multilingual corpus: 1141 ECJ *acquis communautaire* judgments in English, French, German and Italian (1953-2011)

- Non-ECJ: Comparable corpus: around 1200 judgments of Austrian, Belgian, French, German, Italian, Irish and UK national courts (1953-2011)

Discovery procedure

1. Define the units of analysis

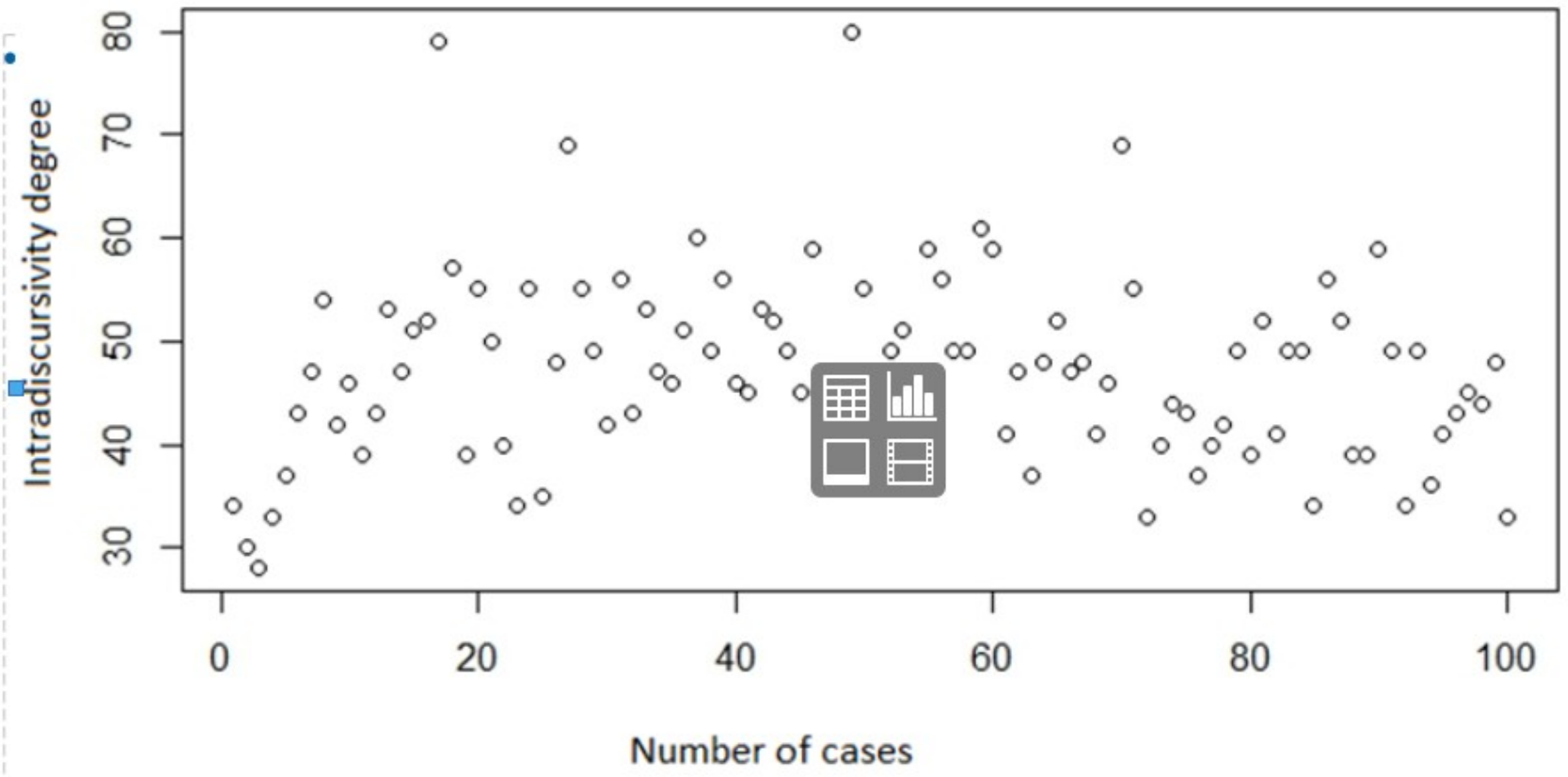
2. Identify repetitive expressions across judgments for CJEU and non-CJEU judgments

3. Compare the degree of repetition across countries and legal systems

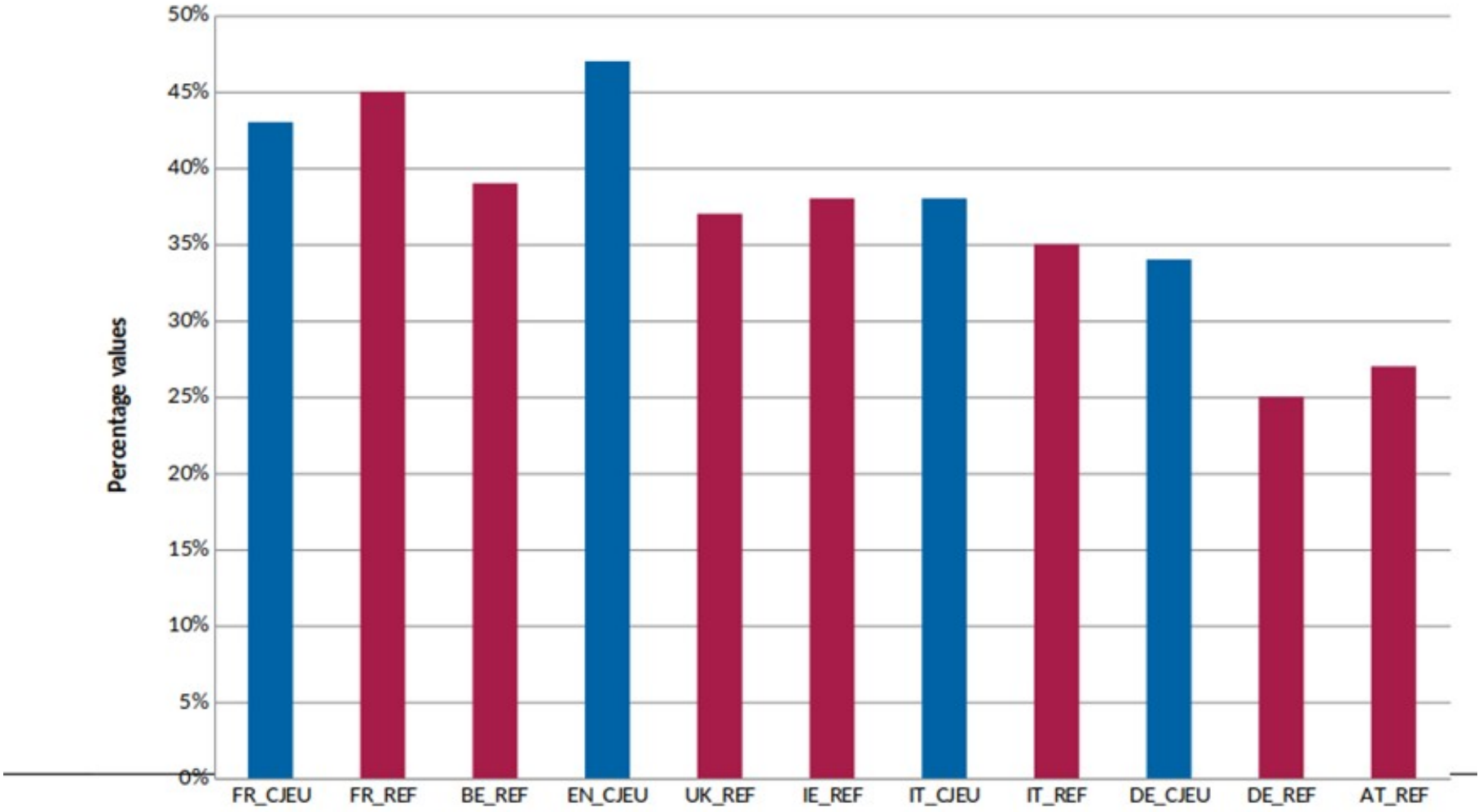
Moreover, as the Court has repeatedly stated, whilst the protection of legitimate expectations i
one of the fundamental principles of the Community, traders cannot have a legitimate expectation th
t an existing situation which is capable of being altered by the Community institutions in the exercise
of their discretion will be maintained;

this is particularly true in an area such as the common organisation of the markets whose purpos
involves constant adjustments to mee
changes in the economic situation (see, in particular, Case C-372/96 Pontillo
ECR I-5091, paragraphs and That necessarily applies with greater force where the hopes purported
entertained by the traders were raised by a publicly distributed leaflet having no legal status...

Degree of repetition in ECJ judgments (English)



Degree of repetition: ECJ/non-ECJ



The requirements to be satisfied by the statement of reasons depend on the circumstances of each case.

- 61982CJ0296
- 61983CJ0041
- 61983CJ0172
- 61985CJ0185
- 61988CJ0303
- 61995CJ0367
- 61996CJ0048
- 61996CJ0301
- 61997CJ0075
- 61997CJ0265
- 61997CJ0372
- 61998CJ0015
- 61998CJ0279
- 61999CJ0017
- 61999CJ0120
- 61999CJ0163
- 61999CJ0280
- 61999CJ0310
- 62000CJ0041
- 62000CJ0057
- 62000CJ0076
- 62000CJ0113
- 62000CJ0114
- 62000CJ0445
- 62001CJ0042
- 62001CJ0076
- 62002CJ0066
- 62003CJ0138
- 62005CJ0266
- 62006CJ0390
- 62007CJ0333
- 62008CJ0089
- 62008CJ0279
- 62008CJ0280
- 62009CJ0194
- 62009CJ0335
- 62009CJ0548
- 62010CJ0014
- 62010CJ0015
- 62010CJ0403
- 62010CJ0539
- 62011CJ0201
- 62011CJ0417
- 62011CJ0439
- 62011CJ0444
- 62011CJ0455
- 62011CJ0629
- 62013CJ0037
- 62013CJ0176
- 62013CJ0200
- 62013CJ0286
- 62013CJ0687

- Formulaicity is one of the defining features of legal judgments
- ECJ judgments tend to be more formulaic than non-ECJ judgments.
- The degree of repetition depends on legal and language systems.

Textual colligation

- the property of expressions “to occur (or to avoid occurring) at the beginning or end of independently recognised discourse units, e.g. the sentence, the paragraph, the speech turn” (Hoey 2005: 115).

PIMFE- Paragraph-initial formulaic expressions

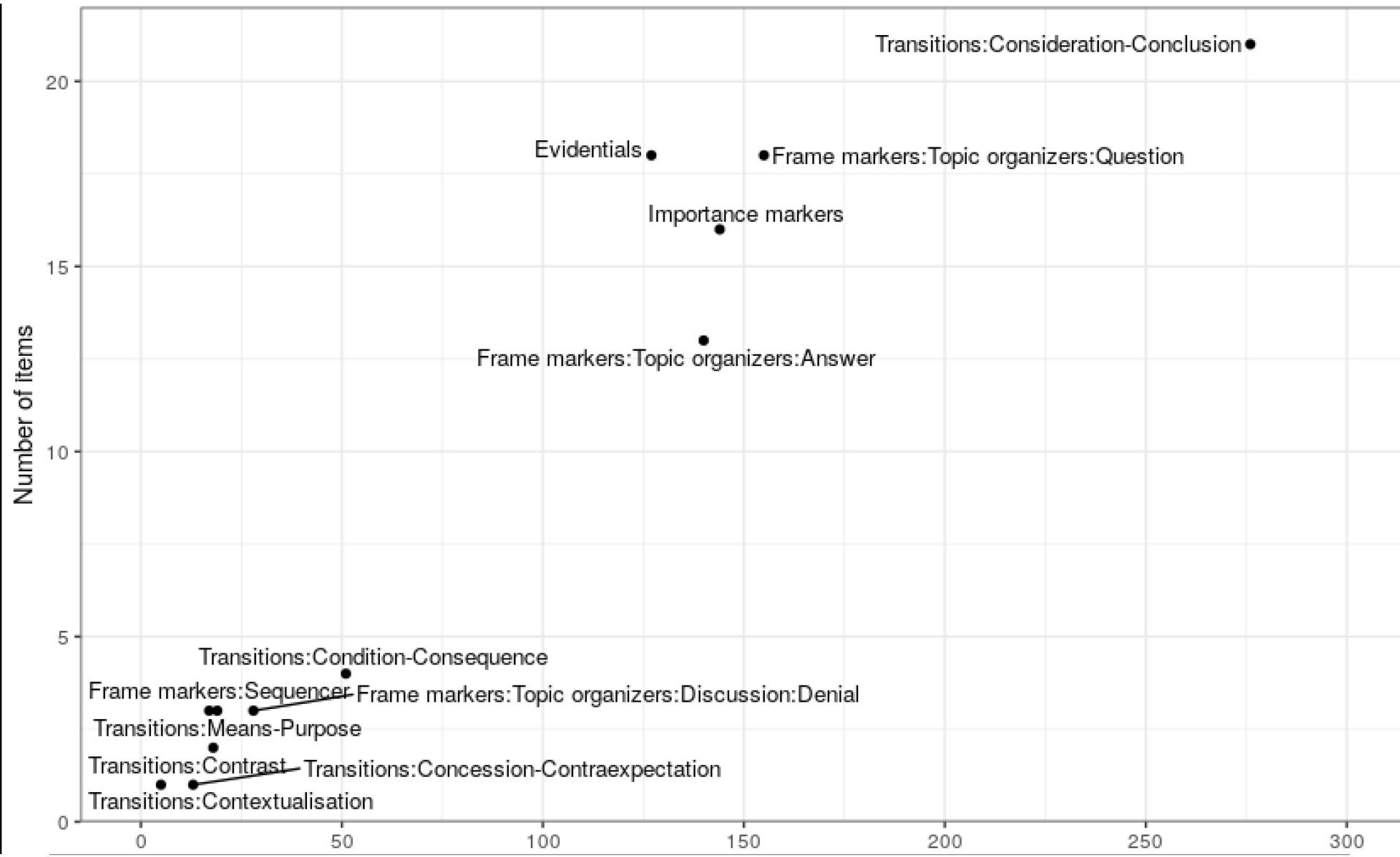
Frequency	PIMFE
58	it follows from the foregoing
47	it is clear from the
29	it is apparent from the
26	the answer to the first
26	if the answer to question
23	the plaintiff in the main
22	it must therefore be concluded
21	as the court has already
20	it is therefore necessary to
21	it should be noted that

Frequency	PIMFE
18	the answer to the question
17	it is common ground that
17	in the light of the
16	the first paragraph of article
16	the answer to the second
15	in that respect it must
15	it should be pointed out
14	the first question asks whether
14	it appears from the file
13	in the view of the

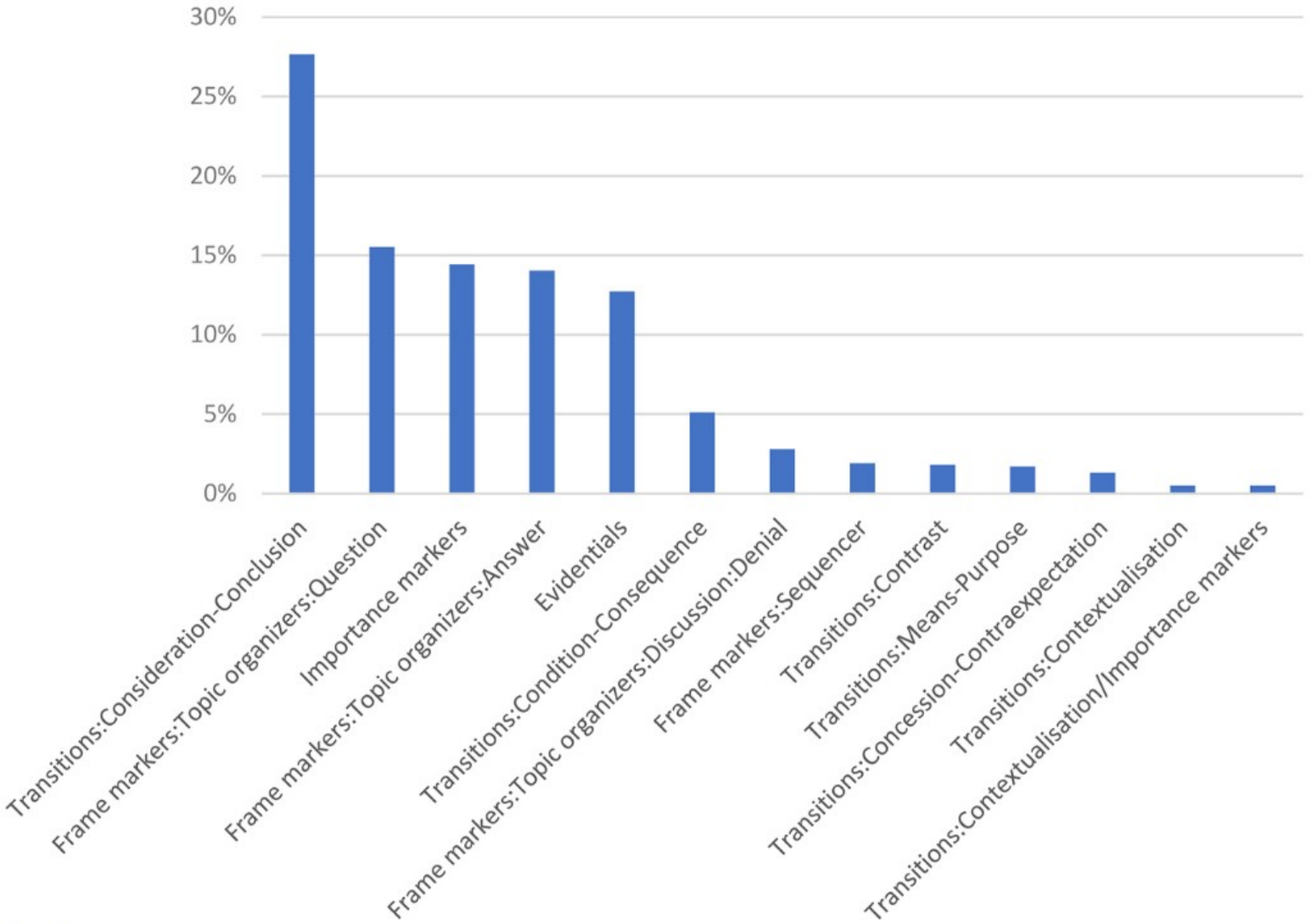
- Language is 'multifunctional' (Halliday & Hasan, 1989, p. 23), and linguistic expressions have three specific functions or meanings: **ideational,** (**representational)**, **interactive (interpersonal)**, and **organisational (discourse-level)**.

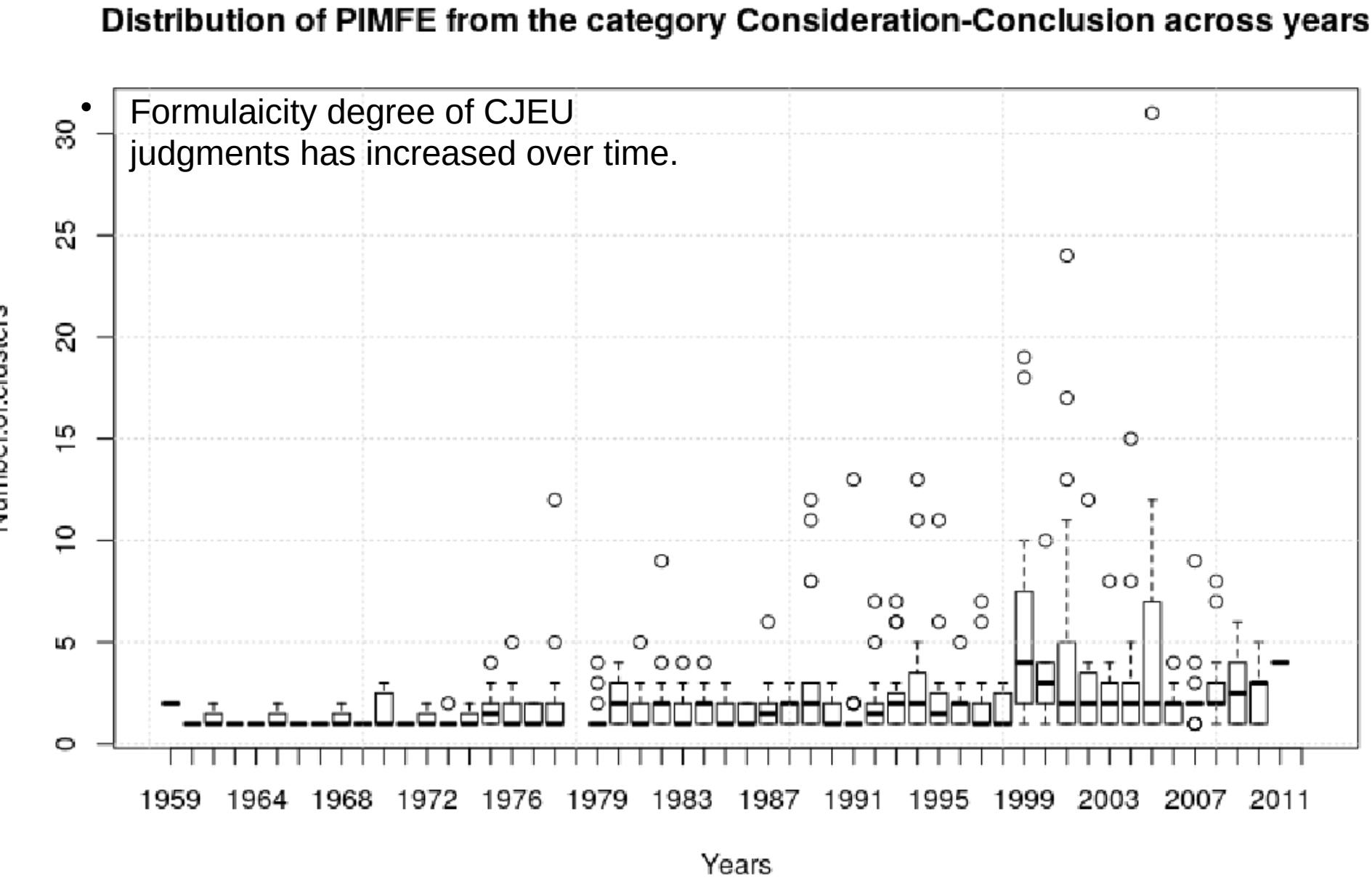
Formulaic metadiscursive devices

- **recurrent, conventionalized expressions** that writers or speakers use to **organise discourse** and **guide the reader's or listener's interpretation** of a text.
  - They are *formulaic* because they occur in **fixed or semi-fixed patterns**, and *metadiscursive* because they **comment on the discourse itself** rather than adding new propositional content.



universität innsbruck


30%
25%
20%
15%
10%
5%
0%
Transitions:Consideration-Conclusion
Frame markers:Topic organizers:Question
Importance markers
Frame markers:Topic organizers:Answer
Evidentials
Transitions:Condition-Consequence
Frame markers:Topic organizers:Discussion:Denial
Frame markers:Sequencer
Transitions:Contrast
Transitions:Means-Purpose
Transitions:Concession-Contraexpectation
Transitions:Contextualisation
Transitions:Contextualisation/Importance markers


Distribution of PIMFE from the category Consideration-Conclusion across years
Formulaicity degree of CJEU judgments has increased over time.
Number of clusters
0
5
10
15
20
25
30
1959
1964
1968
1972
1976
1979
1983
1987
1991
1995
1999
2003
2007
2011
Years

Textual Meaning

- Metadiscursive formulaic expressions can be grouped into semantic classes according to their function
- Those that signal the Consideration-Conclusion pattern tend to be the most frequent in CJEU judgments.

Distributional properties in lexical domains

- Extended distributional hypothesis

The Distributional Hypothesis

- Zellig Harris (1952–1970): meaning correlates with distribution.
  - Words in similar contexts → similar meanings.
  - Equivalence via substitutability.
- Foundation for distributional semantics

Extended distributional hypothesis

- *No two items from one language will correspond to the same item from another language and simultaneously occur in the same context unless they have the same meaning.*
- Lexical items are generated from the corpus

	<give rise to problem>	<give rise to concern>	<give rise to fear>	<give rise to debate>	<give rise to confusion>	<give rise to difficulty>	<give rise to doubt>	<give rise to question>	<give rise to cost>
<zu Problem|Schwierigkeit|Sorge|Verwirrung|Kosten führen>	√	√			√	√			√
<Problem|Schwierigkeit auftreten>	√					√			
<Problem|Sorge|Debatte|Verwirrung auslösen>	√	√		√	√				
<Problem|Sorge|Debatte|Zweifel|Frage hervorrufen>	√	√		√			√	√	
<zu Sorge|Angst|Debatte|Zweifel|Frage Anlass geben>		√	√	√			√	√	
<Problem|Verwirrung|Schwierigkeit|Frage|Kosten entstehen>	√				√	√		√	√
<Problem|Schwierigkeit (mit sich) bringen>	√					√			
<Problem es gibt>	√								
<zu Debatte|Verwirrung kommen>				√	√				
<mit Problem|Schwierigkeite verbunden sein>	√					√			

Lexical domains {CAUSE PROBLEM} and {PROBLEME BEREITEN}

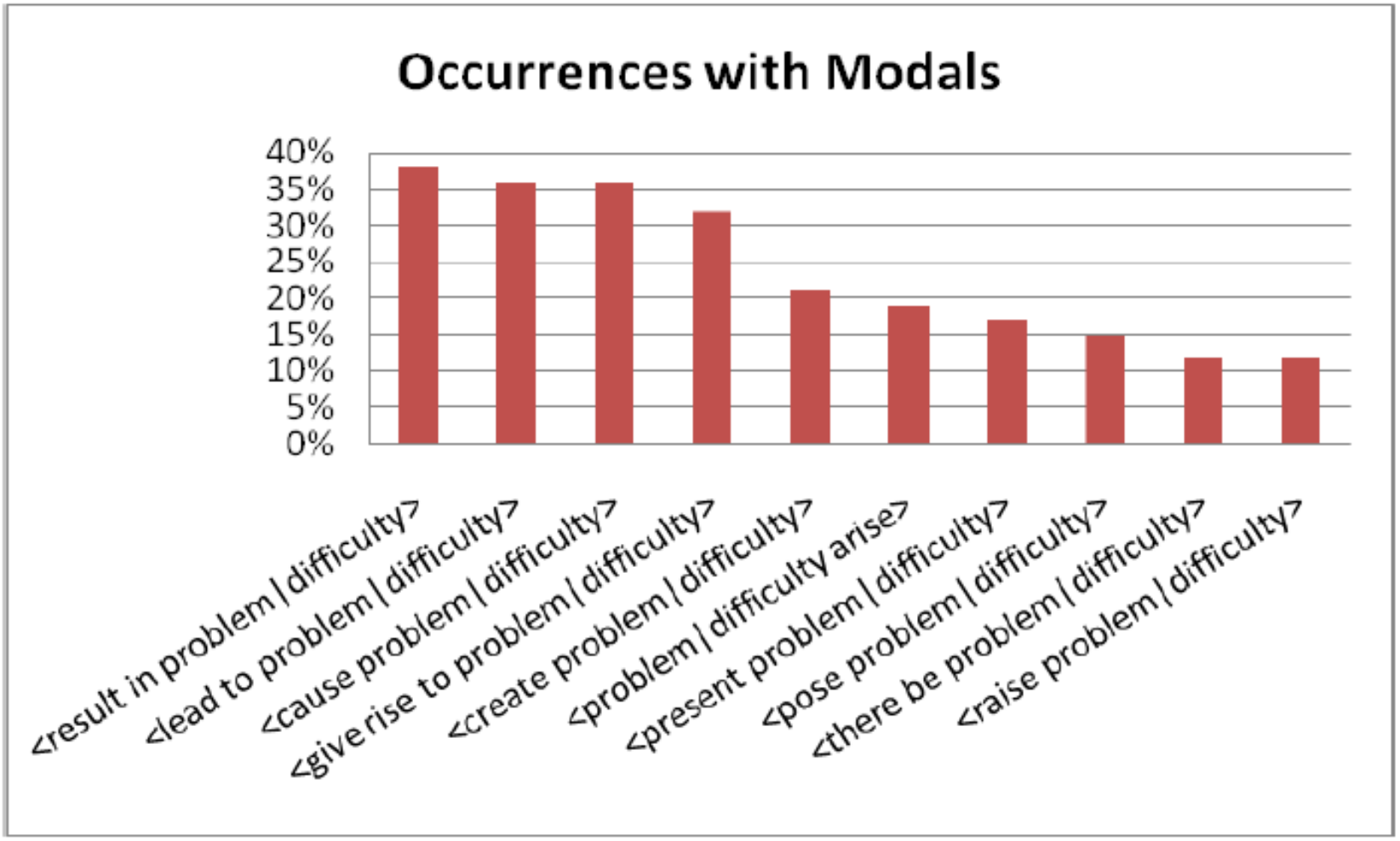
English lexical items	German lexical items
<cause problem>	<es geben Problem|Schwierigkeit>
<create problem>	<Problem|Schwierigkeit auftreten>
<give rise to problem>	<Problem|Schwierigkeit aufwerfen
<lead to problem>	<Problem|Schwierigkeit bereiten>
<pose problem>	<Problem|Schwierigkeit bringen>
<present problem>	<Problem|Schwierigkeit darstellen>
<problem arise>	<Problem|Schwierigkeit entstehen>
<raise problem>	<Problem|Schwierigkeit schaffen>
<result in problem>	<Problem|Schwierigkeit sich ergeben>
<there be problem>	<Problem|Schwierigkeit verursachen>
	<problematisch sein>
	<Ursache GEN|für Problem|Schwierigkeit sein>
	<zu Problem|Schwierigkeit führen>

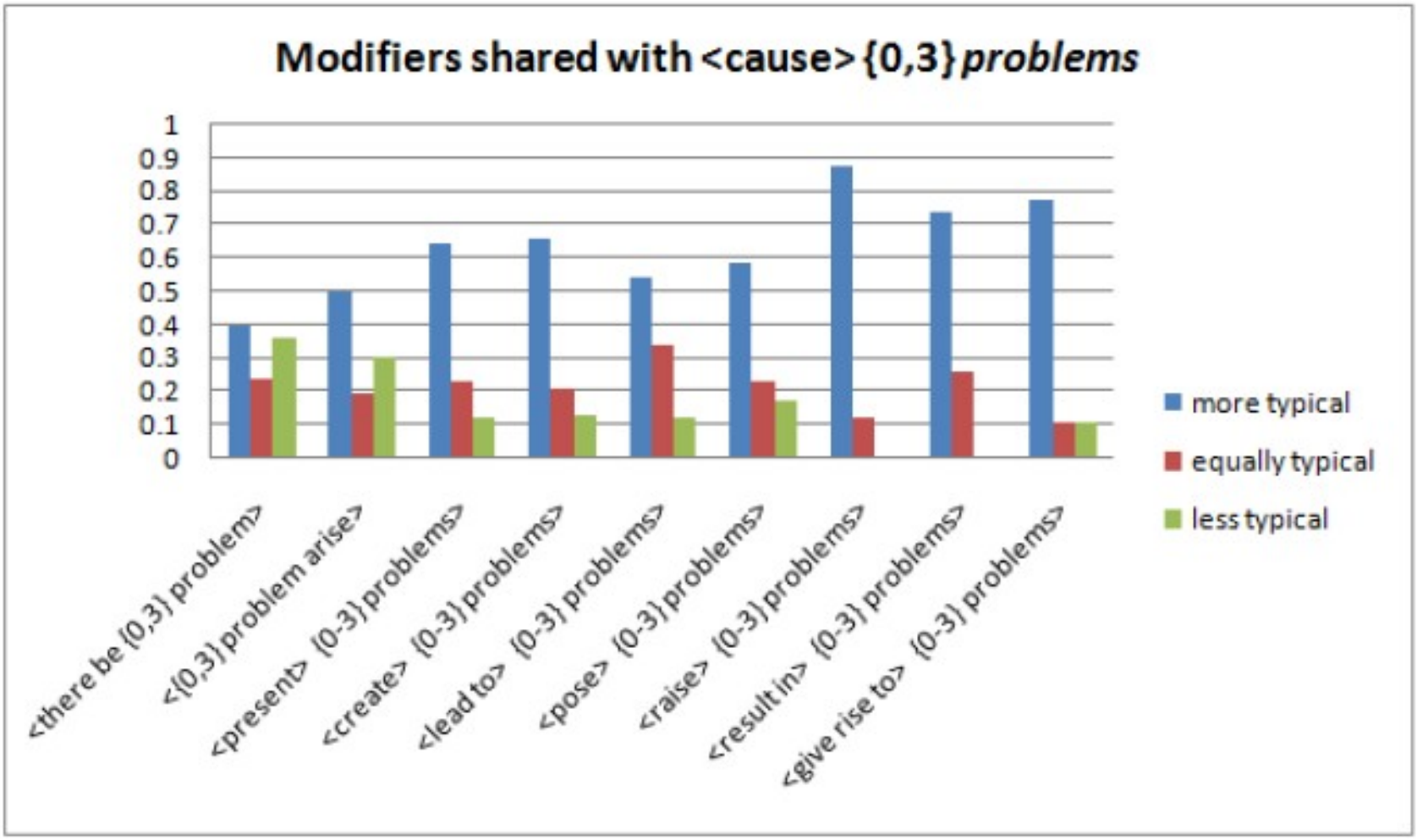
Modifiers that occur with <problem> and <difficulty>

INTENSIFIERS	QUANTIFIERS	SORTALS	COMPARATORS
big	a few	access	additional
considerable	a great range of	behaviour	another
enormous	a lot of	communication	certain
great	a number of	engineering	different
huge	a series of	environmental	distinct
key	a small number of	ethical	further
large	all kind of	financial	new
major	all sort of	health	other
minor	fewer	legal	particular
serious	many	logistical	same
severe	more	management	similar
significant	numerous	noise	special
small	several	operational	typical
substantial	some	performance	unique
subtle		political	various
		pollution	
		practical	
		safety	
		security	

Individual differences in categories {CAUSE PROBLEM}

Lexical items	Transitivity	Passive	RECIPIENT	Patterns
<cause problem|difficulty>	TR	√	DO/<for>+NP	CAUSE
<create problem|difficulty>	TR	√	<for>+NP	CREATE
<give rise to problem|difficulty>	TR			GIVE RISE TO
<lead to problem|difficulty>	TR			GIVE RISE TO
<pose problem|difficulty>	TR	√	<for>+NP	CREATE
<present problem|difficulty>	TR		<for>+NP	PRESENT
<problem|difficulty arise>	INTR			ARISE
<to be problematic>				PROBLEMATIC
<raise *problem*|difficulty>	TR		<for>+NP	PRESENT
<result in problem|difficulty>	TR			GIVE RISE TO
<there be problem|difficulty>	EX			THERE


Occurrences with Modals
40%
35%
30%
25%
20%
15%
10%
5%
0%
<result in problem|difficulty>
<lead to problem|difficulty>
<cause problem|difficulty>
<give rise to problem|difficulty>
<create problem|difficulty>
<problem|difficulty arise>
<present problem|difficulty>
<pose problem|difficulty>
<there be problem|difficulty>
<raise problem|difficulty>


Modifiers shared with <cause> {0,3} problems
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
<there be {0,3} problem>
<{0,3} problem arise>
<present> {0-3} problems>
<create> {0-3} problems>
<lead to> {0-3} problems>
<pose> {0-3} problems>
<raise> {0-3} problems>
<result in> {0-3} problems>
<give rise to> {0-3} problems>
more typical
equally typical
less typical

The main points

- Lexical domains can be identified on the base of **distributional** properties
  - Lexical items that occur in the same lexical domains are associated with the same local grammar template
- Fine-grained semantic categories can be generated from the corpus through the observation of **distributional** properties of lexical items
  - Individual differences can be observed both at the level of selectional properties and statistical tendencies

Functions, use and meaning

Assumptions about exploring meaning

- Aim: to describe patterns of language in context.
- Text linguistics and corpus-driven approaches study meaning through the description of patterns.
- Foundational question: Does describing use equal explaining meaning?

Chomsky's three levels of adequacy applied

- Observational Adequacy – A theory achieves observational adequacy if it accurately fits the observable linguistic facts — that is, if it can identify, classify, and record the data of language use.
- Descriptive Adequacy –A theory achieves descriptive adequacy if it represents the internalized knowledge speakers have — how linguistic structures are mentally organised and related.
- Explanatory Adequacy – accounts for the mental mechanisms and representational capacities that make language possible.?

Textual meaning

- Analysis relies on categorisation of communicative roles.
- Based on human capacity to generalise and categorise.

Textual meaning

- ‘there are no simple linguistic criteria for identifying metadiscourse’ because metadiscourse categories are open and new items can be added or removed depending on data. (Hyland and Tse, 2004: 158)

- Transitions help readers to make ‘connections between preceding and subsequent propositional information’ (Cao and Hu, 2014).
- Logico-deductive relations:
  - 1) Reason-Result
  - 2) Consideration-Conclusion
  - 3) Condition-Consequence
- Associative semantic relations
  - 1) Contrast
  - 2) Statement-Denial
- Tempero-contigual semantic relations
  - 1) Chronological relations

Textual meaning

- Observational: Fully met — empirical text-based descriptions.
- Descriptive: attempts to reach this level by describing *how* sentences and texts are organized in terms of *functions*
  - However, this organization is still **taxonomic** rather than explanatory: it reflects human categorization and the organization of linguistic forms, but it does not specify the **cognitive mechanisms** that generate or interpret meaning.
- Explanatory: fail — describe usage and classification but not the underlying cognitive architecture that enables meaning.

- Categorising data ≠ explaining meaning.

Formal semantics approaches

- Discourse Representation Theory (Kamp & Reyle, 1993) models textual coherence and anaphora resolution.
- Dynamic semantics (Heim, 1982; Groenendijk & Stokhof, 1991) explains context update and presupposition.
- Formal pragmatics and speech act theory (Stalnaker, 1978; Krifka, 2015) integrate interpersonal meaning and speaker intention.

The contextualist view of "meaning"

- Firth (1968): 'You shall know a word by the company it keeps.'
- Wittgenstein (1953): 'The meaning of a word is its use in the language.'
- Meaning = contextual use, not reference.

Corpus-based and Corpus-driven approaches

- Corpora provide empirical access to linguistic behaviour (Hanks, 2008).
  - Corpus-based: applies predefined categories.
  - Corpus-driven: lets categories emerge from data.

The Distributional Hypothesis

- Zellig Harris (1952–1970): meaning correlates with distribution.
- Words in similar contexts → similar meanings.
- Equivalence via substitutability.

Sinclair’s minimal assumption

- “[w]e should only apply loose and flexible frameworks until we see what the preliminary results are in order to accommodate the new information that will come from the text” (Sinclair, 1994: 25).

- “The stance of the observer controls and limits the observations that can be made; for human observers the stance includes their involuntary reactions to language in use, in particular whatever theoretical and descriptive presup positions remain unexamined, and possibly unrecognised, in those reactions. It is therefore essential to adopt a methodology that obliges the observer to distance himself or herself from the experience of running text, in the first instance, and instead look at the linguistic inforrnation as scientific data. Later, of course, once a description arrived at with maximum objectivity has been achieved, the intuitions and responses of the human researcher are essential for interpretation of the phenomena. “ (Sinclair, 1999:2)

- ““We should trust the text. We should be open to what it may tell us. We should not impose our ideas on it.” (Sinclair, 2004, 23)
- Meaning is **textual and discourse-based**. It arises from how words pattern and interact over extended stretches of language.
- Meaning is studied through patterns of co-selection

Pattern grammar

V in the N
V the N N
V to V N
V N
V to V the N
V in A N
V by the N
V on the N
V a N N
V with N
V a N of N
V the N of the N
V at the N
V to V A N
V by A N
V in the A N
V with the N
V with A N

Pattern grammar

II.1 The `talk' group

These verbs are concerned with speaking or writing. This includes:

- verbs that indicate the function of what is said e.g. *argue, ask, complain*
- verbs that indicate how something is said e.g. *mutter, wail*
- verbs that indicate the feeling of the speaker e.g. *enthuse, fulminate*

II.2 The `think' group

These verbs are concerned with thought or feeling, or the expression of thought or feeling. The prepositional phrase indicates the top

II.3 The `learn' group

These verbs are concerned with acquiring knowledge. The prepositional phrase indicates what the knowledge concerns.

- Corpus-driven approaches describe use, not meaning.
  - Observed data must be interpreted.
  - Interpretation requires semantic knowledge.
  - It is at this point that semantics enters linguistic analysis, rather than in a corpus-driven approach or the classification of data.

- What is meaning then?

- Meaning is representational (mental content).
- Meaning depends on internal structures and extralinguistic parameters.
  - Meaning = mental representation (narrow content) + context (broad content).
- Interpretation requires prior semantic knowledge.

Meaning ≠ function 1

- Function refers to what something is for — its purpose, role, or teleological end in a system.
  - “The heart’s function is to pump blood.”
- Function is explained by contribution to a system or by evolutionary purpose (biological, social, or communicative). Meaning, by contrast, refers to what something stands for or represents.
  - “The word dog means ‘dog’.”

Meaning ≠ function 1

- Meaning is semantic and representational — it establishes a relation between a symbol and its referent, not a purpose.
- Thus, while a thermostat has a function, it doesn’t have meaning; it doesn’t understand “temperature.”

Meaning ≠ function 2

- Millikan (1984) Language, Thought, and Other Biological Categories distinguishes proper function (biological purpose) from intentional content (what a representation means).
- Function explains how meaning can arise, but is not itself meaning.

Meaning ≠ function 3

- A sentence has meaning insofar as it expresses a proposition that can be true or false. A function or purpose, by contrast, is not truth-evaluable — it doesn’t represent a state of affairs.

Meaning

- Meaning is a representational and contextually determined property of the mind–world interface.
- Corpus evidence reflects linguistic behavior but not the intentional or truth-conditional content that constitutes meaning.
- Hence, corpus-driven and usage-based models are epistemologically useful but ontologically incomplete as theories of meaning.

Thank you for your attention!