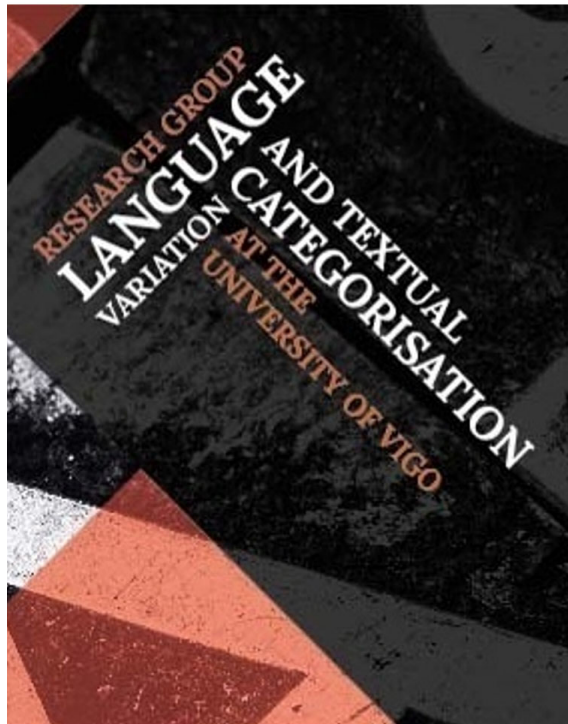


lvtc



Universidade de Vigo

# (Systemic Funcional) Theme as an indicator of register variation in English

**Javier Pérez-Guerra**

[jperez@uvigo.es](mailto:jperez@uvigo.es)



21 Apr 2022

# Vigo



LVTC

*LVTC (Language Variation and Textual Categorisation)  
research group*

UniversidadeVigo

*lvtc*

LANGUAGE VARIATION  
AND TEXTUAL CATEGORISATION

<http://lvtc.uvigo.es>

before 1990: LOB, Brown, Helsinki Corpus (1/1.5 million words)

1994: first release of BNC (100 million words)

**B** RITISH **N** ATIONAL **C** ORPUS

2008: first release of COCA (450+ million words)

**Corpus of Contemporary American English**

...

2018: enTenTen15 (21.9 billion words)



enTenTen – English corpus from the web

today: talk based on a 1-million-word corpus



today: talk based on a 1-million-word corpus

- large-scale case study
- combination of...



# Outline

- Rationale
  - SFL
  - Textual metafunction
    - Themes
    - Registers
- Goal
  - Theme taxonomy
    - First-element hypothesis
    - Preverb hypothesis
  - RQs
- Data
  - Corpus
  - Database
  - DAT analyser
- Analysis of the data
  - Cluster
  - Random Forest
  - Snake plot
- Conclusions



# Rationale

## Pillar 1

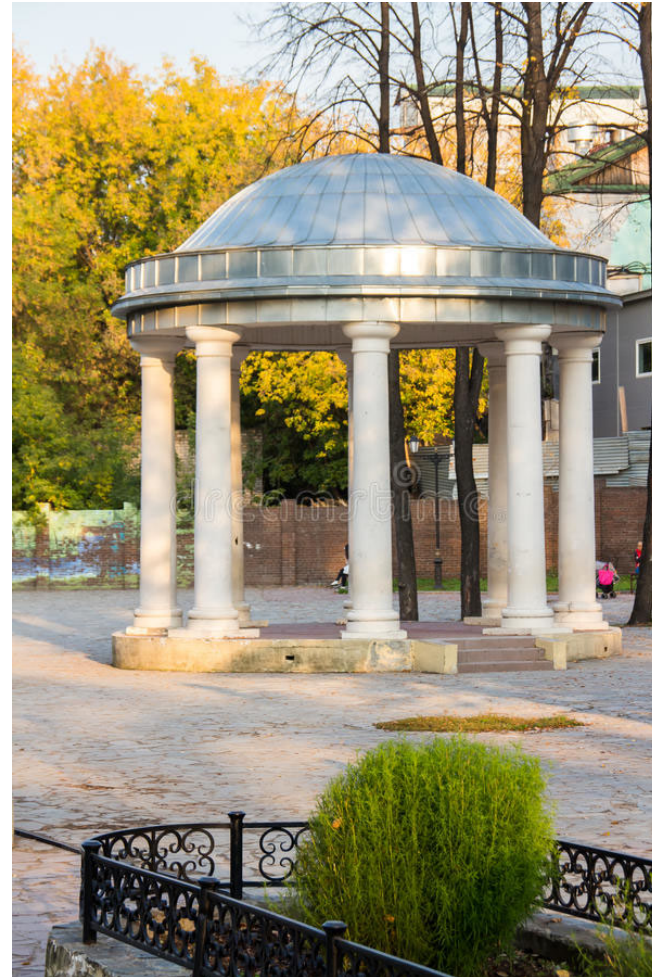
data

&

register categorisation

&

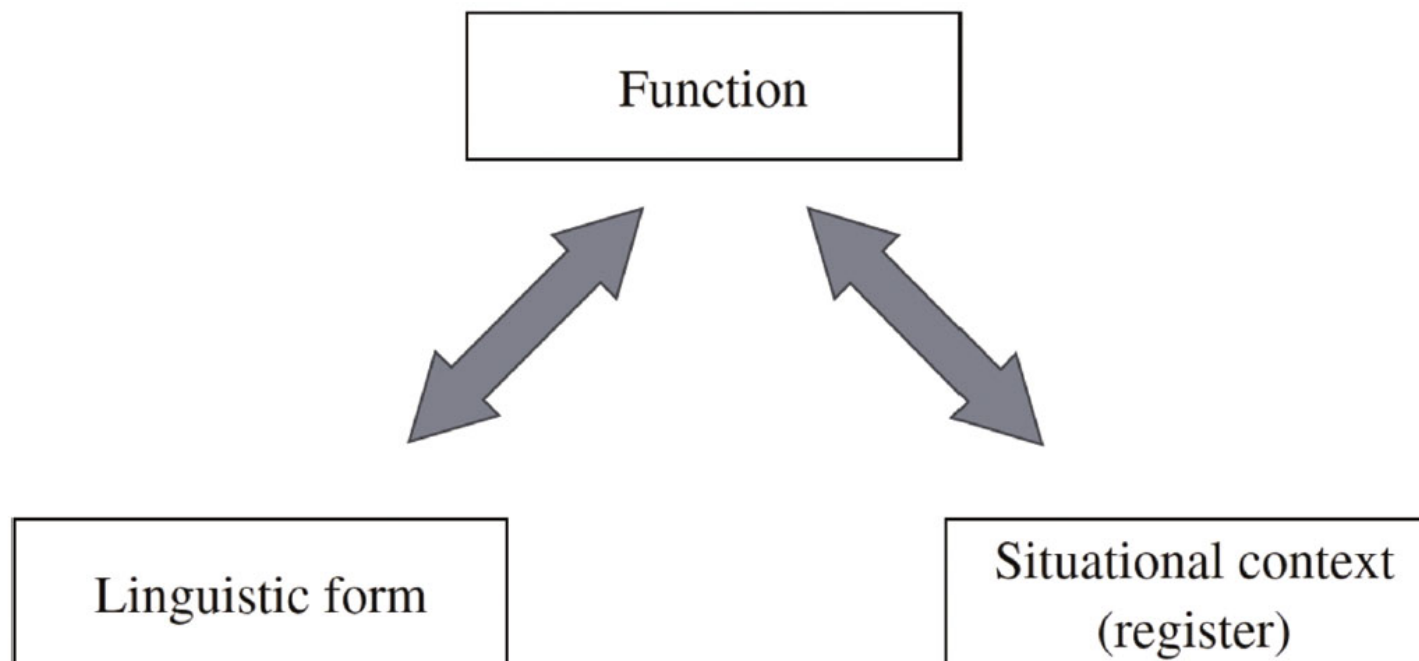
register interpretation



Pillars in rotunda in the autumn park 'Garden of Eden' (Perm)



# Rationale



(Egbert and Biber 2018: 238, based on Biber and Conrad 2009: 6-10)

“Linguistic co-occurrence can be regarded as a special type of linguistic variation; that is, **linguistic features co-occur in texts because they serve related functions**” (Egbert and Biber 2018: 239, my boldface)

# Rationale

## Pillar 2

data

&

register categorisation

&

register interpretation

&

Systemic Functional Linguistics (SFL)



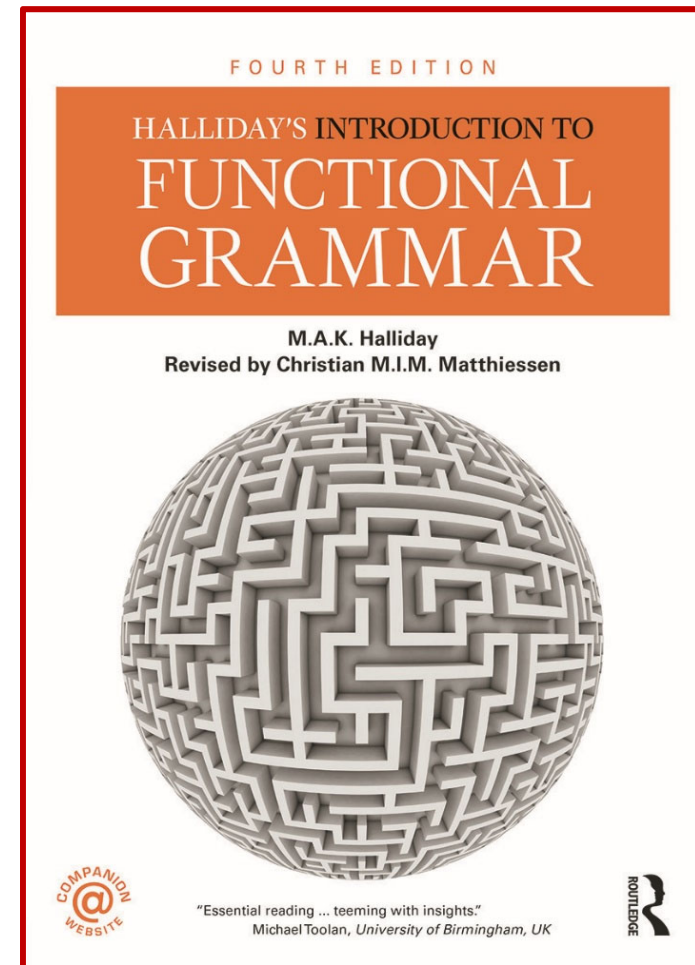
Sculpture Kama River (Perm)

# Rationale

## Systemic Functional Linguistics

Halliday and Matthiessen's  
*Introduction to Functional Grammar*,  
4th edition (IFG4), 2014

in 5(?) minutes...

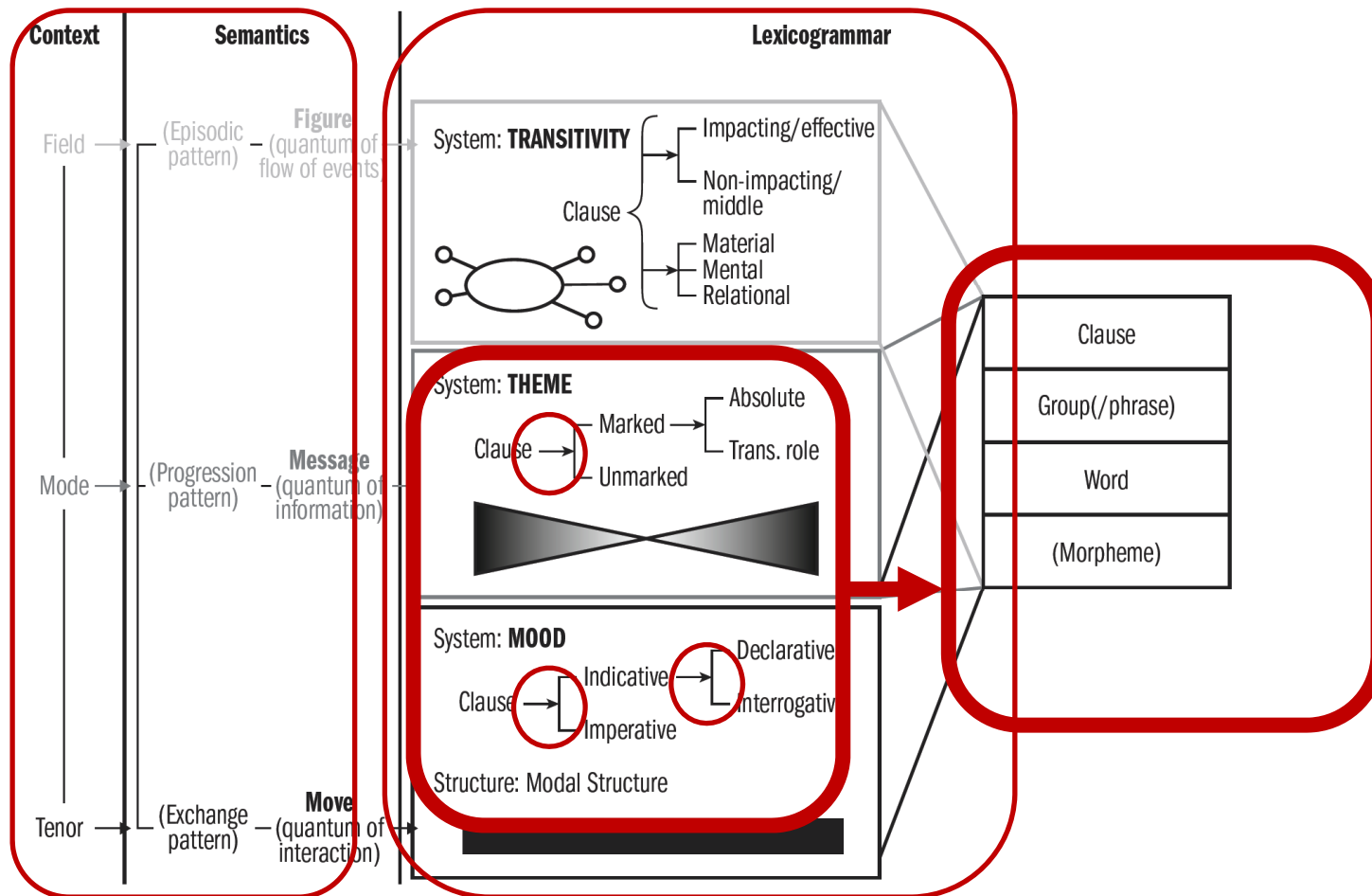


# Rationale

## **Systemic Functional Linguistics**

Assumptions:

- interdependence of language structure and language function
- language as a system of choices of meaning and not simply of choices of structures
- meaning originates in systemic patterns of choice:  
choice in a system > specific structure > specific function



- context and semantics run parallel to lexicogrammar
- Theme and Mood are materialised via 'constituency' (Clause, Group/Phrase, Word)
- arrows: systems > choices

# Rationale

## Systemic Functional Linguistics

Language defined...

- as sound, as writing and as wording
- as system
- as structure
- resource: **choices** among alternatives (IFG4: 20)



# Rationale

## Systemic Functional Linguistics

Fawcett (2008: 10, 2017: 58): “five great innovations” in Halliday’s SFL:

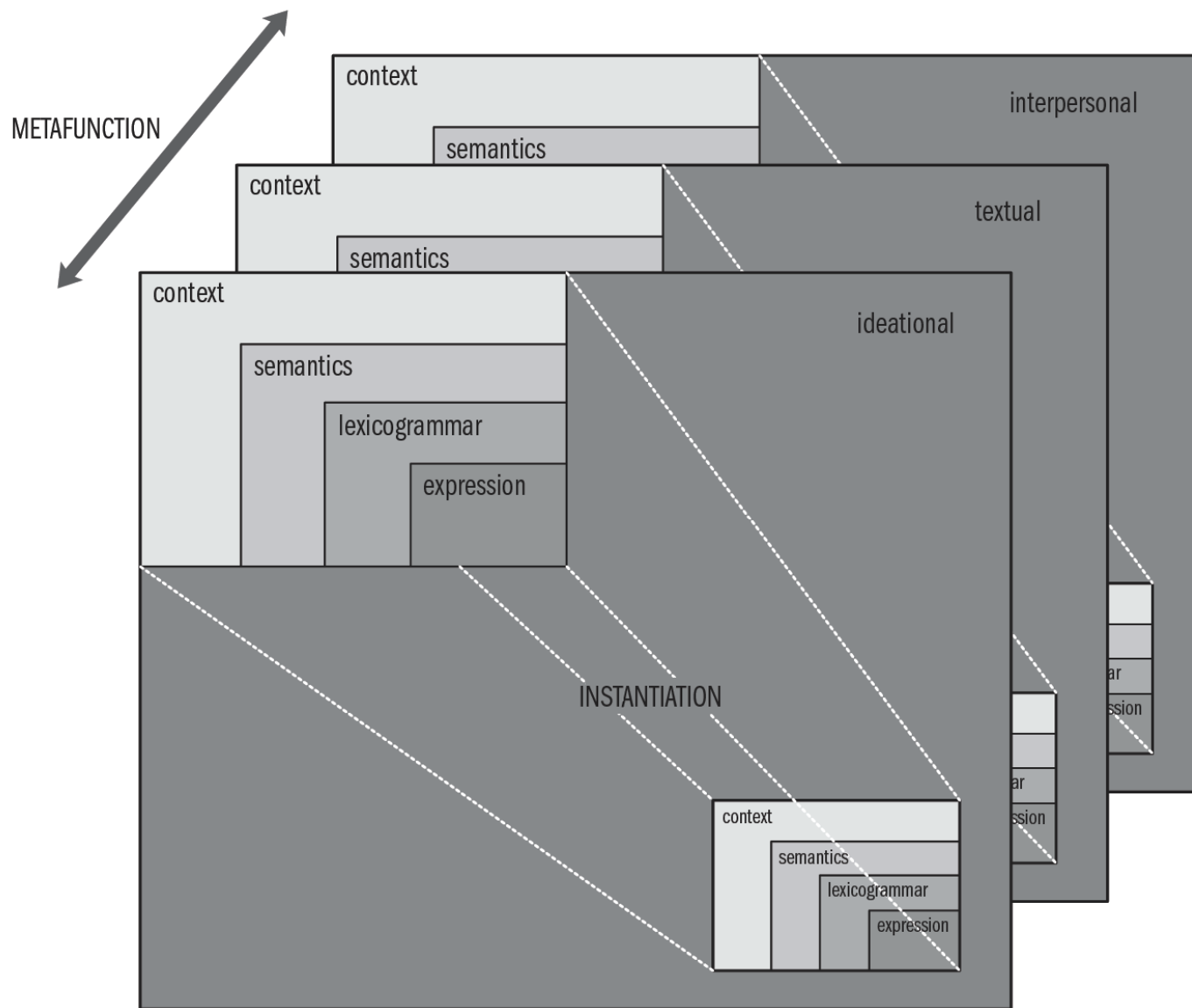
- concept of **system** and therefore the imposition of paradigmatic relations or system **choice** over structure
- **system networks** (Transitivity, Mood and Theme) as choices between meanings
- clause simultaneously realising several **metafunctions**

# Rationale

## Systemic Functional Linguistics

Language functions ~~functions~~ **metafunctions** (IFG4: 30):

- basic functions:
  - making sense of our experience: **ideational** metafunction
  - acting out our social relationships: **interpersonal** (*interactive & personal*) metafunction
- and a third function “that relates to the construction of text”:  
**textual** metafunction



# Rationale

## Systemic Functional Linguistics

Metafunction	Clause as ...	System	Structure
<b>textual</b>	message	THEME	Theme ^ Rheme
<b>interpersonal</b>	exchange	MOOD	Mood [Subject + Finite] + Residue [Predicator (+ Complement) (+ Adjunct)]
<b>experiential</b>	representation	TRANSITIVITY	process + participant(s) (+ circumstances), e.g. Process + Actor + Goal

# Rationale

## Textual metafunction > Theme

Theme as a structural notion, whose identification does not (necessarily) depend on the context:

“[w]e should rather **restrict theme and rheme to linguistic expressions** and take the context for what it is: a necessary prerequisite for all our utterances” (Fries 1984: 187)

In English “the Theme is indicated only by **position** in the clause. In speaking or writing English we signal that an item has thematic status by **putting it first.**” (IFG4: 88)

# Rationale

## Textual metafunction > Theme

But Theme choice implies context choices:

“The Theme is the element that serves as the point of departure of the message; it is that which **locates and orients the clause** within its context.” (IFG4: 89)

“The speaker chooses the Theme as his or her point of departure to **guide the addressee in developing an interpretation of the message**; by making part of the message prominent as Theme, the speaker enables the addressee to process the message.” (IFG4: 89)



# Rationale

Textual metafunction > Theme > Registers

Themes and registers go hand in hand

**Definition of register:**

variety of language associated with “particular social contexts [or situation types] in which language is used in specific ways” (van Leeuwen 1998: 278), ie. patterns of instantiation related to the contexts

“we describe registers as subpotentials (...) as recurrent patterns found in texts operating in contexts of situation at the instance pole of the cline of instantiation” (Matthiessen 2019: 28)

# Rationale

Textual metafunction > Theme > Registers

Themes and registers go hand in hand

“different varieties [registers] of English make use of this important first position [Theme] in different ways” (Berry 1995: 58)

“the experiential content of Themes do not occur randomly in (...) texts (...) [T]he experiential content of the Themes varies with genre type” (Fries 1995: 355)

“investigating such intertwinings [relationships between textual, interpersonal and ideational Themes, their positions and their progressions] might help distinguish genres” (Vande Kopple 1991: 331)

# Rationale

## Textual metafunction > Theme > Registers

- IFG4 (mostly in §3.1): **lexico-grammatical Theme trends** in...
  - guidebook (written, monologic, reporting)
  - taxonomic report (written monologic, expounding-categorising)
  - admission interview (spoken, dialogic, reporting)
  - novel (written, monologic, recreating-narrative) [in §3.4]
  - causal conversation (spoken, dialogic, sharing) [in §3.7]
  - constitution of an association (written, monologic, enabling-regulating) [§4.5.2]

# Rationale

## Theme variation(s) across text types and registers

### Semantics

- Berry's (2013) analysis of written and oral texts: strong **correlation:**
    - **contentful Subject Themes and formal written English**
    - **contentlight Subject Themes and informal spoken English**
- “[m]aybe contentlight Subject Themes in written English could be regarded as wasting words; every part of a clause should be used to convey the maximum amount of content, the part before the verb as well as the part after it” (Berry 2013: 259)

# Rationale

## Theme variation(s) across text types and registers

### Semantics

- **semantic types of (grammatical) subjects as indicators of text-type** ideological and epistemological positioning: MacDonald (1992, 1994) in psychology, history and literature research papers
- **(semantic) types of Themes across disciplines**: Taylor (1983) in science versus history school textbooks – conditional versus circumstantial adjuncts

# Rationale

**Theme variation(s) across text types and registers**

**Semantic and lexico-grammar features**

- McCabe (1999) in history texts
- El-Issa (2016) in tourist guides



# Rationale

## Theme variation(s) across text types and registers

### Discourse purpose and subject matter

- Lavid (2000): chaining strategies in expository, descriptive, narrative, instructive, persuasive texts:  
“statistically significant **correlations between the global chaining strategies** which characterize specific text types, **and the semantic type of themes** selected to signal those strategies. For example, texts which select a temporal chaining strategy to organize information globally tend to signal that strategy by means of temporal themes, while texts globally structured by means of a spatial strategy tend to signal it by means of locative themes”  
(p.44)

# Rationale

## Theme variation(s) across text types and registers

### Discourse purpose and subject matter

- North (2005: 449):  
“‘arts’ students (...) use **interpersonal orienting themes** [versus] ‘science’ students, who tended to make more use of **unqualified assertions**. These differences in the students’ discursive practices may derive from the different views of knowledge in soft and hard disciplines”

### (Simple/constant/derived) **Thematic Progression**

- Nwogu and Bloor (1991): professional/popular medical texts
- McCabe (2004): history textbooks

# Goal

So...

If SFL **Theme** is relevant to register characterisation,  
let's investigate **Themes in different registers**.

We need...



1. a theory, a taxonomy of Themes
2. data

# Goal

## 1. Theme taxonomy

Let's identify Theme(s)...

Main segmental hypotheses (Berry 1996: 29–31)

- Halliday/IFG4 (also 1985: 54, 1994: 53):  
‘first ideational element’ hypothesis >  
‘first-element’ hypothesis 
- Enkvist (1973): ‘subject’ hypothesis
- Berry (1995): ‘preverb’ hypothesis 
- Berry (1996): ‘lexical verb’ hypothesis



# Goal

## 1.1. 'First-element' hypothesis (IFG4)

### Theme constituency

- simple (ideational, topical, experiential) Theme:  
“the Theme of a clause ends with the first constituent that is either participant, circumstance or process” (IFG4: 105)
- multiple: topical (ideational or experiential) Theme plus previous elements fulfilling a textual or an interpersonal function

# Goal

## 1.1. 'First-element' hypothesis (IFG4)

Simple (topical, experiential) Theme:

MOOD		THEME		
indicative	declarative	unmarked	Subject	London Bridge # is fallen down What I want # is a proper cup of coffee There # were three jovial Welshmen
		marked	Adjunct	On Saturday night # I lost my wife
			Complement	This # they should refuse How dreadful # she sounds
	interrogative: yes/no	unmarked	Verbal operator + Subject	Didn't it # smell terrible? Are they # still together?
		marked		On the right # is it?
interrogative: WH		WH-element	Where # did you get that from?	
imperative		unmarked	Verb (Predicator)	Turn # it down Don't # do that Let's # do lunch at the Ivy Let me # send Lesley a photocopy
		marked		Don't you # open it

# Goal

## 1.1. 'First-element' hypothesis (IFG4)

Multiple Theme(s):

textual	continuative	<i>yes, no, well, oh, now</i>
	conjunction	<i>and, or, neither, but, then, when, while, because, if, although, in order to, in spite of the fact that, that, whether, with (in With all the doors # being locked, ...)</i>
	conjunctive Adjunct	<i>for example, by the way, in short, but, , as I was saying</i>
interpersonal	modal/comment Adjunct	<i>in my opinion, no doubt, of course, probably, sometimes, on the whole</i>
	vocative	
	finite verbal operator	(yes/no interrogatives)

well	but	then	surely	Jean	wouldn't	the best idea	be to join in
cont	stru	conj	modal	voc	finite	topical	
Theme							Rheme

# Goal

## 1.1. 'First-element' hypothesis (IFG4)

Yesterday Ivy bought a new pair of shoes.

Perhaps she is feeling rich.

[I've seen King Lear several times–] (but) Coriolanus I haven't seen.

In the morning, quite unexpectedly, she told her parents that she was leaving home.

[Last month Ike lost four pounds in weight,] but this month, sadly, because he's been on holiday in France, he's put it all back on again.

On the table stood a lamp.

In fact it is love that makes the world go round. [predicated Themes (~clefts), expletive]



# Goal

## 1.2. 'Preverb' hypothesis (Berry)

experiential (topical)	lexical subject
	expletive subject
	(marked) adverbial
	(marked) complement: object, prepositional complement, predicative
textual	conjunction
	conjunct
interpersonal	sentence modifier (~disjunct)
	interjection
	wh-argument/adverbial
	(main/auxiliary) verb in imperative, inverted clauses

# Goal

## 1.2. 'Preverb' hypothesis (Berry)

Yesterday Ivy bought a new pair of shoes.

Perhaps she is feeling rich.

[I've seen King Lear several times–] (but) Coriolanus I haven't seen.

In the morning, quite unexpectedly, she told her parents that she was leaving home.

[Last month Ike lost four pounds in weight,] but this month, sadly, because he's been on holiday in France, he's put it all back on again.

On the table stood a lamp.

In fact it is love that makes the world go round. [predicated Themes (~clefts), expletive]

# Goal

## 2. Data

a multi-register corpus? But... does SFL like **corpora** at all?

- corpora as “fundamental to the enterprise of theorizing language” (IFG4: 51)
- “three plusses” (IFG4: 53):
  - + authentic data
  - + spoken language included
  - + “the corpus makes it possible to study grammar in quantitative terms. (...) grammatical systems are probabilistic in nature”

# Goal

## 2. Data

Need of **large-scale empirical research**:

- Real texts:

In many cases [in SFG] texts have not been naturally occurring texts, but they have been especially **constructed** by the analyst for the purpose of the study (Berry 1987: 72)

- Size:

the amount of material examined has been **very small** (Berry 1987: 79)

it would be foolish to draw too firm conclusions from so small a study (Berry 1987: 84)

# Goal

## Theoretical goal

- Hypothesis: registers are determined by formal, structural features
- RQs:
  - Are SFL Themes suitable for register characterisation?
  - Are some SFL Themes more suitable for register characterisation than others?

## Empirical aim

- large-scale analysis of Themes in PDE based on:
  - authentic written textual sources
  - a bunch of text types

# Data

- **Crown corpus** (Xu and Liang 2013)
  - Present-Day (2008–2011) American English
  - Brown family:

	Genre	Sub-corpus tokens	Total tokens		Sub-corpus tokens	Total tokens
<b>Brown 1961</b>	Fiction	259,467	1,027,021	<b>LOB 1961</b>	258,722	1,018,785
	General prose	423,160			418,137	
	Learned	163,309			162,322	
	Press	181,085			179,604	
<b>Frown 1992</b>	Fiction	260,414	1,027,323	<b>FLOB 1991</b>	260,664	1,024,643
	General prose	421,933			419,990	
	Learned	163,228			163,286	
	Press	181,748			180,703	
<b>Crown 2009</b>	Fiction	259,250	1,026,226	<b>CLOB 2009</b>	259,484	1,023,466
	General prose	422,799			421,163	
	Learned	163,197			163,139	
	Press	180,980			179,680	

# Data

- **Database**

- 91,267 clauses, as identified and analysed by the parser
- 1,010 examples of incomplete (fragment, verbless) clauses disregarded
- definitive database: **90,257 clauses**

# Data

- **Database**
  - text types:

	Crown
A	Press: Reportage
B	Press: Editorial
C	Press: Reviews
D	Religion
E	Skill and hobbies
F	Popular lore
G	Belles-lettres
H	Miscellaneous
J	Learned
K	Fiction: General
L	Fiction: Mystery
M	Fiction: Science
N	Fiction: Adventure
P	Romance
R	Humour



# Data

So... we have:

- two definitions of Theme
- a multi-register/genre/text-type (raw) corpus

<file= AmE06\_A01>

City councilors scored a victory over the mayor last week, but it's a vote that may come back to haunt them.

The council overrode Mayor Jeannette A. McCarthy's veto of a zoning change that allows more intensive development of two properties in the Totten Pond Road-Route 128 area, to the dismay of more than three- dozen residents who turned out on one of the hottest nights of the summer for the special session.

Supporters of the change say it will spur revitalization of an area dotted with aging office buildings. Opponents see it as a symbol of big business trumping neighborhood interests.

The change paves the way for renovation or replacement of buildings that are at least 35 years old with structures that could rise up to six stories and house offices, shops, and restaurants.

When the council first voted on the proposal on June 26, supporters said it would potentially benefit six properties. As it turns out, only two parcels would be eligible, both owned by real estate giant Boston Properties.

# Data

And... we need to **identify Themes**.

**DAT (Detection and Annotation of Themes) analyser**



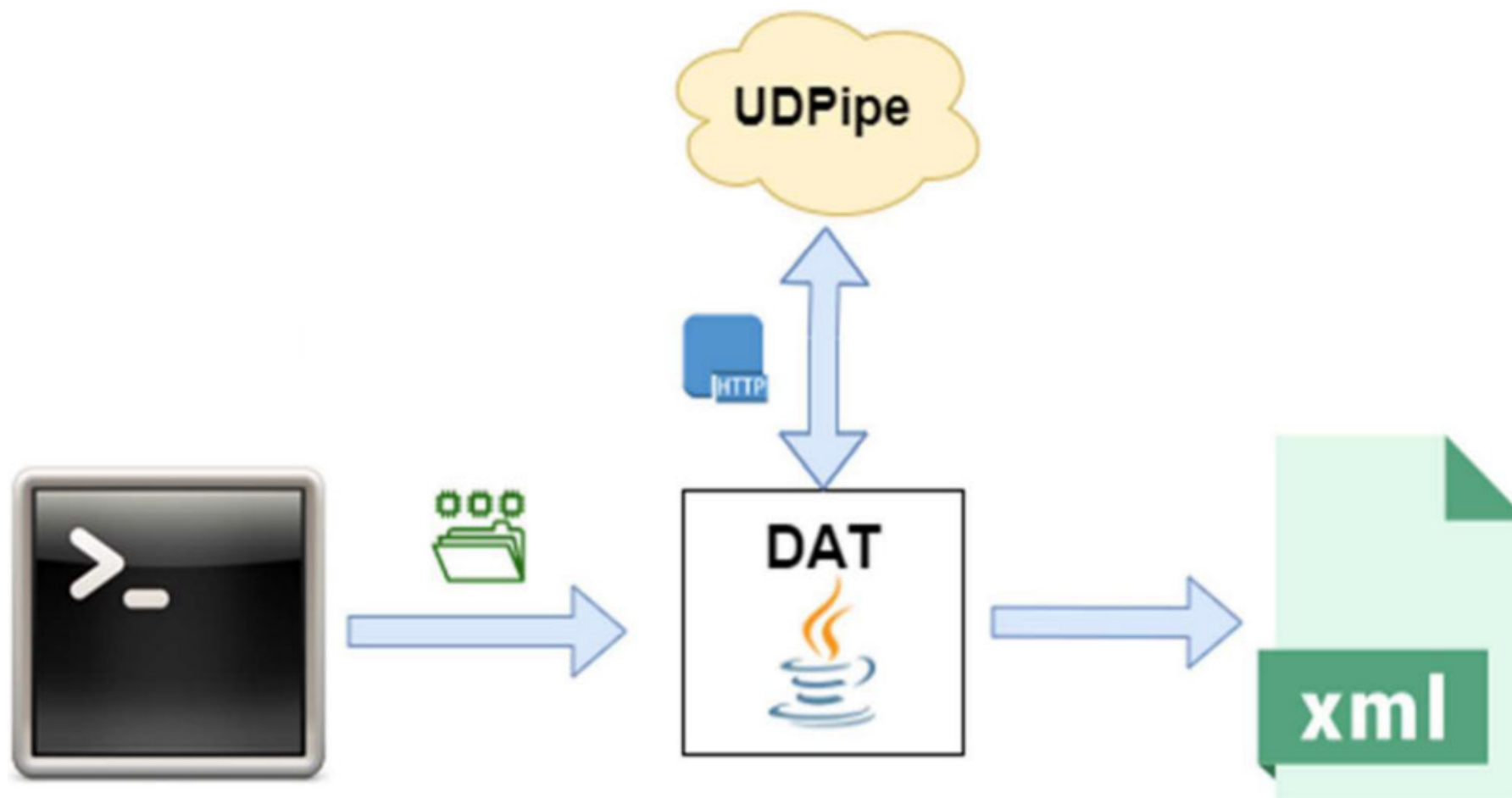
Phases

- **reader**: reads txt files, splits them into 1MB-files, as required by the API REST in UDPipe (parser)
- **parser**: established HTTP connection to the UDPipe's API REST (REpresentational State Transfer), which provides the morphosyntactic parsing of the texts

UDPipe (Institute of Formal and Applied Linguistics, Charles University, Czech Republic, english-ewt-ud-2.4-190531)

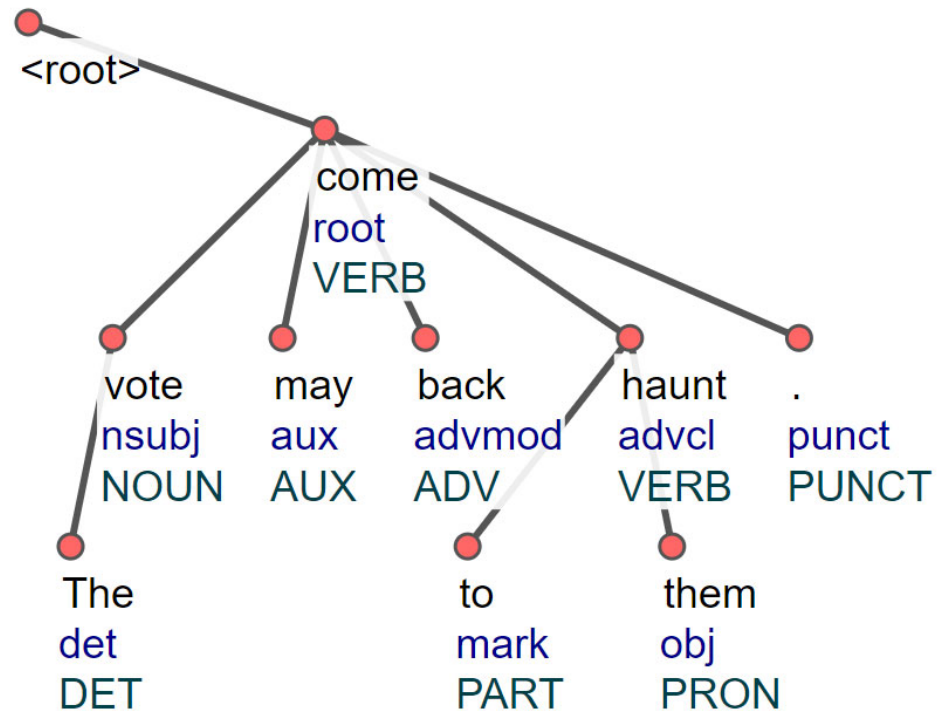
<http://lindat.mff.cuni.cz/services/udpipe/>

# Data



# Data

The vote may come back to haunt them .



	Nominals	Clauses	Modifier words	Function Words		
<b>Core arguments</b>	<a href="#">nsubj</a>	<a href="#">csubj</a>				
	<a href="#">obj</a>	<a href="#">ccomp</a>				
	<a href="#">iobj</a>	<a href="#">xcomp</a>				
<b>Non-core dependents</b>	<a href="#">obl</a>	<a href="#">advcl</a>	<a href="#">advmod*</a>	<a href="#">aux</a>		
	<a href="#">vocative</a>		<a href="#">discourse</a>	<a href="#">cop</a>		
	<a href="#">expl</a>			<a href="#">mark</a>		
	<a href="#">dislocated</a>					
<b>Nominal dependents</b>	<a href="#">nmod</a>	<a href="#">acl</a>	<a href="#">amod</a>	<a href="#">det</a>	<b>Head</b>	<b>DepRel</b>
1	<a href="#">appos</a>			<a href="#">clf</a>	2	det
2	<a href="#">nummod</a>			<a href="#">case</a>	4	nsubj
<b>Coordination</b>	MWE	Loose	Special	Other	4	aux
3	<a href="#">conj</a>	<a href="#">fixed</a>	<a href="#">list</a>	<a href="#">orphan</a>	0	root
4	<a href="#">cc</a>	<a href="#">flat</a>	<a href="#">parataxis</a>	<a href="#">goeswith</a>	4	advmod
5	<a href="#">compound</a>		<a href="#">reparandum</a>	<a href="#">dep</a>	7	mark
6	to	to	TO	-	4	advcl
7	haunt	haunt	VERB	VB		
				VerbForm=Inf		
			<b>Open class words</b>	<b>Closed class words</b>		
			<a href="#">ADJ</a>	<a href="#">ADP</a>	<a href="#">PUNCT</a>	number=Plur Per
8	them	they	<a href="#">ADV</a>	<a href="#">AUX</a>	<a href="#">SYM</a>	Type=Prs
9	.	.	<a href="#">INTJ</a>	<a href="#">CCONJ</a>	<a href="#">X</a>	
			<a href="#">NOUN</a>	<a href="#">DET</a>		
			<a href="#">PROPN</a>	<a href="#">NUM</a>		
			<a href="#">VERB</a>	<a href="#">PART</a>		
				<a href="#">PRON</a>		
				<a href="#">SCONJ</a>		

# Data

## DAT

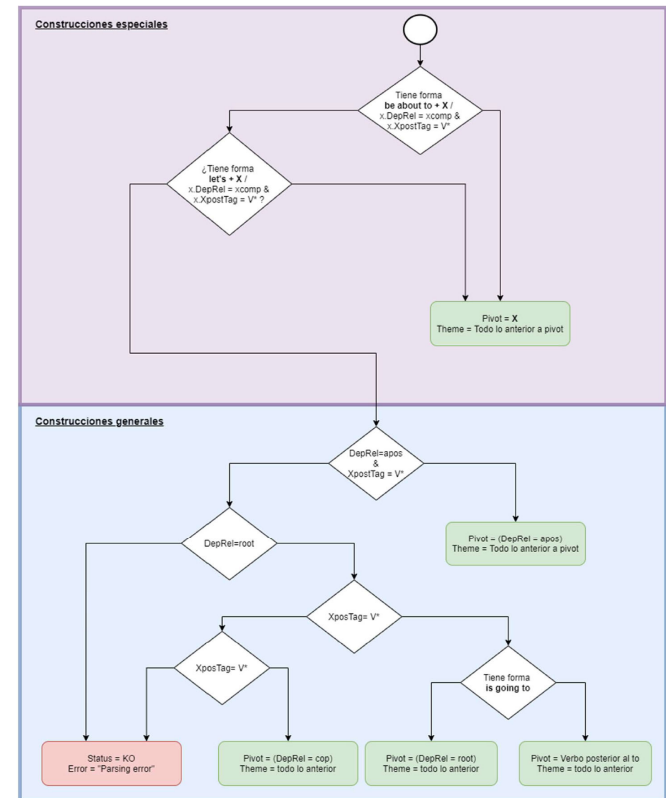
### Parser > Processor

- first, recognition of so-called ‘special’ constructions that require ad-hoc decision by lemma. Eg.:

*Children, the story is about to continue.*

*Let's cook at the Ivy*

- second, recognition of so-called ‘regular’ constructions
- third, identification of Themes by way of the so-called ‘Pivots’, ie. elements whose recognition automatically determines the thematic structure of the example



# Data

## DAT

### Parser > Processor

- if the parser analysed a form as DepRel=appos and as XPosTag=V\*,
  - categorise this form as Pivot
  - categorise all the forms previous to the Pivot as Theme
- otherwise, if the parser analysed a form as DepRel=root and as XPosTag=V\*,
  - if the form is *going*,
    - if *going* is followed firstly by a form categorised as XPosTag=TO
    - and if *going* is followed secondly by a form with XPosTag=V\*
      - categorise the form with XPosTag=V\* following the form with XPosTag=TO as Pivot
      - categorise all the forms previous to the Pivot as Theme
  - (...)

# Data

## DAT

### Parser > Processor

- if the form is not *going*,
  - categorise the form with DepRel=root as Pivot
  - categorise all the forms previous to the Pivot as Theme
- if there are no forms compliant with DelRel=root and XPosTag=V\*,
  - if there is a form parsed as DepRel=cop,
    - categorise the form with DepRel=cop as Pivot
      - if there are preceding forms, categorise all the forms previous to the Pivot as Theme
      - if there are no preceding forms, categorise the Pivot as Pivot and Theme.
- etc., with primary and (all) secondary Pivots (> Themes)



# Data

## DAT

### Parser > Processor

Id	Form	Lemma	XPosTag	Head	DepRel			
# text = Well guess I'll go and put my washing on the line.								
1	Well	well	UH	2	discourse	theme		
2	guess	guess	VB	0	root	pivot		
3	I	I	PRP	6	nsubj	theme		
4	'	'	.	6	punct			
5	ll	will	MD	6	aux			
6	go	go	VB	2	ccomp	pivot		
7	and	and	CC	8	cc			theme
8	put	put	VB	6	conj	pivot		
9	my	my	PRP\$	10	nmod:poss			
10	washing	washing	NN	8	obj			
11	on	on	IN	13	case			
12	the	the	DT	13	det			
13	line	line	NN	8	obl			
14	.	.	.	2	punct			

# Data

## DAT

Parser >

Processor >

Writer

```
<sentence ref="25" text="They live outside the law but pay taxes." state="OK">
- <units>
  - <unit ref="1">
    - <pivot>
      <word head="0" depRel="root" xPosTag="VBP" id="2">live</word>
    </pivot>
    - <theme length="1" scheme="nsubj">
      <word head="2" depRel="nsubj" xPosTag="PRP" id="1">They</word>
    </theme>
  </unit>
  - <unit ref="2">
    - <pivot>
      <word head="2" depRel="conj" xPosTag="VBP" id="7">pay</word>
    </pivot>
    - <theme length="1" scheme="">
      <word head="7" depRel="cc" xPosTag="CC" id="6">but</word>
    </theme>
  </unit>
</units>
- <words>
  <word head="2" depRel="nsubj" xPosTag="PRP" id="1">They</word>
  <word head="0" depRel="root" xPosTag="VBP" id="2">live</word>
  <word head="5" depRel="case" xPosTag="IN" id="3">outside</word>
  <word head="5" depRel="det" xPosTag="DT" id="4">the</word>
  <word head="2" depRel="obl" xPosTag="NN" id="5">law</word>
  <word head="7" depRel="cc" xPosTag="CC" id="6">but</word>
  <word head="2" depRel="conj" xPosTag="VBP" id="7">pay</word>
  <word head="7" depRel="obj" xPosTag="NNS" id="8">taxes</word>
  <word head="2" depRel="punct" xPosTag="." id="9">.</word>
</words>
</sentence>
```

# Data

## DAT (Detection and Annotation of Themes) analyser

### Operationality:

- pilot: AmE06-A, AmE06-C, AmE06-B, AmE06-D: **210,794 words**
- machine: Intel Core i7-7700 @ 2.90 GHz
- total time: **04m 23s 460ms**

### Reliability:

- correctly parsed units: 88.29%
- incorrectly parsed units: 11.71%
  - due to UDPipe: 11.69%
  - due to DAT: 0.02%
- So... **DAT reliability: 99.85%**

# Analysis of the data

## Case study: Let's recap

- RQ1: Can the SFL concept Theme be a proxy for the identification of textual categories?
- Theory: two concepts of Theme:
  - first-element
  - preverb
- Data:
  - 90,257 examples
  - 15 textual categories

textual category	clauses
belles_letr	13636
fiction_advent	6411
fiction_gen	6487
fiction_mystery	5719
fiction_science	1302
humour	1983
learned	11628
misc_governm	3386
pop_lore	8600
press_edit	4446
press_report	7708
press_review	2762
religion	2693
romance	7422
skill	6074
<b>Total</b>	<b>90257</b>

# Analysis of the data

## Case study: Let's recap

- Technology:
  - parser > syntactic analysis
  - DAT > identification of Theme types according to:
    - first-element hypothesis
    - preverb hypothesis
- RQ2: If RQ1 has an affirmative answer, which is the winning hypothesis (first-element vs preverb)?

# Analysis of the data

## Theme taxonomy

- association of syntactic functions to Theme types

parser	database
advcl	exp_marked_a
advp	exp_marked_a
subjcl	exp_subj
modx	inter_inter
modap	inter_inter
conjunct	text
there	exp_expl
mod	inter_inter
dobj	exp_marked_a
interjection	inter_inter
spred	exp_marked_a
advwh	inter_w
compwh	inter_w
subjwh	inter_w

<b>experiential (or topical)</b>	exp_subj	no. of unmarked experiential themes (lexical subjects)
	exp_expl	no. of unmarked experiential themes (expletive subjects)
	exp_marked	no. of marked experiential themes
<b>textual</b>	text	no. of textual themes
<b>interpersonal</b>	inter_w	no. of interpersonal themes (wh-words)
	inter_inter	no. of interpersonal themes (vocatives, interpersonal modifiers)
	inter_verb	no of interpersonal themes ((main/auxiliary) verbs in subjectless_imperative, inv_main and inv_subord clauses: value=1)

# Analysis of the data

## Theme taxonomy: data: first-element Themes

	exp_subj	exp_expl	exp_marked_a	exp_marked_c	text	inter_w	inter_inter	inter_verb	
belles_letr	9733	185	2373	257	503	1046	100	42	
fiction_advent	4798	95	832	214	176	453	35	19	
fiction_gen	4821	92	899	198	316	462	45	15	
fiction_mystery	4356	62	713	203	235	371	24	14	
fiction_science	950	21	193	40	50	94	6	4	
humour	1454	28	309	50	78	135	7	7	
learned	8143	150	2189	224	316	902	81	20	
misc_governm	2392	37	685	29	106	237	12	6	
pop_lore	6098	124	1546	149	356	649	61	34	
press_edit	3242	75	710	69	116	343	54	7	
press_report	5529	93	1262	282	176	523	80	19	
press_review	1966	49	458	49	88	227	17	13	
religion	1911	30	478	50	67	216	27	8	
romance	5636	78	826	308	450	550	44	24	
skill	4341	80	1084	94	241	462	42	13	
<b>Total</b>	<b>65370</b>	<b>1199</b>	<b>14557</b>	<b>2216</b>	<b>3274</b>	<b>6670</b>	<b>635</b>	<b>245</b>	<b>94166</b>

# Analysis of the data

## Theme taxonomy: data: preverb Themes

	exp_subj	exp_expl	exp_marked_a	exp_marked_c	text	inter_w	inter_inter	inter_verb	
belles_letr	13007	281	2918	292	503	1028	100	283	
fiction_advent	6131	103	1228	115	176	447	35	156	
fiction_gen	6214	115	1165	146	316	510	45	123	
fiction_mystery	5459	96	985	123	235	416	24	138	
fiction_science	1247	23	239	42	50	120	6	27	
humour	1883	42	388	39	78	121	7	45	
learned	11125	191	3140	285	316	907	81	259	
misc_governm	3256	47	1105	80	106	254	12	67	
pop_lore	8236	140	1948	266	356	577	61	185	
press_edit	4264	70	1202	136	116	288	54	90	
press_report	7371	118	1596	246	176	574	80	184	
press_review	2623	55	673	50	88	185	17	72	
religion	2571	45	680	81	67	178	27	66	
romance	7062	148	990	183	450	654	44	181	
skill	5827	106	1522	138	241	412	42	112	
<b>Total</b>	<b>86276</b>	<b>1580</b>	<b>19779</b>	<b>2222</b>	<b>3274</b>	<b>6671</b>	<b>635</b>	<b>1988</b>	<b>122425</b>



# Analysis of the data

Let's **cluster** the textual categories:

- per frequencies (RQ1) of Theme types:
  - experiential (unmarked) subjects (lexical versus expletive)
  - (experiential) marked adverbials, complements (objects, predicatives, prepositional complements)
  - textual
  - interpersonal (wh, verbs, clausal modifiers, interjections)
- per (first-element / preverb) Theme hypothesis (RQ2)

# Analysis of the data

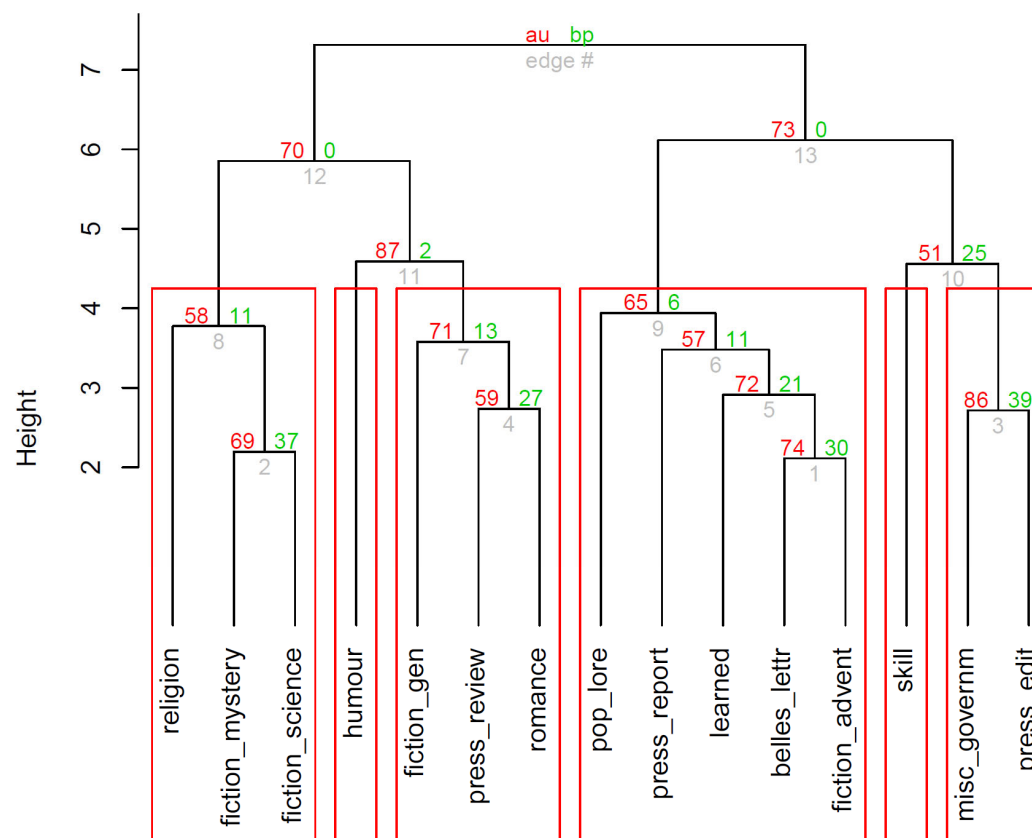
## Behavioural Profiles approach

- Data represented as (BP) vectors of proportions of the different values in the variables (R, 'bp' function):
- Computation of numerical differences between vectors as 'distances' (R, 'dist' function)
- Hierarchical agglomerative clustering (R, 'hclust' function, 'ward.D2' method)
- Optimal number of clusters determined by function 'silhouette' (R)
- Multiscale bootstrap (R, 'pvclust' function) to validate the clusters through the AU (Approximately Unbiased)  $p$ -value (the closer to 1, the greater the statistical support)
- Identification of (sub-)clusters supported at the level of AU  $p \geq 0.95$



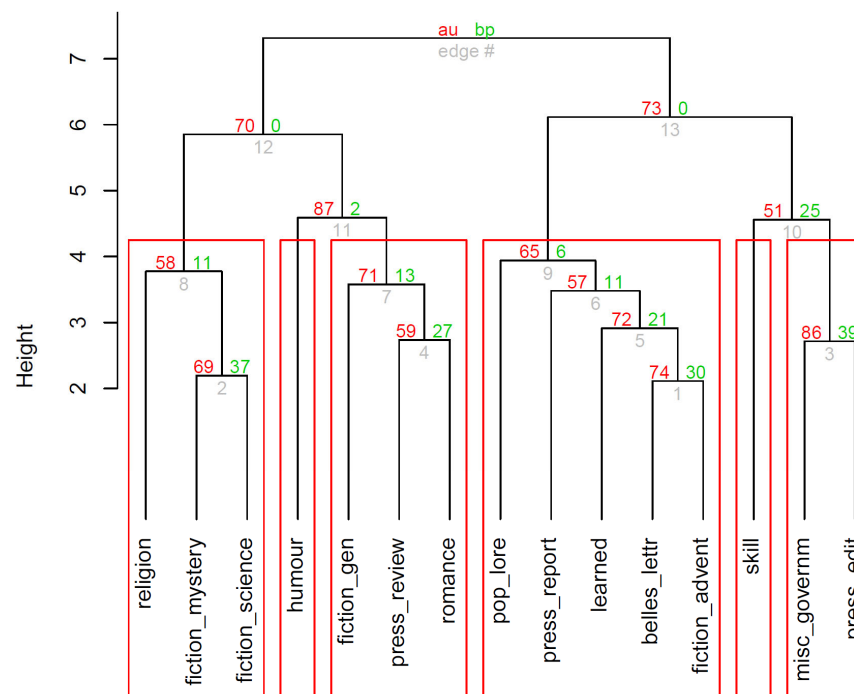
# Analysis of the data

## Preverb hypothesis



# Analysis of the data

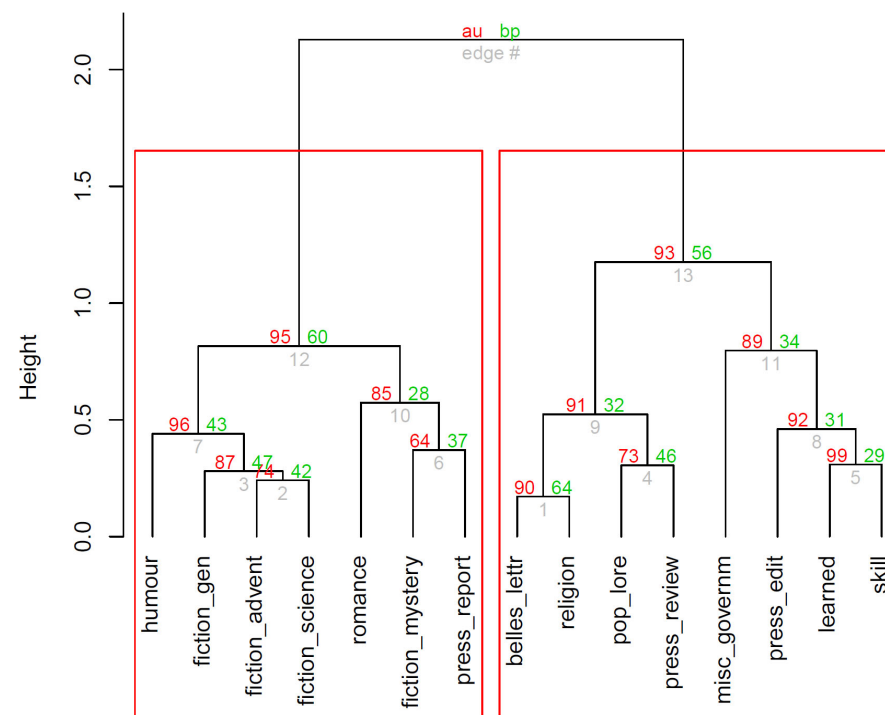
## Preverb hypothesis



- 15 textual categories organised into 6 statistically optimal clusters, as determined by 'silhouette' (R)
  - AU  $p$ -values: rather low ( $\leq 0.9$ )
  - little relation to conventional register division
- So... clustering based on preverb Themes not fruitful

# Analysis of the data

## First-element hypothesis



- 15 textual categories organised into two major clusters
- AU  $p$ -values  $\sim > 0.95$  (best fit)
- 2 clusters: (mostly) popular vs learned/formal registers

So... clustering based on first-element Themes is methodologically suitable and reflects a valid situational or communicative divide among texts

# Conclusions

## Assumptions

- different registers target different audiences and serve different communicative purposes
- register variance implemented through different formal features and strategies

## Hypothesis

- thematic design of clauses may predict register membership, ie. SFL Theme may be a proxy for the identification of registers

# Conclusions

## Case study

- empirical analysis of the thematic components of:
  - >90,000 clauses of written PDE
  - 15 registers
- two major definitions of Theme in SFL:
  - ('orthodox' IFG) 'first-element' approach
  - Berry's 'preverb' hypothesis



# Conclusions

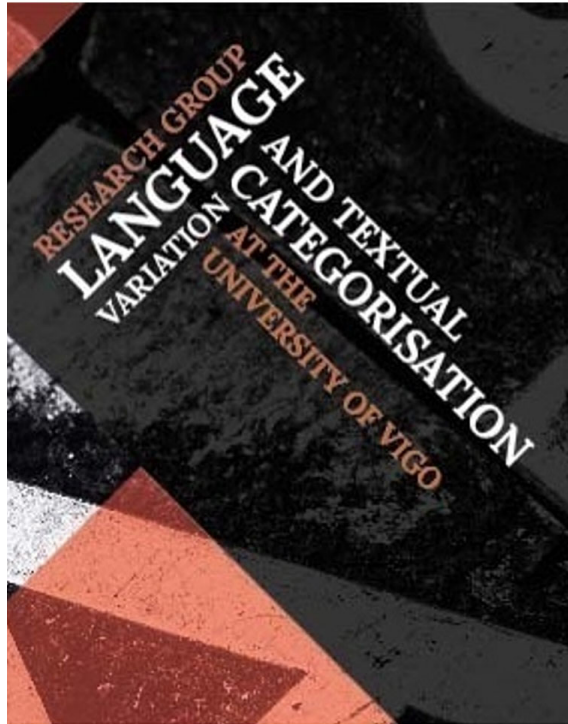
## Case study

- clustering of registers:
  - SFL Theme proved to be suitable for the categorisation of registers
  - first-element Theme is a better dissimilarity metric
    - methodology: better statistical indices
    - interpretation: two clusters:  
popular registers vs learned-formal registers

So...

language users adopt specific linguistic strategies to meet specific situational demands 👍

lvtc



Universidade de Vigo

# (Systemic Funcional) Theme as an indicator of register variation in English

Javier Pérez-Guerra  
jperez@uvigo.es

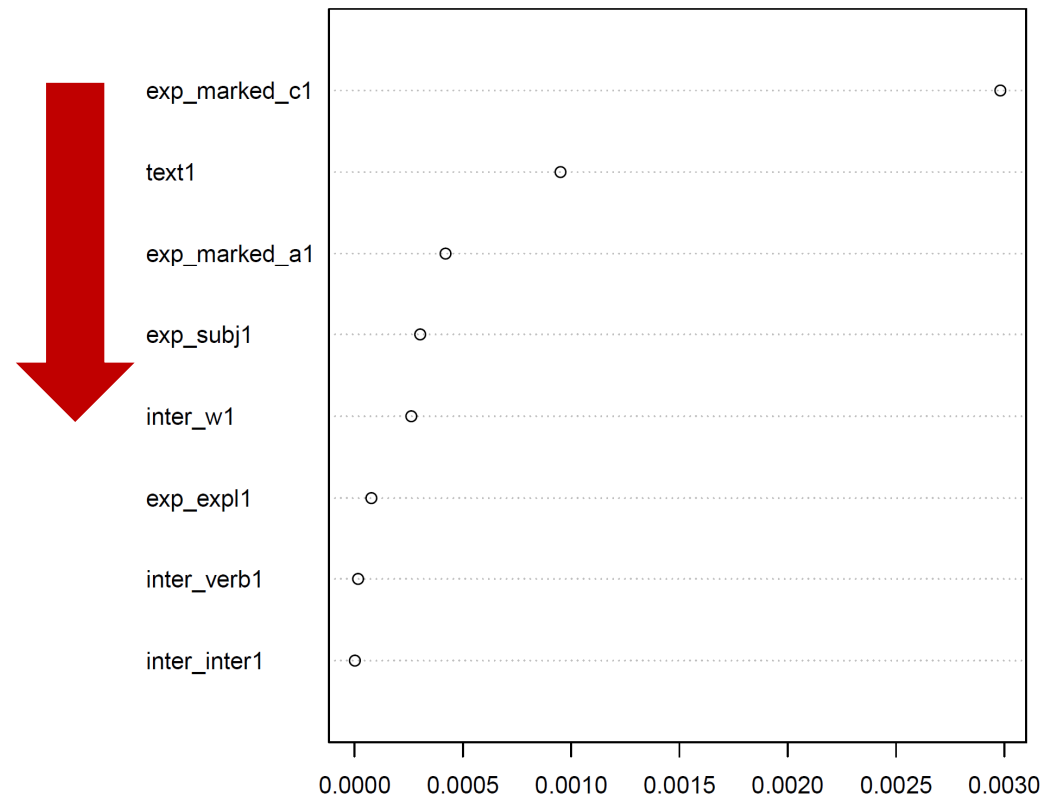


# Further research

- only main clauses? distinction main vs bound clauses? – bound clauses: “not presented by the speaker as being open for [thematic] negotiation” (IFG4: 170):
  - sometimes nonfinite and not translatable into finite:  
*Marisa wants me to stop* > \**Marisa wants that I stop*
  - sometimes subjectless (nonfinite):  
*Marisa wants  $\emptyset$  to open the discussion period*
- additional registers, also speech-based/related and spoken texts
- other varieties
- method:
  - individual texts as random predictors
  - fine-grained study of experiential, interpersonal and textual Themes within the clusters

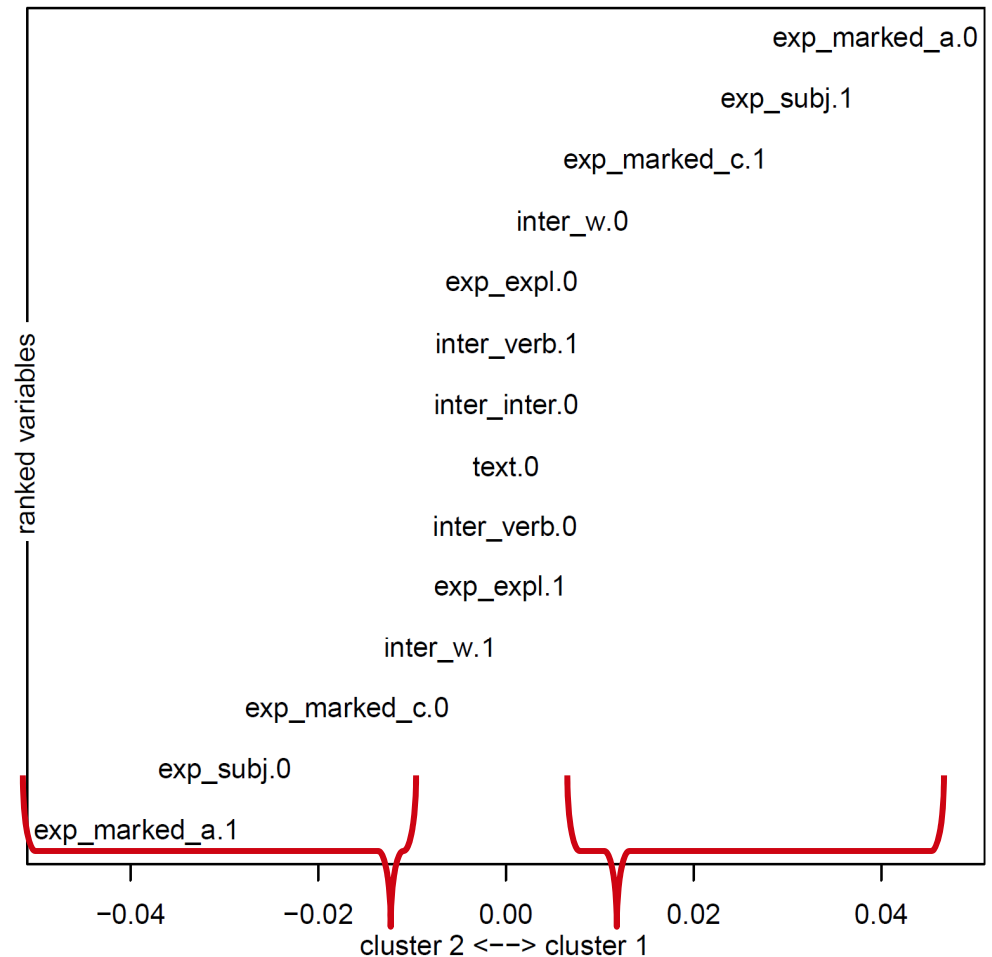
# Analysis of the data

Conditional importance of predictors (Random Forest):  
first-element Themes



# Analysis of the data

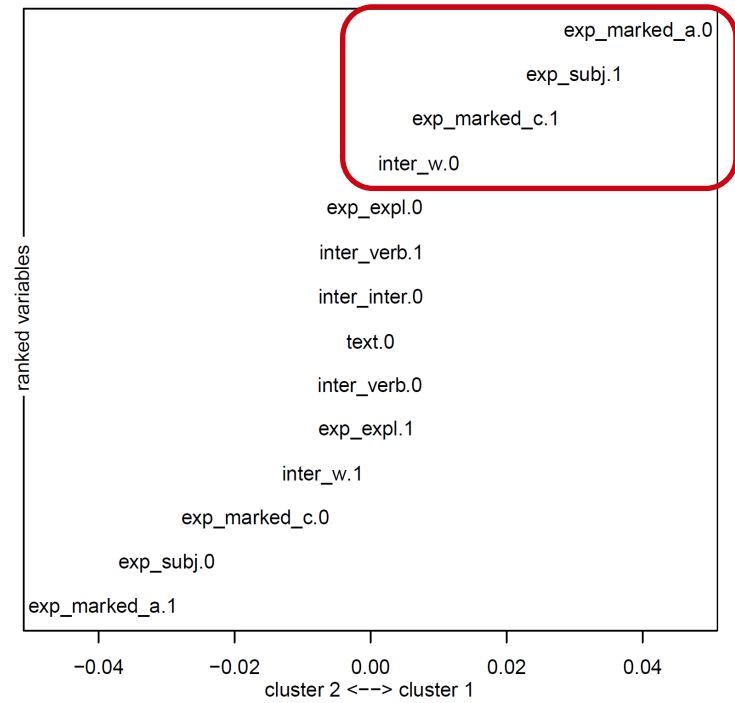
Snake plot of predictors  
after **first-element** clustering



- x-axis: sorted scores (ascending order)
- y-axis: variables according to their ranks in clusters:
  - cluster 1 (popular)
  - cluster 2 (learned)

# Analysis of the data

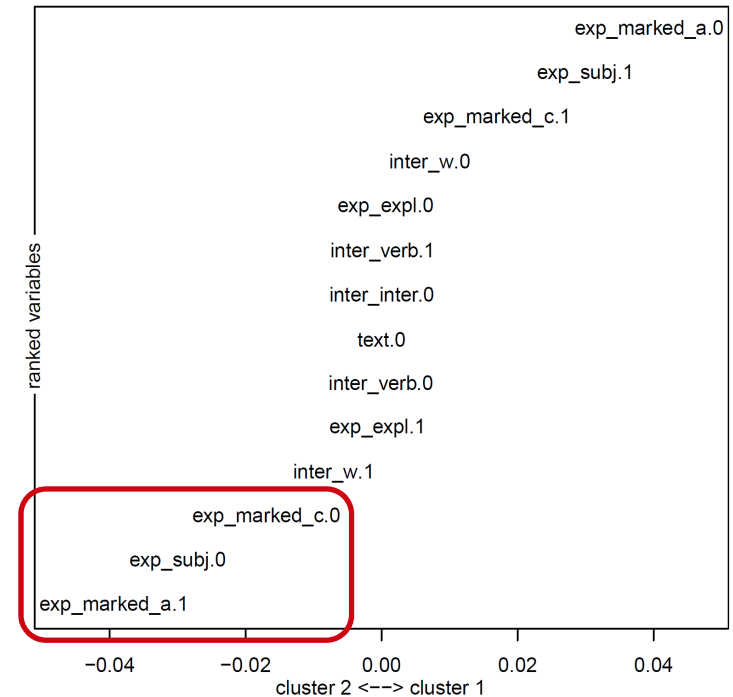
Snake plot of predictors  
after **first-element** clustering



- **simpler and more marked syntax** of Themes in popular registers (cluster 1):
  - pervasiveness of subject Themes
  - success of ‘exp\_marked\_c(omplement)’ Themes  
(By objective, I mean ...)

# Analysis of the data

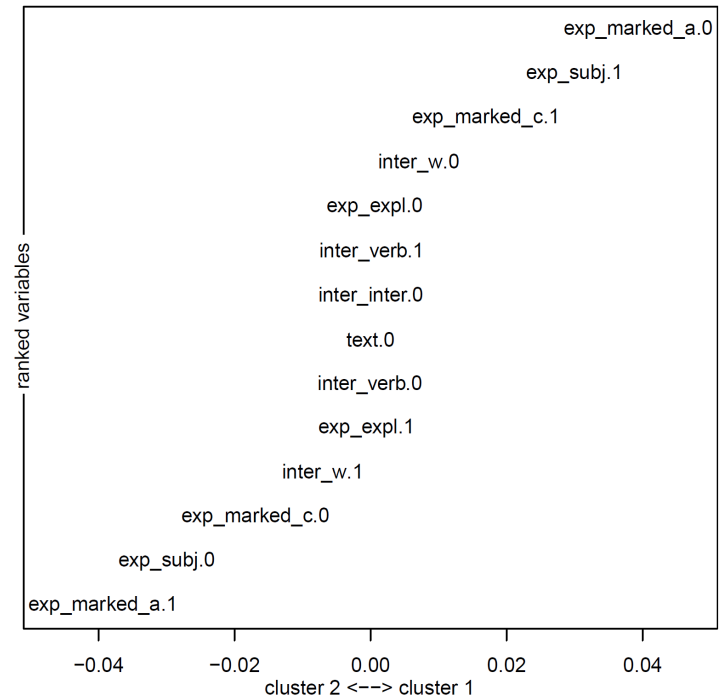
Snake plot of predictors  
after first-element clustering



- more elaborate and syntactically less marked Themes in learned registers (cluster 2):
  - lower proportion of subjects
  - much lower proportion of (marked) thematic complements
  - success of ‘exp\_marked\_a’ Themes (*A year ago former vice president Al Gore threw...*)

# Analysis of the data

Snake plot of predictors  
after **first-element** clustering



So...

- preference for subjects and marked (complement) Themes in popular registers
- higher ratios of subordination (and coordination) with a more unmarked clausal design in learned registers (given support by O'Donnell 1974, Kroll 1977)