Models - predictions - data: an (un)problematic relationship?
The example of Systemic Functional Linguistics (SFL)

Erich Steiner

GECCo-homepage at:
http://www.gecco.uni-saarland.de/GECCo/index.html

1. Introduction: SFL as empirical linguistics?

2. SFL as a theoretical background, predictions and data in terms of low-level linguistic categories

3. Information-theory and SFL as a theoretical background, predictions in terms of low-level linguistic categories

4. Increasing abstractness of annotations and creating an interface between theorizing and data

5. The abstractness of SFL-theorizing: problematic, but inherent?

Phenomenon under investigation:

***Explicitness* and other properties of translated texts**

(Hansen-Schirra, Neumann and Steiner 2012)

Features	Contrast (C1-n)	Phenomenon: Indicator	Explanation
Lexical Density (LD), Type-Token-Ratio (TTR), Parts-of-Speech proportionalities (PoS)	C1 (Reference Corpora ER vs. GR)	- Experiential explicitness: LD (E>G) - Strength of lexical cohesion other than repetition: TTR (G>E) - experiential and referential density: PoS (G>E in nominal orientation)	Language System
PoS proportionalities, reflecting "nominal orientation"	C2.2 (8 Registers within languages E and G)	- Experiential density: nominal orientation	Register, Language
		English: TOU > SHARE > WEB > ESSAY > INSTR > SPEECH > POPSCI> FICTION	
		German: TOU > WEB > SHARE > ESSAY > INSTR > SPEECH > POPSCI > FICTION	
LD, TTR, PoS (Nominal Orientation)		- referential and experiential density: spread of language-internal variation (G>E for TTR and nominal orientation; E>G for LD)	
	C2.1 (EO vs. GO by register, with ER/GR differences factored out)	- experiential and referential density: LD, TTR, PoS	Register
LD, TTR, PoS	C3 (Translations vs. originals within a language and within a register)	- Experiential explicitness: (LD) (ORI>TRANS) - lexical variation: TTR (ORI>TRANS) - referential density: nominality (ORI>TRANS, with exceptions)	Translation Process, De-Metaphorization

Table 1. Summary of shallow statistics used as operationalizations for 'explicitness' (Hansen-Schirra, Neumann and Steiner 2012:263f) Registers are TOU (Tourism), SHARE (Letters to our shareholders), WEB (Websides), ESSAY (Essay), INSTR (Instructions), SPEECH (Speeches), POPSCI (Popular Science), FICTION (Fiction), ORI (Originals), TRANS (Translations)

In additional studies based on CroCo, less shallow annotation types have been used, for example alignment patterns between words, phrases and grammatical functions (Hansen-Schirra et al 2012: ch.6-8), or some less shallow register features (Neumann 2014, Evert and Neumann 2017).

These studies may belong into the type of studies discussed in Chapter 4 below

Phenomenon under investigation:

(Lexical) Cohesion by Language, Register, Mode

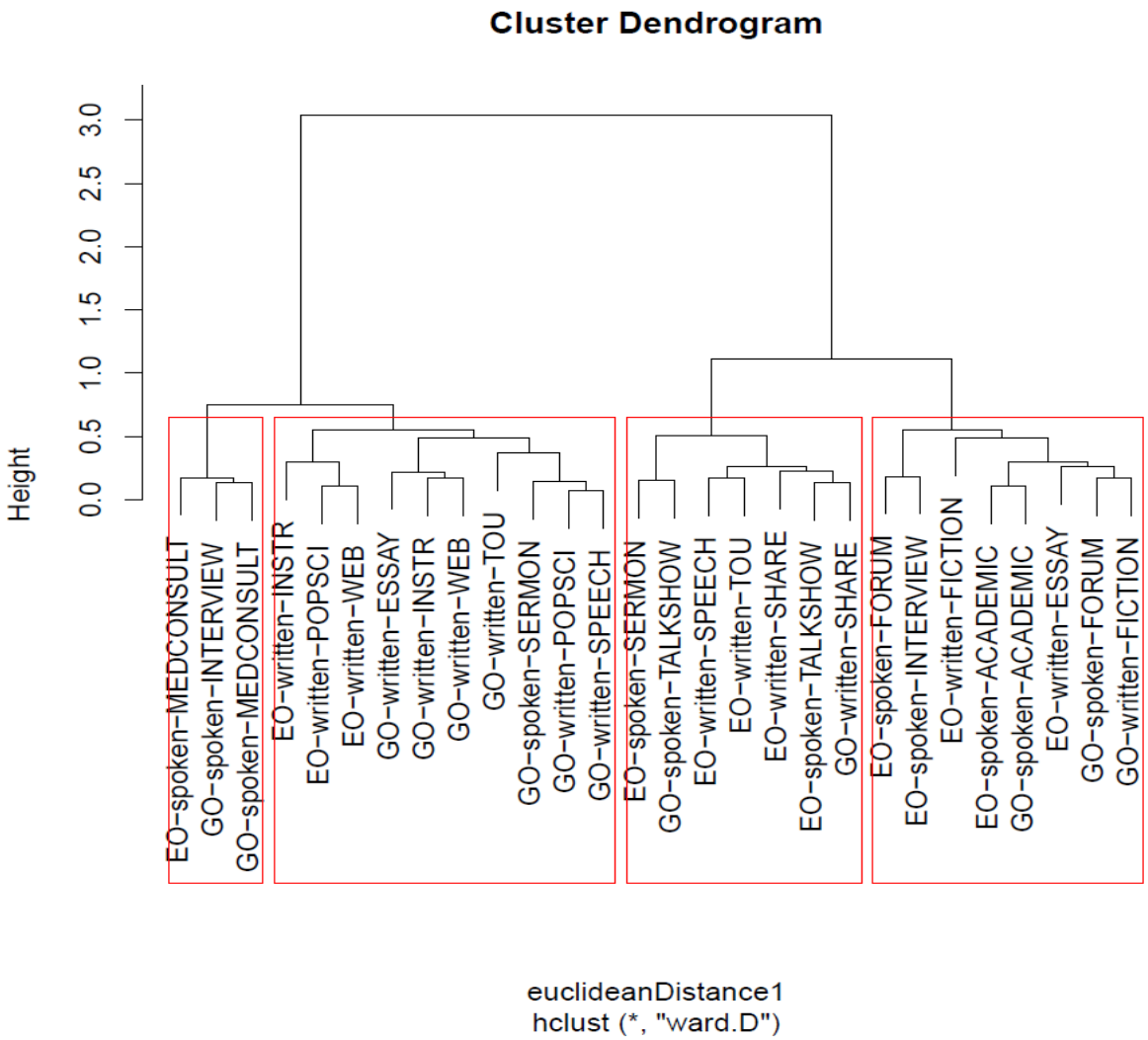
Summary of findings from Kunz et al 2017b

Shallow statistics on the lexis of corpus texts (Kunz et al 2017b)

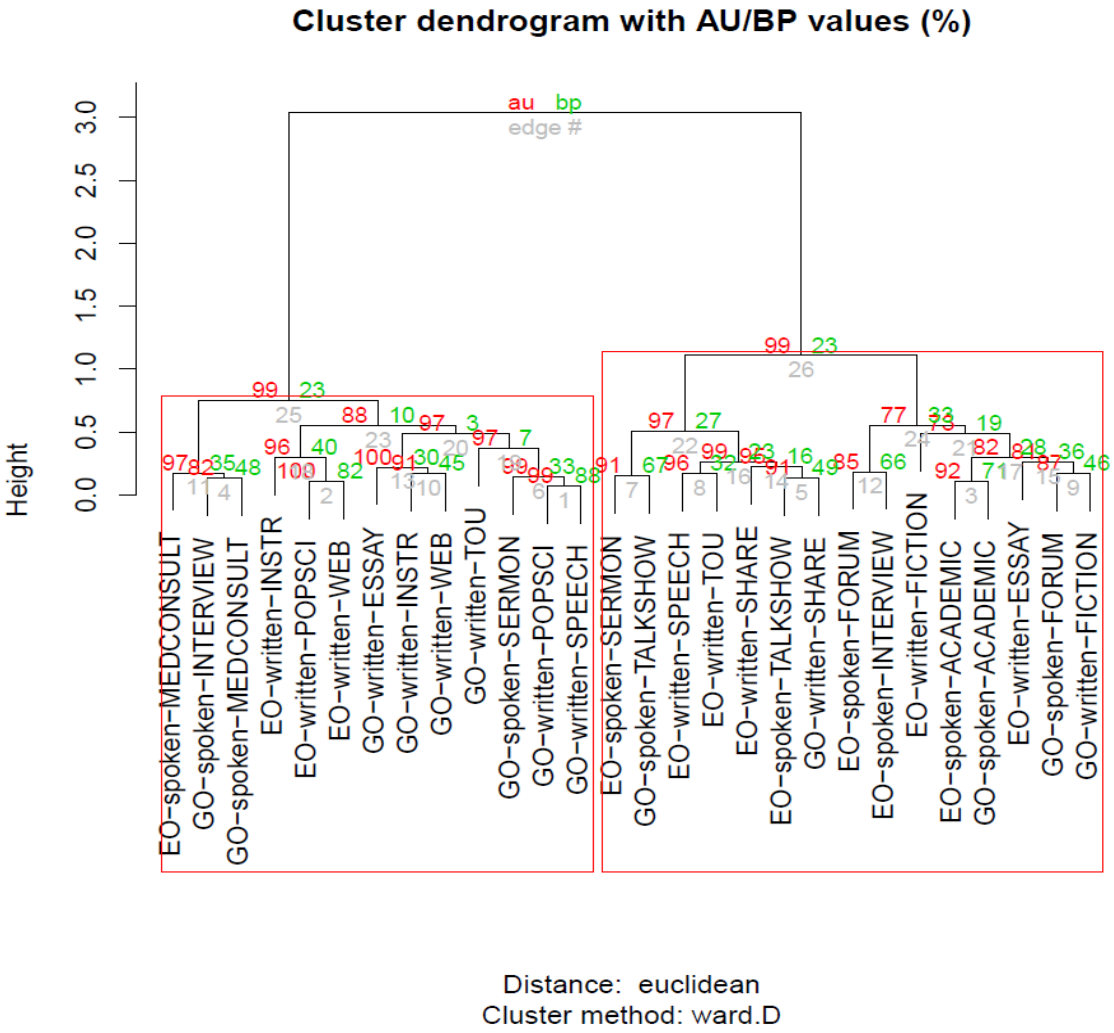
- the role of highly frequent content words in texts (MFCW);
- MFCW as a subset of the top-frequent content words of their register and of the language generally?;
- lexical density;
- standardized type-token-ratio;
- the role of Latinate words.

Interpreted as indicators of:

- semantic variability of relations within chains;
- cohesive strength of lexical elements;
- number and length of chains;
- degree of specification of lexis;
- degree of variation within texts, registers and modes in terms of these properties.

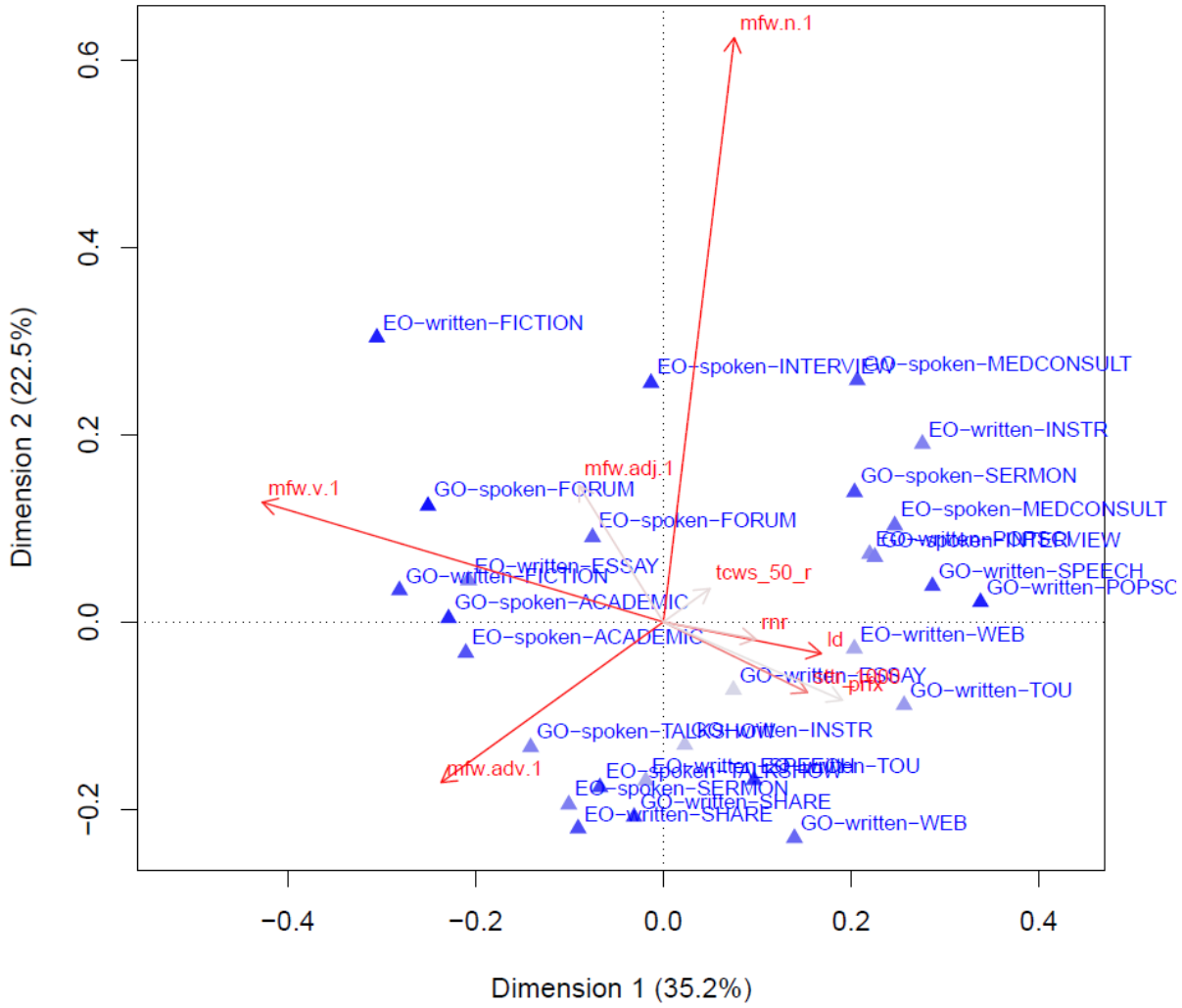


GECCo Lexical Cohesion *Hierarchical Cluster Analysis (HCA*);
shallow features
(LD, STTS, MFW, TCW, RW)



GECCo Lexical Cohesion (HCA);
shallow features

Approximately unbiased (AU);
Bootstrap Probability (BP)



GECCo Lexical Cohesion; shallow features

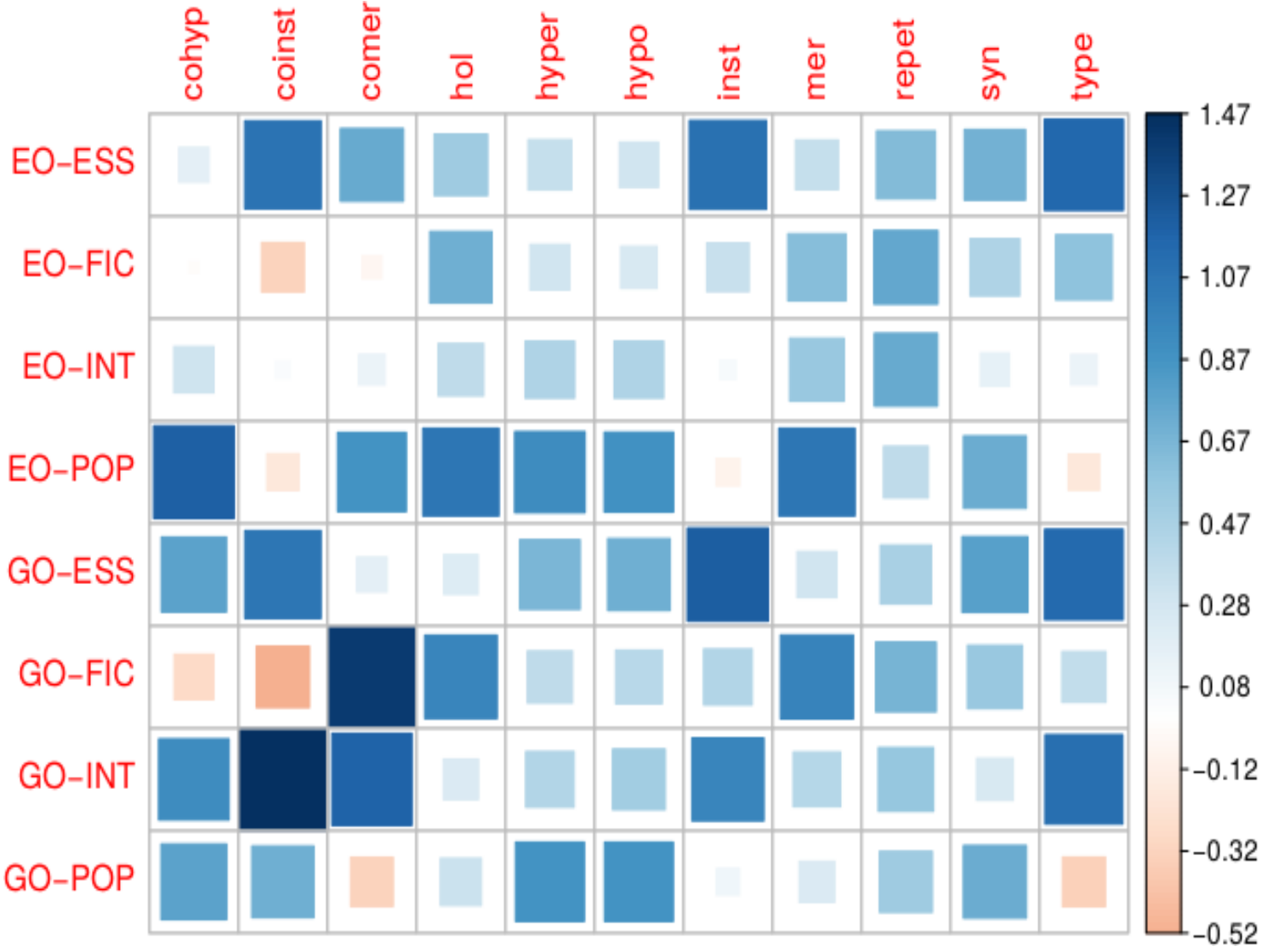
Correspondence Analysis (CA)

	MFCW	LD	STTR	TCW	LAT
language	E > G (v + n)	E > G	E < G	E > G	E > G
	E < G (adj + adv)				
mode	E: S > W (v + adv)	S < W	S < W	S > W	S < W
	G: S > W (v + adv)				
	S = W (adj)				
	S < W (n)				
register ranking	E = G (n)	E = G most written	E ≠ G	E ≠ G	E ≠ G
	E ≠ G (adj + v + adv)	E ≠ G most spoken			
variation	E = G	E > G	E > G	E = G	E > G

Summary of findings from Kunz et al 2017b; cells shaded grey disconfirm hypotheses

contextual parameter/	feature category	feature subcategory	
FIELD	term patterns	NN-of-NN, N-N, ADJ-N	
	verb classes	activity (e.g., *make, show*) aspectual (e.g., *start, end*) causative (e.g., *let, allow*) communication (e.g., *note, describe*) existence (e.g., *exist, remain*) mental (e.g., *see, know*) occurrence (e.g., *change, grow*)	
TENOR	modality	obligation/necessity (e.g., *must*) permission/possibility/ability (e.g., *can*) volition/prediction (e.g., *will*)	
MODE	theme	experiential theme (e.g, *The algorithm...*) interpersonal theme (e.g., *Interestingly...*) textual theme (e.g., *But...*)	
	conjunctive cohesive relations	additive (e.g., *and, furthermore*) adversative (e.g., *nonetheless, however*) causal (e.g., *thus, for this reason*) temporal (e.g., *then, at this point*)	
TECHNICALITY	type-token ratio lexical vs. function words	STTR no. of lexical PoS categories	
INFORMATION DENSITY	lexical density grammatical intricacy	lexical items per clause/sentence clauses per sentence wh-wrds per sentence sentence length	
ABSTRACTNESS	PoS distribution	no. of nominal vs. verbal categories	
CONVENTIONALIZATION	n-grams on PoS basis length of sections	2-to-6-grams overall/per section tokens per section	

Table 3: Linguistic features used in analysis (from Degaetano et al 2014: 1329)



The association table above shows association between lang/registers and sem. relations (likelihood ratio):
If ratio < 1 => log(ratio) < 0 (negative values) => red color
If ratio > 1 = > log(ratio) > 0 (positive values) => blue color

Figure 1: Association table for registers and semantic relations from Lapshinova-Koltunski et al. 2016

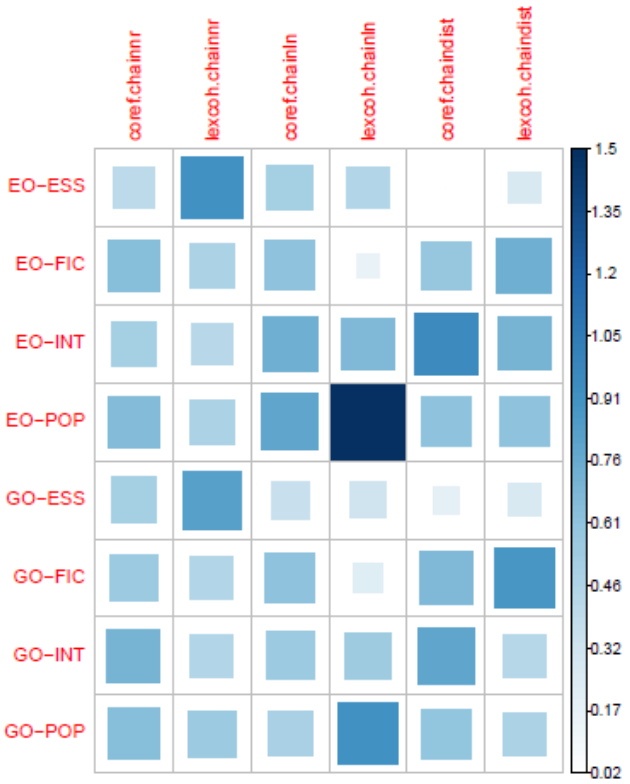


Figure 16: Associations between registers and chain propertiess based on Log Likelihood Ratio

GECCo Lexical Cohesion; semantic features

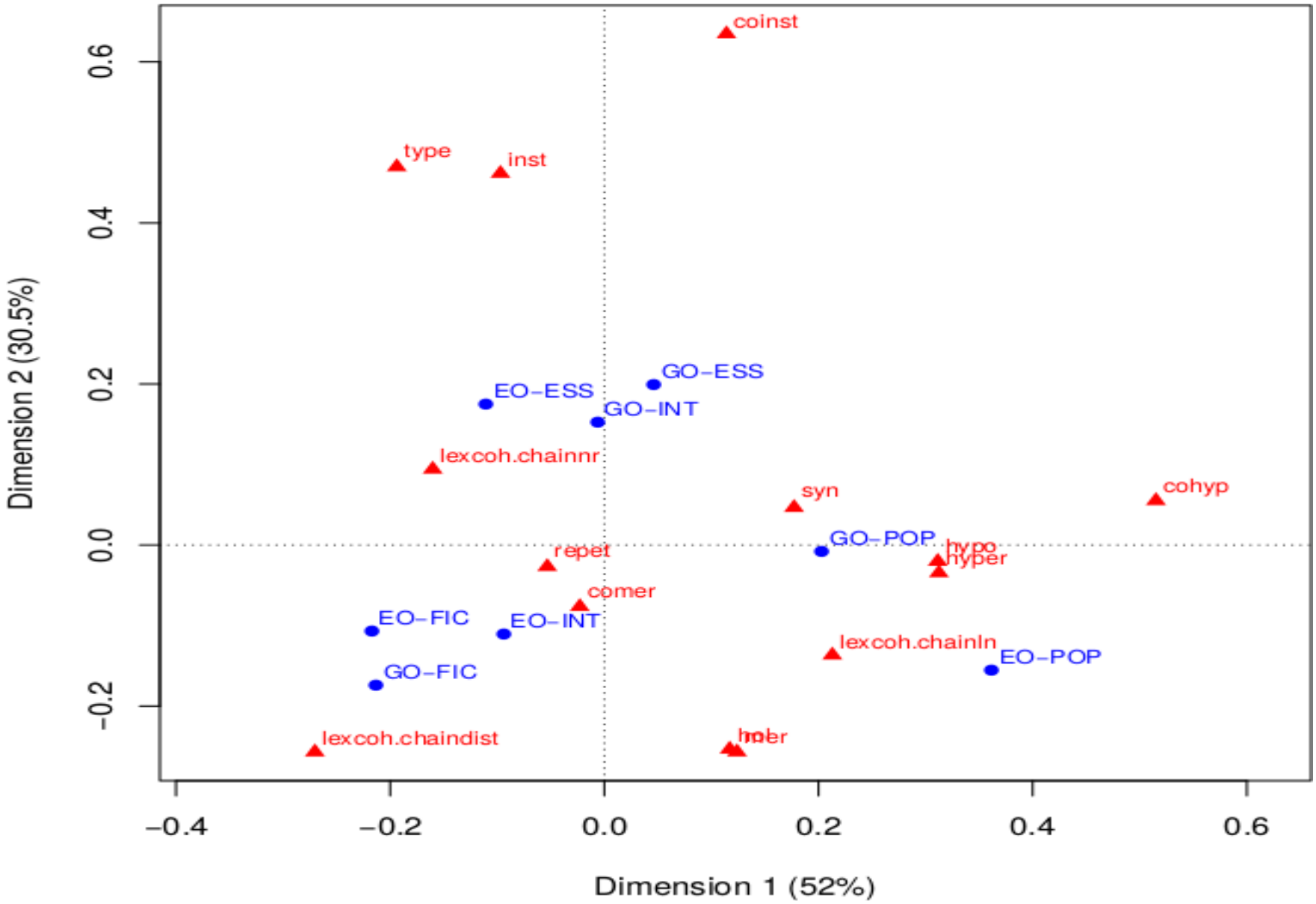


Figure 2: Correspondence Analysis for Registers, chain properties and Semantic Relations from Lapshinova-Koltunski et al 2016

General dimensions of cohesive relations

1 Degree of Cohesion:

What is the average proportion of cohesive devices per text (frequency)?

2 **Strength of Cohesive Relations:**

How explicit are cohesive devices?

How close are elements in cohesive chains?

How long are cohesive chains?

3 **Types of Meaning Relations:**

Which meaning relations are characteristic in languages/modes/registers (frequency)?

Which meaning relations are most distinctive of languages/modes/registers

4 Breadth of variation:

How much cohesive variation is there for languages, modes and registers