Domain and Genre differences

Martin Smrt
martin@martinsmrt.com

Contents

- Motivation
- Cross-register differences
  - Grammatical issues
  - Lexicographic issues
- Multidimensional differences
- Applications

Approaches to Corpus Building

- "Analyses of a large corpus could be generalized to the entire language."

vs.

- "Analyses must be based on a diversified corpus representing a wide range of registers."

Douglas Biber

Photo: http://english.pte.hu/hirek/376

Different domains and genres

- What domains and genres actually are?
- Biber uses the term register:
  - Using Register-diversified Corpora for General Language Studies
- What shall we account for?


TELEPHONE
TELEPHONE


The New York Times
OBAMA
RACIAL BARRIER FALLS IN DECISIVE VICTORY
Democrats in Congress Strengthen Grip

Genre

Social setting


tenor
field
mesolect
diatype
type
acrolect
register
genre
text
and
style
basilect
made with www.wordle.net

As Biber puts it …

- … there are systematic and important differences among the registers of English.
  - Grammatical differences
  - Lexical differences

Clause frequencies

- Mean frequencies of three dependent clause types (per 1,000 words) in four registers

	Relative Clauses	**Causative Adverbial Subordinate Clauses**	***that* Complement Clauses**
Press Reports	4.6	0.5	3.4
Official documents	8.6	0.1	1.6
Conversations	2.9	3.5	4.1
Prepared speeches	7.9	1.6	7.6

Grammatical issues

- Individual linguistic features are distributed differently across registers.
- The same (or similar) linguistic features can have different functions in different registers.
- This has important implications for probabilistic tagging and parsing techniques.

Probabilistic tagging and parsing

- We usually use probabilities to assign
  - Grammatical categories to ambiguous lexical items
  - Sequences of tags to ambiguous word groups

- Do grammatically ambiguous words have different distributions across registers?

Dictionaries based on Exposition and Fiction

- Total lexical entries in the Fiction dictionary: 22,043
- Total lexical entries in the Expository dictionary: 50,549

- Total words occurring in the Fiction dictionary only: 6,204
- Total words occurring in the Expository dictionary only: 31,476

- Words having probability differences of > 50%: 1,010
- Words having probability differences of > 30%: 980

Grammatical Category Probabilities

Word	Grammatical Category	Fiction %	Exposition %
admitted	Past tense	77	24
	Passives	17	67
	Perfects	6	0
	Adjectives	0	9
known	Passives	26	65
	Perfects	65	13
	Adjectives	6	15

Tag Sequence probabilities

First word	Second word	Fiction %	Exposition %
Singular noun	Preposition	23	31
	Singular noun	4	8
	Plural noun	1	4
	.	18	12
	,	15	11
Present tense verb	Indefinite article	12	18
	Adverb	13	9
	preposition	15	19

Lexicographic issues

- Differences between spoken and written discourse
- “There are striking differences across written registers in the use of … words.”

Written vs. Spoken

- Frequencies of ***X + certainty adjective***

	X + certain	**X + sure**	**X + definite**
Written text	259.0	234.0	34.9
Spoken text	292.5	426.9	19.4

- *Written text* - Longman/Lancaster Corpus
- *Spoken text* - London/Lund Corpus

Register differences

- In social science:
  - *certain* is quite common
  - *sure* is relatively rare
  - *definite* is common relative to its frequency in the whole written corpus.
- Fiction shows the opposite pattern:
  - *certain* is relatively rare
  - *sure* is relatively common
  - *definite is* quite rare.

Register differences (continued)

- Actual patterns of use are even more complex

- *certain* is commonly used to mark uncertainty rather than certainty
- Certainty is rarely expressed in social science at all
  - The most common collocations for *certain in social* science reflect a kind of vagueness (e.g., a certain kind of…, in certain cases…, there are certain indications that…)
  - These collocations are relatively rare in fiction

Multidimensional differences

- There are systematic patterns of variation among registers
- These patterns can be analyzed in terms of underlying dimensions of variation
- It is necessary to recognize the existence of a multidimensional space in order to capture the overall relations among registers.

Finding dimensions

- Texts were automatically tagged for lingustic features:
  - tense and aspect markers, place and time adverbials, pronouns and pro-verbs, nominal forms, prepositional phrases, adjectives, adverbs, lexical specificity, lexical classes (e.g., hedges, emphatics), modals, specialized verb classes, reduced forms and discontinuous structures, passives, stative forms, dependent clauses, coordination, and questions

Dimension evaluation

- Frequencies of features were counted and normalised
- Factor analysis was run to identify the major co-occurrence patterns among the features

Dimensions

1. Informational versus Involved Production
2. Narrative versus Nonnarrative Concerns
3. Elaborated versus Situation-Dependent Reference
4. Overt Expression of Persuasion
5. Abstract versus Nonabstract Style

Dimensions in detail

- Primary communicative functions
- Major co-occurring features
- Characteristic registers
  - See PDF
  - Table 7, Pages 231 – 232 (13-14/24)
  - Figure 1, Page 230 (12/24)

Further applications

- Automated prediction of registers

  1. Compute the distance of a text from a category
  2. Assign the nearest category

- Cross-Linguistic Comparisons
  - "the written style of English and French tended to be more similar in specialized technical texts than in general language texts" (Kittredge 1982, p. 108)

References

- Using Register-Diversified Corpora for General Language Studies

Douglas Biber
Northern Arizona University