“Up close and personal vs. birds-eye view” of discourse: a corpus study of perspective using Czech data

Václav Cvrček, Charles University, Czech Republic
vaclav.cvrcek@ff.cuni.cz

Masako Fidler, Brown University, USA
masako_fidler@brown.edu

Acknowledgments:
Charles University project Progres Q08 Czech National Corpus
Brown University Humanities Research Grant

ICLC
Nishinomiya, Japan
August 7, 2019
Langacker on perspective in grammar

”the optimal viewing arrangement”:
“the perceiver is so absorbed in the perceptual experience that he loses all awareness of self, and when the object perceived is well-delimited, wholly distinct from the perceiver” (Langacker 2001: 315-317)
Our interpretation of “the optimal viewing arrangement” on the discourse level

optimal viewing arrangement on the discourse level = discourse acuity, ability to discern the main point out of a text

It is maximized when

• the speaker is absorbed in the register (becomes increasingly perceptive of the register = becomes a speaker of this register)
• the reader (perceiver) can thus delimit the target content of the text (object of perception) from the other components that are the conventional parts of the type of text – register + genre (as defined by Biber and Conrad 2009)
First attempt: Keyword analysis

• Keyword analysis of a text (Target text, Ttxt) against the background of two different reference corpora (RefC) (cf. Scott 1999)
• Keywords: What the text is about (Scott and Tribble 2006: 59–60)

Example: Fidler and Cvrček 2015
• Socialist president’s New Year’s Addresses (NYA)
  • RefC1 – representative corpus of contemporary written Czech (SYN2010)
  • RefC2 - corpus of texts from socialist press before 1989
    RefC2 reflects readers’ increased ability to weed out the routine expressions typical of the period

• Modeling inexperienced reader vs. experienced reader of NYA (“time-sensitive” reading of text)
Major drawback of KWA

KW list puts everything into a mixed bag:

a. KWs connected with the main point of the text (connected to what the Ttxt makes distinct from other texts)

b. KWs specific to the genre (those words that occur in a specific position of the text)

c. KWs specific to the register (those words that consistently occur in the entire text)

Q: Can we really zoom in on the main content of the text when we cannot distinguish between a and (b, c)?
A more fine-grained analysis MDA-assisted KWA: Sibling-Aided Keyword-Register Analysis (SAKRA)

1. Take a diverse corpus (Koditex) with different types of text from spoken, web and written modes
2. Run MDA to identify the distribution of register-related devices (morphosyntactic, lexical etc. features)
3. Identify x number of texts that are closest to Ttxt in terms of text metadata and register features (“sibling” texts)
4. Create a RefC that maximally diverges from Ttxt (w.r.t. text metadata and register features)
5. Run KWA of Ttxt as well as the “sibling” texts (against the RefC created in 4)
6. Identify KWS shared by the Ttxt and the sibling texts (aka “key keywords”, Scott 1997): register-related KWS
7. Those KWS that are not shared are associated with the main point – the content that distinguishes the Txts from others
Preliminaries 1: Koditex corpus

- **guiding principles:** *diverse*, contemporary, **text length** control (Zasina et al. 2018)
  - diversified stratified sampling
  - 3334 text excerpts = **chunks** (not whole texts)
  - 10.8 M tokens (9 M words excl. punctuation)

- **annotation:** lemmas, tags, multi-word unit & named-entity recognition

- **3 modes** – written, spoken, web (...the ”extratextual“ perspective)
  - 8 divisions, 45 classes
  - ± 200,000 words per class
Preliminaries 2: Multi-dimensional analysis (MDA)

• Coined by D. Biber (Biber 1995; Biber & Conrad 2009)
  • systemic & functional variability (motivated by context & situation)

• Czech MDA (Cvrček et al. 2018): specific features of Slavic languages/Czech – rich morphology, free word order, diglossia...

• 122 functional features (see next slide) – interrelated choices → dimensions of variation → registers

• Distribution of features – similarity of texts w.r.t. features they employ („intratextual“ perspective)
Preliminaries 2: Features used in Czech MDA

Originally 140+ features, final list 122, e.g.: :

- **phonetics** – narrowing é > í, vowel breaking ý > ej, average word length...
- **morphology** – freq. of cases, numbers, moods, tenses...
- **derivation** – adjectives denoting similarity, verbal nouns, diminutives...
- **lexicon** – indefinite pronouns, reporting verbs, verbs of thinking, semantically bleached nouns...
- **pragmatics** – contact expressions, fillers, intensifiers, downtoners...
- **syntax** – types of attributes, clusters of POS, types of dependent clauses...
- **text/discourse** – questions, phraseology, word repetition...
Hypothesis: Shared KWs are likely to be register-related

- Register-related KWs recruit from evenly distributed words × topic-related KWs are specific, i.e. not dispersed in the corpus
  - Shared KWs (→register) = occurring (as a KW) in the Ttxt and at least 2 sibling texts
  - Unique KW (→topic) = occurring as a KW solely in the Ttxt

- Measure of dispersion – Deviation of proportions (DP; Gries 2008):
  - 0 ... ideally even distribution in a corpus
  - 1 ... maximally uneven distribution
Shared KWs and dispersion

- Statistically significant difference (Wilcoxon test)
- Rank-biserial correlation: 0.589
- Shared (register-related) KWs tend to be more dispersed than unique (topic-related) KWs, although the difference is not clear-cut
Pilot study: a text on pathological physiology

Analyzed sample: a chapter related to blood, blood diseases and their detection

Keywords harvested from the Ttxt and the RefC = topic keywords, keywords related to genre and to register

Compare the next two slides
• KWs shared by Ttxt and at least 3 sibling texts
• KWs unique to Ttxt
Pathological Physiology:
KWs shared by Ttxt and at least 3 sibling texts

1. množství ‘multitude’
2. tvořit ‘to constitute’
3. tzv ‘so-called’
4. počet ‘number’
5. kyselina ‘acid’
6. roztok ‘solution’
7. obsah ‘content’
8. zvýšený ‘raised’
9. ml ‘milliliter’
10. hodnota 'value'
Pathological Physiology:
KW\'s unique to Ttxt and DIN ≥ 99

KW\'s unique to Ttxt, 10 samples (DIN=100)

1. Haemoglobinometer
2. of Warfarin
3. NR (International Normalized Ratio, blood test)
4. leukocyte
5. Hb (Haemoglobin type)
6. calculating
7. methyl alcohol
8. single-layered
9. titration
10. Sahli\'s (haemoglobinometer)

KW\'s with (DIN ≥ 99), larger to smaller DIN, samples

1. hemolytic
2. anemia
3. blood cell
4. of sodium
5. pipette (liquid dispenser)
6. clotted
7. (blood) coating
8. tendency to clot
9. test tube
10. distilled

Blood, blood tests, blood diseases
Czech original available upon request
Implications of SAKRA (1): Modelling perception

• SAKRA models how the reader accesses P with increasing acuity
• The entire set of KWs in Ttxt via KWA against the background of RefC
  o perception of the reader with less experience with the type of text
• Remainder of KWs after removing KWs shared with sibling texts
  o perception of reader who acquired more experience with the register of this text type (able to see Ttxt within the context of similar texts)

See the next slide for visualization...
SAKRA: a dynamic model with an iterative process

Inexperienced reader (V1), becomes exposed to sibling texts (V2) and differentiates KWs as part of the discourse convention and KWs associated with P (the main points of text).
Implications of SAKRA (2): towards a systematic differentiation of genres

- SAKRA on each text → aggregated for each text class
- 4 most “target-content-driven” genres vs. 4 extremely “register-driven” genres
- Baseline – 60 % of topic-related KWs in a text

→ Allows a systematic differentiation of genres in terms of the two types of KWs
Implications of SAKRA (3): Language acquisition of academic writing

• Comprehension of specific types of academic writing is likely to depend on what type of text the student is most exposed to

→ Learner-centered strategies could be built on the results of SAKRA adjust register-related lexical items vs. lexical items associated with target content based on the student language exposure

A potential to develop more individualized strategy to improve comprehension of academic writing for both first and second-language learners
References


THANK YOU
Děkujeme za pozornost!
ありがとうございます。