Usage-based Approach to Discourse through Keyword Analysis
Masako Fidler (masako_fidler@brown.edu) and Václav Cvrček (vaclav.cvrcek@ff.cuni.cz)
Brown University, USA & Charles University in Prague, Czech Republic
ICL at the U. of Alberta, Edmonton, Canada (June 23-28, 2013)

Theoretical objective
Expansion of the usage-based approach (Hopper 1987, Bybee 2010) to the level of overarching themes in discourse.

The main idea
Contrast in word frequencies (target text vs. reference corpus) → Contrast in patterns of language use between the two

Special properties of the text (“keyness”) (Scott and Tribble 1996, Baker and Ellece 2011)

We may be able to unpack changing properties of the society in political discourse over time.

Materials
Target texts: fifteen New Year's Addresses in Czech from 1975 to 1989 by President of the totalitarian Czechoslovakia ~1,000 words each
Reference corpus: Periodicals from 1952-1969

Methods:
- “Needle in a Haystack” Corpus-based detection of thematically prominent word forms

“KWords” software (http://kwords.korpus.cz/) processes a text and provides a list of prominent words (keywords) - Keywords are NOT query words; they are outputs!

Main features of the software:
1. thematically prominent words culled and ranked by DICE’s co-efficient (KWs)
2. thematic links and their visualization (pic. on the right) (KW links)
3. collocation for each KW intended for detecting subtle shifts in topics in the same genre over time (hence the name Needle in a Haystack project)
4. comparison of KW rankings over time

Data for discussion
Recurring/frequent thematic word forms
Near-synonyms (keyness over time) (terms of address, names for the same political entity)
Isolated KWS: KWs that occur once or sporadically

Correspondences are found between the data and important historical events:

Graph 1: Recurrent/frequent KWs: types of 1st pers. (personal and possessive) pronounal forms that were picked as KWs (1975-1989)

Graph 2: Different use of apparently near-synonymous KWs
- “friends!” (prátelé) (“comrades!” (soudruzi!)) (1978, immediately after Charter 77) whereas “friends” (držitel) (“comrades!” (elsewhere))
- “comrades!” (and not “friends”) increase in keyness (the Eastern Bloc regains control)
- “co-citizens!” (spoluobčané) sometimes lines up with “comrades” (76, 79, 80, 81, 83, 84, 85, 86, and 88), but its keyness is a mirror image of “comrades!” (soudruzi) in 1977-1978 and 1982 & 1987 (politically more precarious years)

Graph 3: Recurrent KWs – the number of politically-related KW links to frequent KW “of people” (lidu)
- appearance of plurality: link to “of the (National) front” in the late 1970s vs. homeland (and society) increasing in the strength of links towards the 1980s (Charter 77 human rights movement, demonstrations in Poland)

Isolated KW appear in specific periods marked by domestic and international events.
1975: (with) the USSR, natural resources, prices, crisis, humanism (world-wide economic crisis, oil shock)

Discussion and conclusions
This corpus-based analysis of prominent word forms indicates subtle reactions of the Czechoslovak leadership to the major events in history.

Statistical contrast in language use (target text vs. ref. corpus) reflects topic shifts in texts these topic shifts in turn signal subtle societal changes over time.
The following aspects of corpus-based keyword analysis help shed light on topic and societal changes in text over time:
- Ranking of KWs (degrees of keyness)
- Diversity of syntactic functions to which keyness is attributed
- Changing set of KWs and KW links

References
Bybee, J. 2010. Language, usage and cognition