

NEH Application Cover Sheet

Digital Humanities Start-up Grants

PROJECT DIRECTOR

Dr. Masako Fidler
Professor of Slavic Languages
20 Manning Walk, Box E
Providence, RI 02912-9105
UNITED STATES

E-mail: masako_fidler@brown.edu
Phone(W): 401-863-3933
Phone(H):
Fax:

Field of Expertise: Languages: Slavic Languages

INSTITUTION

Brown University
Providence, RI UNITED STATES

APPLICATION INFORMATION

Title: *A Needle-in-a-Haystack Method: A New Prototype Corpus-Based Keyword Analysis Tool*

Grant Period: From 5/2014 to 10/2015

Field of Project: Languages: Slavic Languages

Description of Project: We will build new prototype open-source software for text analysis based on large and balanced corpora of Czech, to be accompanied by a set of well-articulated Best Practice Guidelines. We will derive a functional beta version for English as the first of the multi-language adaptations. The software will harvest word forms that serve as keys to major concepts of the text. It will complement and strengthen traditional qualitative text analysis by providing systematically extracted data from large volumes of text. Its newly implemented statistical algorithm has the potential to be highly sensitive to nuanced language use that may be difficult to capture by intuition alone, like finding "a needle in a haystack" (hence the name The Needle-in-a-Haystack Method). The Guidelines will provide a much-needed resource to promote data-driven research with minimized researcher bias, bridging raw data and interpretation in a way that will require little knowledge of statistics or corpus linguistics.

BUDGET

Outright Request	\$59,836.00	Cost Sharing	\$24,858.00
Matching Request		Total Budget	\$84,694.00
Total NEH	\$59,836.00		

GRANT ADMINISTRATOR

Cassandra Klinzing
Contract Administrator
164 Angell St
Box 1929
Providence, RI 02912-1929
UNITED STATES

E-mail: cassandra_klinzing@brown.edu
Phone(W): 401-863-1283
Fax:

Project title:
The Needle-in-a-Haystack Method:
A New Prototype Corpus-Based Keyword Analysis Tool

Funding sought: Level II (Grant period: May 2014-October 2015)

1. Table of Contents.....	page 1
2. List of Participants	page 2
3. Abstract and Statements of Innovation and Humanities Significance.....	page 3
4. Project Narrative – The Needle-in-a-Haystack Method (NHM).....	page 4 - 9
a. Enhancing the humanities through inovation.....	page 4 – 6
b. Environmental Scan.....	page 6 – 8
c. History and Duration of project.....	page 8
d. Work Plan.....	page 8 - 9
e. Staff.....	page 9
f. Final products and dissemination.....	page 9
5. Budget.....	page 10 – 13
a. (5 extra sheets: rate agreement)	
6. Biographies and detailed functions of the Project Team.....	page 14 - 15
7. Data Management.....	page 16
8. Letters of Commitment and Support.....	page 17 - 30
9. Appendices.....	page 31 – 40

2. List of Participants

Project Director:

Fidler, Masako, Department of Slavic Languages, Brown University

Co-PIs:

Cvrček, Václav, Institute of the Czech National Corpus, Charles University in Prague, Czech Republic

Fidler, Masako, Department of Slavic Languages, Brown University

Consultants on corpus linguistics

Baker, Paul, Lancaster University, UK

Davies, Mark, Brigham Young University, USA

Technical Support

Vondříčka, Pavel, Institute of the Czech National Corpus, Charles University in Prague, Czech Republic

Advisory Board at Brown U. on political science and further dissemination of the product of the project

Cook, Linda, Department of Political Science, Brown University, USA.

Melson, John, Sheridan Center for Teaching and Learning, Brown University, USA

Mylonas, Elli, Senior Digital Humanities Librarian, Research and Outreach Services, University Library,
Center for Digital Scholarship, Brown University, USA

Takayama, Kathy, Executive Director, Sheridan Center for Teaching and Learning, Brown University,
USA

Targan, David, Associate Dean of the College for Science, Brown University, USA

Assistant

Research Assistant (Brown University)

3. ABSTRACT AND STATEMENTS OF INNOVATION AND HUMANITIES SIGNIFICANCE

We will build new prototype open-source software for text analysis based on large and balanced corpora of Czech, to be accompanied by a set of well-articulated Best Practice Guidelines. We will derive a functional beta version for English as the first of the multi-language adaptations. The software will harvest word forms that serve as keys to major concepts of the text. It will complement and strengthen traditional qualitative text analysis by providing systematically extracted data from large volumes of text. Its newly implemented statistical algorithm has the potential to be highly sensitive to nuanced language use that may be difficult to capture by intuition alone, like finding "a needle in a haystack" (hence the name *The Needle-in-a-Haystack Method*). The Guidelines will provide a much-needed resource to promote data-driven research with minimized researcher bias, bridging raw data and interpretation in a way that will require little knowledge of statistics or corpus linguistics.

Statement of Innovation

NHM is an innovative text analysis prototype tool in Czech and English with the potential for multilanguage application. It implements a new algorithm for analysis, and provides a choice of statistical methods and informative visualization. Unlike other similar tools, NHM is being built specifically for users in the humanities with minimal technical knowledge; it identifies word forms characteristic of the targeted text and helps users bridge the obtained data and the interpretation of the text.

Significance for the humanities

NHM complements and strengthens the qualitative study of texts, which is at the core of the humanities. It responds to an increasing need among students and scholars in the humanities to make informed use of quantitative analytical methods. NHM will help humanists access diverse worldviews reflected in patterns of language use, examine how these worldviews impact interpretation of texts, and arrive at a representative and statistically significant picture of language, culture and society.

4. PROJECT NARRATIVE: THE NEEDLE-IN-A-HAYSTACK METHOD (NHM)

A. Enhancing the Humanities through Innovation

a. Objectives

Our ultimate goal is to develop a new, multilingual, open-source quantitative method (the Needle-in-a-Haystack Method [NHM]). NHM will consist of (a) data-driven software that harvests word forms with grammatical features ("keywords"), which serve as conceptual and semantic anchors for interpreting a text, and (b) a set of well-articulated Best Practice Guidelines that explain the relationship between the obtained raw data and interpretation. We seek funding (Level II) to build the NHM prototype for Czech and to derive a functional beta version for English.

The NHM software provides a new type of keyword analysis: (1) it is a tool that is accessible to users who do not have prior experience with corpus linguistics/statistics; (2) it offers a choice of statistical methods, including a new algorithm that has the potential to be highly sensitive to nuanced word usage, which may be difficult to capture by intuition alone, like finding "a needle in a haystack." The Best Practice Guidelines will ensure informed use of NHM. The products from and beyond the grant period will respond to an increasing need to inform and train students and scholars in the use of maximally objective quantitative analytical methods to arrive at a representative and statistically significant picture of language, culture, and society (Ensslin and Slocombe 2011).¹

b. What are keywords (KWs) and keyword analysis (KWAs)?

In this project we define KWs as word forms² that are obtained statistically (e.g., chi-square and log-likelihood tests) by comparing a target text(s) (T-txt(s)) with a larger corpus that reflects widespread linguistic patterns of the language (the reference corpus, RefC) (Scott 1996, Baker and Ellice 2011). KWs are thus word forms that occur in a text more frequently than expected by chance alone and are often closely connected to the special features and the genre of the text. Unlike many existing tools and approaches that might appear similar (**Section E**), what we mean by KWA is therefore a corpus-linguistic statistical method to extract word forms as *outputs*.

c. Innovation and intellectual goals

The project is being built on two valuable sets of materials to ensure methodologically sound KWA: (1) different subsets of language corpora from the Institute of the Czech National Corpus (CNC) at Charles University in Prague as RefCs (total: 1300 million words), and (2) T-txts for testing the limits of KWA: the New Year's Addresses delivered by the president of the totalitarian Czechoslovakia, Gustáv Husák, from 1975 to 1989 (NYAs)³.

The CNC easily surpasses the data coverage of other Slavic language corpora in terms of size and balance of genres and are comparable to the corpora of commonly taught languages such as English and German.⁴ Unlike other large corpora, which are often genre-specific and drawn primarily from the

¹ The Czech State TV expressed interest in using the beta version of NHM (<http://kwords.korpus.cz/>) to analyze the Czech presidential debates in 2012.

² NHM obtains word forms with grammatical features rather than lexicon.

³ Exploratory case studies using different types of text are being conducted separately (**Appendix 3, sections 2-3** for illustration)

⁴ The number of words in the major corpora in millions: The CNC (1300); the Russian National Corpus (149); the National Corpus of Polish (1500); the British National Corpus (100), the Bank of English (2500 (commercial)); the Corpus of Contemporary American English (450); the German Reference Corpus (5400); the Chinese Internet Corpus (280). Note that the Polish and German corpora contain very large percentages of journalistic texts (often from Internet sources). Measured by the proportion of the corpus size and the number of speakers, it is not an exaggeration to state that the CNC is the best of all the existing corpora.

Internet, the CNC covers a wide range of text types (journalism, fiction, science) and media, both printed and internet sources. Each text in the CNC is labeled as to its origin, author, text-type, genre, and year of publication.

T-txts used for testing NHM are about 30,000 words each⁵ and are sufficiently long to allow harvesting of KWs; the size ensures sufficient KWs and allows comparison between the results of qualitative analysis and the raw data from the software. These texts also provide a controlled environment to test the software's effectiveness, a necessary prerequisite to developing an objective method. Out of the three major factors that can influence the properties of T-txts (genre, topic, and author) over the 15-year span, two of them are maximally constant: the genre and the author. Furthermore, the texts pose maximum challenge to intuition-based analysis; they are an optimal material for testing the sensitivity of the method. These texts appear to be perfectly "flat," repeating the same socialist clichés and winding ritualistic sentences. The degree of ritualistic language nearly reaches the point of the absurd (cf. Václav Havel's *Garden Party* and *Memorandum*), even compared to the press in other East European countries in the 1970s and 1980s.⁶ Reading of these texts is easily clouded by the post-1989 politics as uninformative, but a small group of astute readers during socialism sensed on-going political development behind the scenes, which eventually led to the fall of the regime in 1989. If NHM can identify KWs that point to such nuanced interpretation by the latter group of readers, it will have a high potential to detect subtle nuances in other types of texts. (Details in [Section B](#))

Use of the RefC and T-txts, combined with the recently developed algorithm to rank KWs, is likely to deliver the most appropriate prototype by finding the most meaningful KWs even in texts that resist intuitive interpretation.

d. NHM and humanistic inquiry

The interplay between different RefCs and the T-txt allows for systematic study of how texts are framed by historical/cultural/linguistic constraints.

A well-balanced RefC takes its language data from various genres. It therefore represents the general pattern of language use, which roughly corresponds to native speakers' expectations of how the language is used (Hoey 2005). Language use, in turn, is closely connected with worldview (Underhill 2011)⁷. A well-balanced RefC can be used as a material that reliably reflects the worldview entertained by the majority of native speakers.

When the user drops a text into NHM, the software compares the word frequencies of the T-txt and the RefC, and yields those word forms (KWs) that are statistically prominent in the T-txt against the background of the RefC. These KWs help us identify concepts that readers find striking (and therefore informative). For example, a T-txt about the socialist political system likely contains KWs such as *masses*, *state* and *party* more frequently than usual; these words are like anchors that the reader uses to capture the main concepts in the text (Fidelius 1998).

Language, however, changes over time, and so does the worldview connected with it. RefCs drawn from different points of history are likely to reflect such a change; they can impact the set and/or ranking of KWs in one and the same T-txt and, consequently, its interpretation. NHM is thus expected to mirror changing perceptions of the same T-txt over time. As it is impossible for readers of the 21st century to

⁵ The NHM software has no problems processing large RefCs and small T-Txts. Our current issue to enhance NHM to be able to deal with longer novel-length T-txts with stability and speed.

⁶ The challenging aspect of the T-txt is optimal for capturing subtle textual nuances even compared to texts produced in other East European countries that were subject to censorship: the former USSR *did* reveal changes in the society more explicitly from time to time than its satellite nations because of its leading role in East Europe; in other neighboring states (e.g., Poland and Hungary) the media reports were more expressive about social changes than the former socialist Czechoslovakia. The East German press had moments to reveal its internal problems because of economic pressures from West Germany.

⁷ E.g., politeness expressions in a language that are used in accord with age, gender, and social status signal the importance of these social factors for the reader's worldview.

exactly internalize the language repertoire and the worldview of readers from the past, NHM is an indispensable tool that can approximate how readers likely read the T-txt in their historical context.

Imagine the routine use of words such as *socialism* and *workers* in socialist Czechoslovakia during the 1980s. These may catch the attention of young readers of the 2000s while use of *world peace* may not, although the latter may signal the former socialist government's concern over the stability of the Eastern bloc ([Appendix 3, 1b](#)). In contrast, the reader, who was in day-to-day contact with the communist discourse under the former regime, would have been more likely to filter out the noise (the routine clichés) and focus on *world peace* to suspect possible shifts in politics.

NHM, in short, allows different modes to frame the same T-txt. Below are some possible topics that the user could study by using different configurations of T-txts and RefCs (selected samples in [Appendix 3](#)):

- 1) Studying the prototypical interpretation of a text by readers of the 21st century. This allows non-native speakers to access the current native speaker's interpretation.
- 2) Reconstructing the effects of a text that was intended to be published in the past (but was not because of censorship or other circumstances).
- 3) Studying intertextual relationships, the relationship between the original text and its adaptation to a different genre, and possible motivations for revisions made by the author on the same text.
- 4) Reconstructing how the historical reader might have detected cultural and societal shifts over time (i.e., reconstructing how s/he "read between the lines")
- 5) Predicting socio-cultural shifts from a set of texts
- 6) Comparing the interpretation of a text in the past and now
- 7) Accounting for the longevity or readability of a literary text over time
- 8) Comparing the effects of a text in the original language and the effects of the translation

e. The English NHM and support by Consultants, the Advisory Board, and the institutions

The project includes creation of a beta version of the English NHM based on the Czech prototype. This will be the first step towards multilingual application of NHM. Experts in corpus linguistics will therefore play an important role in the project: two prominent scholars in corpus linguistics (Paul Baker and Mark Davies) will evaluate our interim progress and advise on the implementation of the NHM for English and beyond. As the testing material concerns politics, we also seek consultations on the interpretation of political texts from a specialist in political science (Linda Cook). Both Brown University and the ICNC have the infrastructure for public dissemination and preservation of the research outcomes and the necessary technical assistance. The collaborating faculty from both institutions will liaise with the Brown Advisory Board consisting of specialists engaged in cross-disciplinary digital literacy, literature, education, and political science.

B. Environmental Scan

KWA is used in both linguistic and literary research (cf. analysis of Shakespeare's plays in Scott and Tribble 2006). It is considered as a method to track the meaning of a text (Baker 2010) and widely used in different types of linguistic research: e.g., language and society (Baker 2011) and analysis of discourse (e.g. Duguid 2010), a study of the literary style of authors (Čermák and Cvrček 2009, 2010), and the language of a specific time period (Čermák, Cvrček and Schmedtová 2010). However, there has not been a definitive version of the method or a set of best practice guidelines. We intend to develop both.

There are several major differences between NHM and other software that provides similar functionality. NHM has the most rigorous and focused KWA functions: inclusion of the newest algorithm for ranking KWs (DICE), visualization of KW links, and adjustable contexts to search for KW links. NHM contains many functional features that do not require knowledge of corpus-/computational linguistics (no complicated automatic pre-processing of T-txts, no required installation on a local computer)⁸ ([Appendix 5a](#)).

⁸ Use of a large RefC is desirable for text analysis to be based on a sufficient amount of data. This,

There have been debates on how best to rank KWs (Mujis, 2010; Andrew et al. 2011; Rosenfeld & Penrod 2011). KW ranking is important since KWA studies tend to focus on the top 5-10% of KWs in interpreting a text (especially when harvested KWs exceed several hundreds). NHM provides a choice in ranking: one of them is a variation of the existing DICE's Coefficient (Cvrček and Fidler 2013), combining the best features of Kilgarrif (2009) and Gabrielatos and Marchi (2012). DICE has been shown to reflect KW strength (keyness, or the salience of KWs) better than carrying out log-likelihood tests, which are often used for KW ranking in existing software ([Appendix 5b](#)); NHM is the first of its kind to include DICE, and will continue to incorporate newer techniques as they emerge.

Furthermore, NHM will include a set of extensive Best Practice Guidelines that will bridge the method, the data from NHM, and the interpretation. The Guidelines should not be confused with the User's Manual (which will also be prepared). The former is based on the premise that the raw extracted data is only the beginning of rigorous interpretation of text (Phillips 2013). Technology allows systematic work with large language data and prevents over-dependence on individual subjective judgments in analysis, but it is ultimately the user who must arrive at the interpretation; this interpretation must be based on what the data can tell us. In other words, the user must know how much we can speculate, using the extracted data based on the T-txts, RefC, and pros and cons of each statistical method. The Guidelines will explain what each function does (e.g., reasons for excluding or including certain grammatical words from the analysis; motivations for expanding or limiting the contexts to search for KWlinks; possible uses of features such as the concordance, collocation and distribution plot ([Appendix 2](#)); advantages and disadvantages of using particular RefCs and T-txts rather than others). The relationship among the method, data, and interpretation will be explained with case studies (including "bad practice" samples, which we will intentionally create) and will include sample T-txts so that the user can replicate the results themselves. It is also noteworthy that the NHM Best Practice Guidelines are built on exploratory use of data in a highly inflected language (Czech). The mainstream corpus linguistic studies based on English tend to focus on lexicon rather than word forms with grammatical features, although word forms contain valuable information about varying degrees to which individuals play an active role in an event (e.g. actor, patient, recipient, experiencer, possessor, partner). The Guidelines will inform users of potential ways to interpret such grammatical features.

The material used for testing the effectiveness of NHM is in and of itself unique for a study of discourse. Linguistic study of political speeches in Czech from the totalitarian period using KWA has not been extensively carried out before. It is necessary to emphasize, however, that description of specific texts is not the goal but a path towards developing an appropriate method. Our project therefore should not be confused with Corpus-Assisted Discourse Study (CADs), for which the target of investigation is not the method but specific societal and cultural phenomena. In a similar vein, this project does not aim to describe how grammar evolves over time (e.g., Heine and Traugott 1991, Bybee et al. 1994, Hopper and Traugott 2003, Bybee 2010). Quantitative methods in political science (e.g., Benoit and Lowe 2012 and Däubler et al. 2012) specifically focus on political positioning of speakers. NHM is more holistic than this method, as it aims to investigate textual indicators of societal-cultural shifts as well as the trajectory in which the society might be headed.

NHM is distinct from other statistical text analysis methods. In contrast to the Wordhoard project, Google Ngram Viewer and Topic Modeling, NHM uses a well-defined and balanced RefC to contrast with the T-txt. The RefC used by the Google Ngram Viewer is drawn from the web, and its design is not known. The other two tools aim to find patterns within a set of texts ("a single corpus"); these patterns are presumed to be informative about these particular texts without a reference point. Unlike the data-driven NHM, the Wordhoard project and the Google Ngram Viewer assume the *preexistence* of KWs; the user inputs search words for queries because they are presumed to be important. Topic Modeling (Steyvers and Griffiths 2007, Blei and Lafferty 2009, Block 2010, the MALLET project [mallet.cs.umass.edu], Jockers

however, does not mean that NHM is completely dependent on preloaded RefCs. NHM does have the option for the user to input his/her own RefC. Such a flexibility is yet another strength of NHM ([Appendix 2](#)).

2013) has been widely used to identify groups of word forms that may represent important overarching themes within a corpus. While NHM produces links among KWs as outputs, however, Topic Modeling assumes that the user can predetermine the number of clusters to which KWs will belong; the user then places a semantic label on each cluster. Most importantly, none of these methods produces outputs as a result of comparison between two corpora. None of them deals with the interpretation of text from more than one viewpoint. The NHM is therefore quite unlike existing applications used for text analysis: it is more data-driven, is tested on challenging materials to ensure the sensitivity to nuanced extraction of KWs, has more functionality than others, and expects that a text might be read differently when it is subject to different frames.

C. History and Duration of the Project

This project was launched in 2012. Some tasks towards the final version of the Czech prototype have been completed. We built and uploaded the beta version of the Czech NHM in January 2013. The initial coding for the English NHM has begun⁹. Since then, NHM has undergone several revisions. At various stages of its development we have presented the Czech beta version at US and international conferences on Corpus-Assisted Discourse Studies (Cvrček and Fidler 2012), Corpus linguistics (Cvrček and Fidler 2013), and Slavic studies (Cvrček 2012, Fidler and Cvrček 2013ab). Using Husák's New Year's Addresses, we showed that NHM can reflect: different viewpoints from which to read the same T-txt; the relationship between language use and societal changes; and the advantage of KW ranking by DICE. We also had three small-scale planning meetings with members of the Advisory Board. The co-PIs are currently preparing articles on NHM and analysis of New Year's Addresses by Husák. Recently we asked a literature and digital humanities specialist to join the Advisory Board (Melson).

D. Work Plan

The table below represents the activities to be covered by the grant (for more details see [Appendix 4ab](#)). Abbreviations: Ast=Assistant, B=Baker, Co=Cook, Cv=Cvrček, D=Davies, F=Fidler, Me=Melson, MI=Mylonas, Tk=Takayama, Tr=Targan, V=Vondříčka, Mtg = Meeting

TASKS	Participants	Completion
Testing of the software		
Comparing data from the qualitative analysis and NHM data as we make improvements to the software; discussions for further enhancements based on the tests	F, Cv	5/2015
Application of NHM on other types of texts to explore the value of NHM to the humanities (information to be included in the Best Practice Guidelines)	F, Cv	9/2015
Best Practice Guidelines		
The 1 st draft of the Best Practice guidelines for NHM	F, Cv	9/2014
Public-domain release of the guidelines for Czech and English	F, Cv	9/2015
Completion of software development and dissemination of project		
1. Discussion on added functionality to the Czech NHM prototype	F, Cv	8/2014
2. Coding to improve robustness and stability of the software, adding functions to the Czech NHM as the prototype for other languages and completion of the English beta version, the first application based on the Czech prototype	Cv, V	6/2015
3. Public-domain release of the final version of the Czech prototype	Cv, V	6/2015
4. Public-domain release of the NHM beta version for English	Cv, V	8/2015
5. Completion of the project website (Brown research website)	F, Cv, Ast	9/2015
6. Conference presentations (Section F)	Cv, F	10/2014 & 7/2015
7. Submission of journal articles on NHM	F, Cv	11/2014, 8/2015
Interim and final evaluation and future-planning meetings¹⁰		
1: Evaluation and implementation of NHM for English and beyond	B, Cv, D, F (Providence)	2/2015

⁹ This NHM will be connected to word-frequency lists from the Corpus of Contemporary American English.

¹⁰ The meeting in Prague will include hands-on discussions of NHM as well as extensive planning with the ICNC staff on the Norway grant.

2: Presentation and consultation with the Advisory board: use of technology in the humanities and social sciences for students and faculty. **Cv, Co, F, Me, MI, T, Tk (Providence) 4-5/2015**

3: Looking forward (funding, scope of NHM) **B, Cv, D, F (Prague) 10/2015**

a. Potential risks

Creating beta versions of additional languages may encounter unforeseeable technical and interpretive difficulties. For the future we may also need to consider how to coordinate other multilingual versions of NHM with the necessary reference corpora. We built in consultations with experts in corpus linguistics who are knowledgeable in English, Portuguese and Spanish¹¹, and are committed to taking risks and tackling problems as they arrive during the next phase of the project. Issues in interpreting political texts and strategies to improve accessibility of NHM to a wide audience will be addressed by consulting experts in political science, literature, digital humanities, and education on the Advisory Board.

b. Part of the project to be funded by subsequent grants:

We will seek further funding (e.g., NEH Digital Humanities Implementation Grant, ACLS Collaborative Research Fellowship, the Norway Grant¹²) to finalize the English NHM and expand our scope beyond Czech and English.

E. Staff

Name and title	Time committed to this project	Funding
Fidler , Project Director and Co-PI: 10% of academic year time		cost shared
Cvrček , Co-PI: approximately 6 hours/week for 18 months (in Prague and Providence) as a Visiting Assistant Professor at Brown (pending NEH funding)		requested
Vondříčka , Technology Consultant: 5 hours/week for 9 months		requested
Consultants on corpus linguistics:		
Baker and Davies : two full-day meetings (Prague and Providence)		requested
Advisory Board members:		
Cook, Melson, Mylonas, Takayama, and Targan : one full day meeting		in kind
Cook : bi-weekly consultations with Co-PIs over Skype and/or in person		in kind
Mylonas : 1-2 times/month consultations over Skype and/or in person		in kind
Research Assistant	10 hours/week for 5 months (1-4/2015, 9/2015)	requested

F. Final Products and Dissemination

There are several venues where we plan to disseminate NHM.

- Open-source Czech NHM (prototype software) and the beta version of the English NHM and the Best Practice Guidelines (code, data stored at the Brown research website and ICNC website).
- Test use of the Czech NHM and English NHM in the classroom (Cvrček's course on corpus linguistics, Fidler's course on sociolinguistics and a new course on discourse analysis (in preparation))
- paper presentations (American Corpus Linguistics Conference, 2014, Corpus Linguistics Conference, 2015)
- article submission to refereed journals (e.g. *Corpora*, *Journal of Digital Humanities*)
- dissemination of the final results via appropriate lists (e.g. linguistlist, corpora list).
- public presentations on venues related to training and professional development of students and faculty, including those held at the Sheridan Center for Teaching and Learning and the Brown Library
- the white paper describing the role of the NHM as a significant step towards developing a multilingual quantitative method to interpret texts for scholars and students in the humanities and social sciences.

¹¹ We are also informally starting conversations with Serge Sharoff from University of Leeds to plan the Russian NHM.

¹² www.naep.cz/index.php?a=view-project-folder&project_folder_id=103&view_type_code=program&#record_5684



Budget Form

Applicant Institution: *Brown University*

Project Director: *Masako Fidler*

Project Grant Period: *05/01/14-10/31/15*

[click for Budget Instructions](#)

	Computational Details/Notes	(notes)	Year 1	(notes)	Year 2	Project Total
			05/01/2014- 04/30/15		05/01/2015- 10/31/2015	
1. Salaries & Wages						
Masako Fidler (Project Director and Co-PI), Professor of Slavic Languages, Brown U.	Year 1 salary- 10% May 2014; Sept, 2014-May, 2015; Sept-Oct,2015 Cost shared		\$12,220		\$4,089	\$16,309
Vaclav Cvrcek (co-PI), Assistant Professor of Research, Brown U.	Year 1 salary- approximately 6 hours per week Year 2 salary approximately 6 hours per week		\$10,000		\$4,500	\$14,500
TBD (grad or undergrad) student	5 months; 10 hours/week; \$15 hour	%	\$3,248	%		\$3,248
2. Fringe Benefits						
Fidler, Masako	Fringe Benefits - 30.50% effective 5/1/14 - 6/30/14, 31% 7/1/14 - 10/31/15 Cost shared		\$3,782		\$1,268	\$5,049
Cvrcek, Vaclav (co-PI)	Fringe Benefits - 8% effective 5/1/14 - 10/31/15		\$800		\$360	\$1,160
						\$0
3. Consultant Fees						
						\$0

Davies, Mark	Providence fall 2014, Prague fall 2015 \$800 each meeting		\$800		\$800	\$1,600
Baker, Paul	Providence fall 2014, Prague fall 2015 \$800 each meeting		\$800		\$800	\$1,600
Vondricka, Pavel (technical consultant)	9 months; 9 hours/week		\$7,500			\$7,500
						\$0
						\$0
4. Travel						\$0
Fidler, Masako	planning meeting at the NEH offices in Washington DC (Wash DC airfare)		\$300			\$300
Fidler, Masako per diem	Washington per diem (1 day)		\$273			\$273
Fidler, Masako	10/2014 American Corpus Linguistics Conference (airfare Providence-Los Angeles and per diem), cost- shared (Faculty Travel Fund) \$1,500 (transportation and 2 days per diem)		\$1,500		\$0	\$1,500
Fidler, Masako	7/2015 Corpus Linguistics Conference, UK (airfare Boston-Lancaster and 2 days per diem), cost- shared (Faculty Travel Fund + International Travel Fund) \$2,000				\$2,000	\$2,000

Cvrcek, Vaclav	10/2014 American Corpus Linguistics Conference (airfare Providence-Los Angeles and 2 days per diem) \$1500		\$1,500			\$1,500
Cvrcek, Vaclav	Conference, UK (transportation) from Prague		\$429			\$429
Cvrcek, Vaclav per diem	7/2015 Corpus linguistics, UK airfare and per diem (2 days)				\$800	\$800
Cvrcek, Vaclav conference	Corpus Ling. Conference fee		\$400			\$400
Davies, Mark	Utah - Prague (airfare) meeting in Prague				\$1,485	\$1,485
Davies, Mark per diem	Prague per diem (4 days)				\$250	\$250
Davies, Mark	Utah - Providence (airfare) meeting in Providence		\$493			\$493
Baker, Paul	Lancaster - Providence (airfare), meeting in Providence		\$800			\$800
Baker, Paul	Lancaster - Prague (airfare), meeting in Prague				\$235	\$235
Baker, Paul per diem	Prague per diem (4 days)				\$250	\$250
						\$0
						\$0
5. Supplies & Materials						\$0
						\$0
						\$0
6. Services						\$0
						\$0
						\$0

7. Other Costs						\$0
						\$0
						\$0
						\$0
						\$0
						\$0
8. Total Direct Costs	Per Year		\$44,844		\$16,837	\$61,680
9. Total Indirect Costs						
62.5%MTDC, DHHS Agreement dated 1/7/13	Per Year		\$17,088		\$5,925	\$23,013
10. Total Project Costs	(Direct and Indirect costs for entire project)					\$84,693
11. Project Funding	a. Requested from NEH					
				Outright:		\$59,836
				Federal Matching Funds:		\$0
				TOTAL REQUESTED FROM NEH:		\$59,836
	b. Cost Sharing					
				Applicant's Contributions:		\$24,858
				Third-Party Contributions:		\$0
				Project Income:		\$0
				Other Federal Agencies:		\$0
				TOTAL COST SHARING:		\$24,858
12. Total Project Funding						\$84,693

Total Project Costs must be equal to Total Project Funding ----> (\$84,693 ?)

Third-Party Contributions must be
greater than or equal to Requested Federal Matching Funds ----> (

ORIGINAL

COLLEGES AND UNIVERSITIES RATE AGREEMENT

EIN: 05-0258809

DATE:01/07/2013

ORGANIZATION:

FILING REF.: The preceding
agreement was dated
01/30/2012

Brown University
164 Angell Street, Box J
Providence, RI 02912-

The rates approved in this agreement are for use on grants, contracts and other agreements with the Federal Government, subject to the conditions in Section III.

SECTION I: INDIRECT COST RATES

RATE TYPES: FIXED FINAL PROV. (PROVISIONAL) PRED. (PREDETERMINED)

EFFECTIVE PERIOD

<u>TYPE</u>	<u>FROM</u>	<u>TO</u>	<u>RATE(%)</u>	<u>LOCATION</u>	<u>APPLICABLE TO</u>
PRED.	07/01/2012	06/30/2013	62.00	On-Campus	Research
PRED.	07/01/2013	06/30/2015	62.50	On-Campus	Research
PRED.	07/01/2012	06/30/2015	23.00	On-Campus	Other Sponsored Activities
PRED.	07/01/2012	06/30/2015	26.00	Off-Campus	Research
PROV.	07/01/2015	Until Amended			Use the same rates and conditions as those cited for fiscal year ending June 30, 2015.

ORGANIZATION: Brown University

AGREEMENT DATE: 1/7/2013

***BASE**

Modified total direct costs, consisting of all salaries and wages, fringe benefits, materials, supplies, services, travel and subgrants and subcontracts up to the first \$25,000 of each subgrant or subcontract (regardless of the period covered by the subgrant or subcontract). Modified total direct costs shall exclude equipment, capital expenditures, charges for patient care, student tuition remission, rental costs of off-site facilities, scholarships, and fellowships as well as the portion of each subgrant and subcontract in excess of \$25,000.

ORGANIZATION: Brown University

AGREEMENT DATE: 1/7/2013

SECTION I: FRINGE BENEFIT RATES**

<u>TYPE</u>	<u>FROM</u>	<u>TO</u>	<u>RATE(%)</u>	<u>LOCATION</u>	<u>APPLICABLE TO</u>
FIXED ✓	7/1/2012	6/30/2013	31.00	All	Full-Time Employees
FIXED	7/1/2012	6/30/2013	8.00	All	Part-Time Employees
FIXED	7/1/2013	6/30/2014	30.50	All	Full-Time Employees
FIXED	7/1/2013	6/30/2014	8.00	All	Part-Time Employees
PROV.	7/1/2014	Until amended	31.00	All	Full-Time Employees
PROV.	7/1/2014	Until amended	8.00	All	Part-Time Employees

** DESCRIPTION OF FRINGE BENEFITS RATE BASE:
Salaries and wages.

ORGANIZATION: Brown University

AGREEMENT DATE: 1/7/2013

SECTION II: SPECIAL REMARKS

TREATMENT OF FRINGE BENEFITS:

The fringe benefits are charged using the rate(s) listed in the Fringe Benefits Section of this Agreement. The fringe benefits included in the rate(s) are listed below.

TREATMENT OF PAID ABSENCES

Vacation, holiday, sick leave pay and other paid absences are included in salaries and wages and are claimed on grants, contracts and other agreements as part of the normal cost for salaries and wages. Separate claims are not made for the cost of these paid absences.

1. The rates in this Agreement have been negotiated to reflect the administrative cap provisions to OMB Circular A-21 published by the Office of Management and Budget on May 8, 1996. No rate affecting the institution's fiscal periods beginning on or after October 1, 1991 contains total administrative cost components in excess of that 26 percent cap.

2. For all activities performed in facilities not owned by the organization and to which rent is directly allocated to the project, the off-site rate will apply. Grants or contracts will not be subject to more than one indirect cost rate. If more than 50% of the project is performed off-site, the off-site rate will apply to the entire project.

3. Fringe Benefits: Annual fringe benefit rates consisting of retirement expenses, Social Security Taxes, Tuition Remission, Education Assistance, Dental Plan, Worker's Compensation, Health Insurance, Long-Term Disability, Group Life Insurance, Unemployment Insurance, Sabbaticals, and Benefits Administration Expenses.

4. Equipment means an article of nonexpendable tangible personal property having a useful life of more than one year and an acquisition cost of \$5,000 or more per unit.

5. Effective 7/1/10 the Faculty and Administrative Staff / Weekly Staff fringe benefit rates have been combined.

* Effective 7/1/99 tuition support for dependents of Brown University employees will no longer be an allowable fringe benefit expense in the approved rates.

This rate agreement updates fringe benefit rates only.

ORGANIZATION: Brown University
AGREEMENT DATE: 1/7/2013

SECTION III: GENERAL

A. LIMITATIONS

The rates in this Agreement are subject to any statutory or administrative limitations and apply to a given grant, contract or other agreement only to the extent that funds are available. Acceptance of the rates is subject to the following conditions: (1) Only costs incurred by the organization were included in its facilities and administrative cost pools as finally accepted; such costs are legal obligations of the organization and are allowable under the governing cost principles; (2) The same costs that have been treated as facilities and administrative costs are not claimed as direct costs; (3) Similar types of costs have been accorded consistent accounting treatment; and (4) The information provided by the organization which was used to establish the rates is not later found to be materially incomplete or inaccurate by the Federal Government. In such situations the rate(s) would be subject to renegotiation at the discretion of the Federal Government.

B. ACCOUNTING CHANGES

This Agreement is based on the accounting system purported by the organization to be in effect during the Agreement period. Changes to the method of accounting for costs which affect the amount of reimbursement resulting from the use of this Agreement require prior approval of the authorized representative of the cognizant agency. Such changes include, but are not limited to, changes in the charging of a particular type of cost from facilities and administrative to direct. Failure to obtain approval may result in cost disallowances.

C. FIXED RATES

If a fixed rate is in this Agreement, it is based on an estimate of the costs for the period covered by the rate. When the actual costs for this period are determined, an adjustment will be made to a rate of a future year(s) to compensate for the difference between the costs used to establish the fixed rate and actual costs.

D. USE BY OTHER FEDERAL AGENCIES

The rates in this Agreement were approved in accordance with the authority in Office of Management and Budget Circular A-21, and should be applied to grants, contracts and other agreements covered by this Circular, subject to any limitations in A above. The organization may provide copies of the Agreement to other Federal Agencies to give them early notification of the Agreement.

E. OTHER

If any Federal contract, grant or other agreement is reimbursing facilities and administrative costs by a means other than the approved rate(s) in this Agreement, the organization should (1) credit such costs to the affected programs, and (2) apply the approved rate(s) to the appropriate base to identify the proper amount of facilities and administrative costs allocable to these programs.

BY THE INSTITUTION:

Brown University

(INSTITUTION)

(SIGNATURE)

(NAME)

(TITLE)

(DATE)

ON BEHALF OF THE FEDERAL GOVERNMENT:

DEPARTMENT OF HEALTH AND HUMAN SERVICES

(AGENCY)

(SIGNATURE)

Darryl W. Mayes

(NAME)

Director, Northeastern Field Office

(TITLE)

1/7/2013

(DATE) 0944

HHS REPRESENTATIVE:

Michael Stanco

Telephone:

(212) 264-2069

6. BIOGRAPHIES AND DETAILED FUNCTIONS OF THE PROJECT TEAM

- Project Director and Co-PI

Masako Fidler is a professor at the Department of Slavic Languages, Brown University. Her specialization is discourse-cognitive linguistics with focus on Czech. As Project Director she will be responsible for managing the progress of the project. She will participate in the qualitative analysis of texts, the interpretation of the quantitative analysis and the discussions of the necessary features in the new version of the software; supervise the Assistant; and coordinate the meetings with corpus linguistics Consultants and the Advisory Board. Brown fully supports the academic year salary of the PI, but makes no specific commitment of her time or salary to this particular research project. As Co-PI of this project Fidler will be responsible for the output of the final products with Cvrček.

- Co-PI

Václav Cvrček, Chair of the Institute of the Czech National Corpus, Charles University in Prague, Czech Republic. Cvrček's appointment as a Visiting Assistant Professor in research (with salary contingent on funding) is currently being processed at Brown. His specialization is corpus linguistics with focus on Czech, for which he has numerous publications. Cvrček will prepare and refine the NHM software in consultation with Fidler and Vondříčka, and participate in the testing of the software on T-txts. He will work for approximately 18 months in Prague and Providence. As Co-PI of this project Cvrček will supervise the Assistant, and will be responsible for the output of the final products with Fidler.

- Consultants on corpus linguistics in English

Paul Baker, Professor, Lancaster University, UK. Baker is a prominent scholar in corpus linguistics, sociolinguistics, and corpus-assisted discourse analysis. His research interests include corpus linguistics, language and gender/sexual identities, and critical discourse analysis. Baker has published extensively, among other topics, on keyword analysis in English.

Mark Davies, Professor, Brigham Young University. Davies is likewise a prominent scholar in corpus linguistics and has expertise in building large American English language corpora and corpora of other languages, for which he has been funded by the NEH.

Both Consultants will critique and evaluate the progress of the project, advise on the future multilingual implementation of the NHM in Providence and Prague. They will discuss the methodology, including on-site experimentation with coding with Vondříčka, and participate in the discussions on future funding with potential participants at the ICNC in Prague.

- Technology Support

Pavel Vondříčka, Institute of the Czech National Corpus, Charles University in Prague, has been engaged in software development and maintenance of the Institute server. He has helped build the beta version of the NHM software as well as other applications for CNC users. Vondříčka will provide technical consultation, 5 hours/week for 9 months, in developing the prototype NHM for Czech and a beta version for English.

- Advisory Board

The following Advisory Board members from Brown University will dedicate their time at a one-day meeting with the Co-PIs in Providence.

Linda J. Cook (http://research.brown.edu/myresearch/Linda_Cook) Cook received her Ph.D. from Columbia University in 1985. She is currently a professor in the political science department at Brown University, leader of the 2013-14 Pembroke Seminar, "Socialism and Post-Socialism" at Brown, and an associate of the Davis Center for Russian and Eurasian Studies at Harvard University. She has authored "The Soviet Social Contract and Why it Failed" (Harvard, 1993), "Postcommunist Welfare States: Reform

Politics in Russia and Eastern Europe," (Cornell, 2007) and numerous other publications. Her main research interests are in the comparative politics of the Russian Federation, East-Central Europe and Eurasia, electoral-authoritarian regimes cross-regionally; comparative welfare states, labor, gender, and representation. She is currently working on a book, "Russia's Fragmented Welfare State," and a cross-regional study of the politics of welfare in electoral-authoritarian regimes. Cook will be advisor to the qualitative analysis of the T-txts used for testing of NHM with her expertise in East Europe.

Elli Mylonas (library.brown.edu/cds/about/staff/elli-mylonas), Senior Digital Humanities Librarian in the Data Curation subgroup of Research and Outreach Services at Brown University. Mylonas' areas of expertise lie in hypertext, XML, structured text, and digital rhetoric. As senior digital humanities librarian at the Brown University Library, Mylonas will provide technical advice and support primarily on dissemination and preservation and on broader applications in digital humanities for the project's results.

Kathy Takayama (research.brown.edu/myresearch/Kathy_Takayama), Executive Director of the Sheridan Center for Teaching & Learning (brown.edu/Administration/Sheridan_Center/), Brown University. Takayama will advise on how best to utilize our project for humanities scholarship and pedagogy in depth and breadth. She will create opportunities to present our work as a case study in a featured program series on digital humanities at Brown. She will also advise on potential opportunities to disseminate our work on a large scale through her leadership of Brown's online pedagogy initiatives (including Massive Open Online Course, or MOOCs). In collaboration with the Brown Library, she will provide multiple opportunities for the project team to participate in discussions on teaching and curricular planning with graduate students and faculty in the humanities and cross-disciplinary studies.

John Melson (brown.edu/about/administration/sheridan-center/people/staff/john-melson), Instructional Designer, Sheridan Center for Teaching & Learning, Brown University. Melson was formerly the Project Manager and Textbase Editor of the Women Writers Project and has expertise in literary scholarship and digital humanities, particularly in the areas of text encoding, data visualization for the humanities, electronic publishing, digital pedagogy, and interface design. He will advise the project on potential research and teaching applications within digital humanities fields, especially in relation to literature.

David Targan, Associate Dean of the College for Science, Brown University. Targan will be an advisor in the future implementation of the NHM as part of digital literacy education across disciplines. As some of the prominent features of NHM have been inspired by science (cf. the current visualization method of KWlinks), his expertise in high performance computing and science education will help bring a fresh view on the development and fine-tuning of the application.

Research assistant

To be determined (advanced undergraduate or graduate student). S/he will work as web-designer (80%) for the project dissemination website (20%) (5 months; 10 hours/week). We are in the process of identifying a candidate that fits the following criteria:

Required:

- enthusiasm for working with new technologies
- interest in linguistics, especially sociolinguistics and corpus linguistics
- advanced skills in web designing

Preferred:

- familiarity with Czech
- advanced knowledge of corpus technology
- basic knowledge of linguistics and statistics, CSS and PHP

7. DATA MANAGEMENT PLAN

Three types of data will be used in our project: (1) the corpora and the texts used for analysis, (2) the code for the NHM application and the Best Practice Guidelines, and (3) the results of our analysis on the specific language material:

(1) Corpora and texts used for analysis:

The texts for analysis during the initial testing (Husák's New Year's Addresses, NYAs) are about 30,000 words in length. Texts from different genres will also be provided to showcase how NHM can be used for various types of text analysis and for users to replicate the results. The NYA texts in print are publicly available in libraries and their scans are online [<http://archiv.ucl.cas.cz/index.php?path=RudePravo>], but we will upload the texts, in plain text and HTML format (ready for use in NHM), and the scans for use by future researchers who wish to repeat or enhance our work. All texts created for this project that have no other restrictions will also be stored in the Brown Digital Repository which has a commitment to preserving materials and making them available over the long term (<https://repository.library.brown.edu/studio/policies/>).

All the corpora we will use in the analyses (corpus of contemporary written Czech – SYN2005, SYN2010, and corpus of totalitarian language – Totalita and its subcorpora and others (in a scrolled down menu under "Select reference corpus" at <http://kwords.korpus.cz/#z> are the property of the Institute of the Czech National Corpus (ICNC), and they are maintained and hosted on ICNC's servers (www.korpus.cz). The corpora are available to researchers with the caveat that adjustments and additions may occasionally be made in the future.

(2) Code of the application

The NHM application is intentionally made small to increase speed (192MB including reference corpora data). The application and the Best Practice Guidelines will be stored and hosted on the servers of the Institute of the Czech National Corpus as well as the Brown Digital Repository. All the servers are regularly checked and undergo periodic backups. The application itself will be released under an open-source license (GPL) and will be available for use and also for download from the project website. The ICNC is an academic unit of Charles University in Prague, Czech Republic, which is a member of the CLARIN initiative for dissemination and preservation of language data.

(3) Results of our research on specific language material and theoretical aspects of the NHM:

The research results (interim reports, manuscripts of submitted research papers, digitized texts and a copy of the Best Practice Guidelines for the NHM) will be stored in the Brown Digital Repository (<https://repository.library.brown.edu>) as well as the Brown research project website. There will also be links to these sites from the Department of Slavic Languages website. The Slavic Department webpages will contain basic information about the project and its participants with links to the ICNC-hosted software. The Brown Digital Repository has a commitment to preserving and making available scholarship produced by Brown researchers.

**8. LETTER OF SUPPORT (L. JANDA) AND LETTERS OF COMMITMENT
(CONSULTANTS AND ADVISORY BOARD MEMBERS)**

UNIVERSITETET I TROMSØ UiT



HSL FAKULTETET
INSTITUTT FOR SPRÅKVITENSKAP

Deres ref.:
Vår ref.:
Dato: 13.08.2013

National Endowment for the Humanities
1100 Pennsylvania Ave., NW
Washington, D.C. 20506 USA

I am writing to recommend that the proposal "The Needle-in-a-Haystack Method: A New Prototype Corpus-Based Keyword Analysis Tool", submitted by Masako Fidler and Václav Cvrček, be awarded an NEH Digital Humanities Start-up Grant Level II.

In this letter I will first give a brief description of my own relevant qualifications and then turn to the project at hand.

I have been a full professor of Slavic Languages/Russian Linguistics at the University of North Carolina (1996-2007) and the University of Tromsø (2008-present) for over seventeen years, during which time I have served on many committees for the evaluation of grant proposals, both in the US and Europe. I have many years of experience in evaluating scholarly manuscripts, particularly as Associate Editor of *Cognitive Linguistics* (among the top-ranked journals in linguistics worldwide). My CV can be downloaded at: <http://ansatte.uit.no/laura.janda/CV%2013.doc>.

The Needle-in-a-Haystack project is highly meritorious for many reasons. This is exactly the kind of project that we should focus on in building up digital resources for the humanities. It is well grounded both theoretically (sociolinguistics, cognitive linguistics, grammaticalization, lexical priming) and practically (corpus linguistics, statistics, software development). The wedding of linguistics with textual analysis in this way is unique and can lead to potential applications within and beyond literary studies, politics, and sociology. The method and the deliverables are portable to other languages and topics and can thus create a long lasting standard of best practices.

The principle investigators in this project bring to it vast experience in developing and managing linguistic software and presenting it in a way that is easy for users to navigate. Masako Fidler is the author of the Brown University Czech Anthology, which is used by instructors throughout the US and Europe. Václav Cvrček is the director of the Czech National Corpus, one of the most respected digital resources of its type worldwide. Both principle investigators have impressive scholarly track records, with numerous prominent books and articles in the field of linguistics. These two scholars have already presented preliminary results from this project at a number of international conferences (for example the annual meeting of the American Association of Teachers of Slavic and East European Languages in Boston January 2013 and the International Cognitive Linguistics Conference in Edmonton Alberta June 2013), and these presentations have already attracted considerable interest to the project.

It is important to note that Masako Ueda and Václav Cvrček are on the leading edge of a move toward statistical analysis of linguistic data in both Slavic in particular and linguistics in general. In a recent article I presented evidence that 2008 was a turning-point for cognitive linguistics in this respect. Although we have always had some quantitative analyses in the pages of *Cognitive Linguistics*, five years ago we definitively crossed the 50% line, and it is likely that the majority of articles in our field in the future will involve quantitative perspectives.

The Needle-in-a-Haystack project has several remarkable features that make it stand out. One is an innovative use of the concept of keywords. In all studies that I have seen previously, keywords are postulated externally and rather arbitrarily, based on assumptions made by the researchers. In this project, by contrast, no assumptions are made and the keywords are instead extracted by an entirely unbiased objective method. Instead of being assigned, the keywords emerge organically from the data. This study will mark a watershed that all future studies will have to refer to.

Another special feature is the way in which this study takes into account not just lexicon but also grammatical features. Most studies of textual analysis, and indeed an overwhelming majority of corpus linguistic studies, focus only on lexical features, apparently forgetting that grammar plays an enormous role in the expression of meaning in language. Some of the blame for this limitation lays with the fact that so much research is focused on English, which is relatively speaking a morphologically impoverished language (and typologically rather unusual in that regard). Czech is, by contrast, heavily inflected, presenting a large number of interesting language-specific grammatical challenges. It is therefore an excellent laboratory for developing a tool that could be used across various languages.

In short, I find the Needle-in-a-Haystack project to be exciting and groundbreaking. The principle investigators are highly qualified and have proven capacity to deliver quality results. The design guarantees that this project will have a long-lasting impact on linguistics, textual analysis, and digital humanities.

Respectfully submitted,



Laura A. Janda
professor

laura.janda@uit.no

<http://ansatte.uit.no/laura.janda/>



August 12th 2013

Dear Mako and Vaclav,

This letter is to acknowledge my willingness to act as a Consultant on your research project 'The Needle-in-a-Haystack Method: A New Prototype Corpus-Based Keyword Analysis Tool'

In this role I am happy to come to Providence and Prague to evaluate the progress of the project and discuss with you the implementation of keyword analysis method to languages other than Czech, especially English, at meetings during the grant period, as well as discuss future grant applications.

Best wishes,

Professor Paul Baker

Department of Linguistics and English Language

Lancaster University



LINGUISTICS AND ENGLISH LANGUAGE
BRIGHAM YOUNG UNIVERSITY
4064 JFSB
PROVO, UT 84602
(801) 422-2937 / FAX: (801) 422-0906



From: Mark Davies, BYU
To: NEH Digital Humanities Review Committee
Re: Proposal from Václav Cvrček and Masako Fidler
Date: August 19, 2013

I have been asked to be a consultant for the project “The Needle-in-a-Haystack Method: A New Prototype Corpus-Based Keyword Analysis Tool”, which is being proposed by Václav Cvrček (Institute of the Czech National Corpus, Charles University in Prague, Czech Republic) and Masako Fidler (Department of Slavic Languages, Brown University).

As a consultant, I will evaluate the interim progress of the project and advise the PI’s on the implementation of keyword analysis method to languages other than Czech, especially English. This will take place at two meetings during the grant period – once in Providence and once in Prague.

I have received three large, multi-year grants from the National Endowment for the Humanities and I have also served on review committees for other NEH proposals. With this background – and having read the proposal by Cvrček and Fidler in some detail – I am thoroughly convinced that this project has critical relevance for the humanities in general. This is also evidenced by the addition to the project of literary scholars like John Melson, who is now a member of the Advisory Board.

In summary, I am convinced that the PI’s have the necessary skills and knowledge to successfully carry out the project, and I am very pleased to be associated with this project.

Sincerely,

Mark Davies
Brigham Young University
<http://davies-linguistics.byu.edu>



Linda J. Cook
Professor, Dept. of Political Science
Chesler-Mallow Senior Research
Fellow, Pembroke Center
Providence, RI 02912
phone: 401 863-2505
Fax: 401-863-7018
Email: Linda_Cook@brown.edu

August 25, 2013

Professor Masako Fidler
Program Director
Dept. of Slavic Languages
Brown University
Providence, R.I. 02912

Dear Mako,

This Letter of Commitment is to confirm that I would be happy to serve as a member of the Advisory Board for the proposed project, “The Needle-in-a-Haystack Method: A New Prototype Corpus-Based Keyword Analysis Tool.” I understand that you, as co-PI with Vaclav Cvrcek of Institute of the Czech National Corpus, Charles University, Prague, are proposing this project for an NEH Digital Humanities Start-Up Grant Level II.

The study would apply an innovative analysis to textual indicators of societal developments, based on the texts of 1975-1989 New Year’s Addresses by Gustav Husak, then-President of Communist Czechoslovakia. My role would be to advise the project’s PIs on qualitative analysis of these political speeches, through Skype calls two to three times a month for the duration of the project, and participation in a one-day meeting in spring, 2015. My expertise on East European politics during these years – the last decade of Communism and beginning of the transition period, qualify me well for such an advisory role. I have extensive background, research, and field work experience in the Former Soviet Union beginning in the early 1980s, including study of many primary sources (in Russian) and close attention to political developments throughout the East European region. (Please see the short bio and link to my Brown Faculty Research profile, included with this letter.)

The project seems designed exceptionally well to bridge the social sciences and humanities, to make quantitative digital linguistic analysis of key political speeches available to people in the humanities and others who rely mainly on qualitative methods, as well as those who do not speak Czech or Slovak, including contemporary students.

[Type text]

While the major work of the project would of course be done by linguists, I could contribute the perspective of a political scientist with a career-long specialization in the region. I would very much enjoy participating in the project, and anticipate that its results will be both original and valuable across several disciplines. We really still understand little about the causes and underlying historical processes that led to the collapse of Communism, which was especially rapid and peaceful in what is characterized as Czechoslovakia's 1989 'velvet revolution.' The proposed linguistic analysis of the leader's key speeches over what were, in retrospect, the last fifteen years of the regime would be extremely interesting to political scientists, historians, and others working to reconstruct the processes that led the precipitous events of 1989, and transformed all of Eastern Europe.

If you would like any additional information, please e-mail me at: Linda_Cook@brown.edu
Thank you very much.

Sincerely,

Linda J. Cook

Professor Masako Fidler
Department of Slavic Languages
Brown University
Providence, RI
USA

12th August 2013

Letter of Commitment

Dear Mako,

Thank you for inviting me to collaborate with you in the project “The Needle-in-a-Haystack Method: A New Prototype Corpus-Based Keyword Analysis Tool.” I am happy to serve as Co-PI with you during the period May 2014 – October 2015. I understand that my title at Brown will be Visiting Assistant Professor of Research.

Our collaborative research on Husak's New Year's addresses has been a great experience for me, several good conference presentations came out of this project and we are now drafting journal articles. We have a solid beta-version of keyword analysis for Czech and our first attempt at the alpha-version for English. I am especially proud that we have implemented a new method of keyword ranking (DICE) and a visualization technique with D3.js, which cannot be found in any other similar software. As we discussed at Corpus Linguistics Conference in Lancaster this summer, I am committed to our on-going plan for multilingual adaptation of our prototype.

If funded by the NEH, I understand that I will be responsible for several activities: software development with my colleague Pavel Vondříčka, and quantitative and qualitative analysis of texts to evaluate the effectiveness of the software and drafting of the Best-Practice Guidelines with you. I am honored to work in consultation with Paul Baker and Mark Davies, two most recognized corpus linguists, and the members of the Advisory Board at Brown University.

This will be a truly interdisciplinary research opportunity involving literature, political science, digital humanities, technology and education, discourse analysis, and corpus linguistics. I am convinced that such collaboration will lead to the most comprehensive Best-Practice Guidelines that will bridge theory, raw data, and interpretation of text. I am keenly aware of the existing disconnect between technology and its application to text analysis; as more and more users are becoming interested in quantitative analysis of large texts, the time is ripe for creating not only software but software with comprehensive Guidelines that will clearly state (for non-IT users!) the best features as well as the limits of the application.

I realize there is still work ahead, but I am excited about the prospect of providing other humanists with a new way to explore texts in depth and breadth. I believe that the NEH Digital Humanities Start-up Grant will strengthen our chances of securing funds based not only in the United States, but also in Europe (e.g. the Norway Grant). I would like to explore such opportunities with you to continue our project beyond the NEH grant period.

The project is of primal interest to me and to my Institute. I am prepared to devote maximum possible effort to this project, minimally 6 hours a week during the entire grant period.

I verify that the fund from the NEH will be the only source of support for this project.

Looking forward to our continued cooperation with you,



Václav Cvrček, Ph.D.
Chair
Institute of the Czech National Corpus
Charles University in Prague
Czech Republic

Professor Masako Fidler
Department of Slavic Languages
Brown University
Providence, RI
USA

2th August 2013

Letter of Commitment

Dear Mako,

Thank you for inviting me to collaborate with you and Václav Cvrček in your project “The Needle-in-a-Haystack Method: A New Prototype Corpus-Based Keyword Analysis Tool”. I will gladly help as technical support of the project during its period (May 2014 – October 2015).

If funded by the NEH, I understand that I will participate in this project as a technical support for 9 months (5 hours/week) during the grant period. I will be responsible especially for the development of the application for keyword analysis, which we already started to build in co-operation with Václav. The current version of the program provides the user with some basic information about the text (based on keyword analysis), but it lacks many crucial features and enhancements, which should be incorporated in the future, including improving the stability of the application. I will be more than happy to discuss with you other features as we make progress in our analysis.

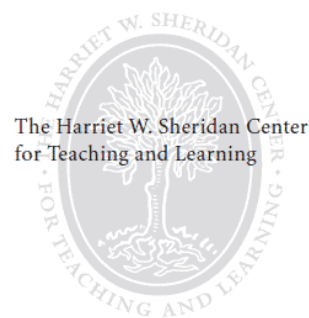
Sincerely,



Pavel Vondříčka, Ph.D.
Institute of the Czech National Corpus
Charles University in Prague
Czech Republic



BROWN



August 20, 2013

Masako Fidler
Department of Slavic Languages
Brown University

Václav Cvrček
Institute of the Czech National Corpus
Charles University in Prague

Dear Masako and Václav,

I am writing to express my enthusiastic support for your NEH Digital Humanities Start-Up proposal, “The Needle-in-a-Haystack Method: A New Prototype Corpus-Based Keyword Analysis Tool.” The development of your “Needle-in-a-Haystack Method” (NHM) and the software based on it represent crucial progress in developing methods of textual analysis that contribute simultaneously to linguistic and literary scholarship.

I am very pleased to contribute to this important project as a member of the Advisory Board and through my participation in the Spring 2015 project meeting. In my position at the Sheridan Center for Teaching and Learning at Brown University, I will be happy to consult on pedagogical applications of your work, particularly in the context of humanities teaching and research. And, as the former Project Manager and Textbase Editor for the Women Writers Project (a recipient of multiple NEH grants, including a Digital Humanities Start-Up Grant during my tenure), I will gladly advise on strategies for making NHM and its tools maximally useful to researchers working in other areas of the digital humanities.

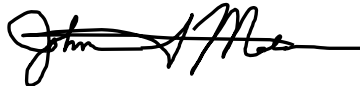
My own background is in both literary studies and digital humanities, so I want to emphasize the significance your project holds for both domains. While there is a growing general awareness of computational methods within “mainstream” literary studies, literary scholars only rarely have the training required to apply such methods with a complete understanding of what they can—and cannot—tell us about language and literature. Clear, accessible resources like the Best Practices Guidelines you aim to produce are vital to sustaining the kinds of intellectual cross-fertilization implicit in contemporary digital literary studies. Currently, the lack of such guidelines constitutes a major impediment for scholars, like myself, who would otherwise be eager to draw on data-driven, corpus-based methods in their own qualitative studies of literary and historical texts.

Beyond these practical guidelines, however, I see enormous long-term potential for NHM to assist literary scholars in evaluating the patterns of language that enact subtle cultural shifts within literary texts. By providing researchers with the ability to track and visualize individual

keywords that stand out, relative to multiple reference corpora, within target texts, NHM and its associated suite of tools provide new opportunities to examine at a large scale how particular texts differ—and how those differences shift at various historical moments and for multiple groups of readers. Such methods of large-scale analysis suggest a range of applications for contemporary literary scholarship, from comparative genre studies to reception history to detailed historicist analysis.

I am very excited to support this project and to see its continued development in the future.

Sincerely,

A handwritten signature in black ink, appearing to read "John Melson". The signature is fluid and cursive, with a long horizontal stroke extending to the right.

John Melson, Ph.D.
The Harriet W. Sheridan Center for Teaching and Learning
Brown University



August 23, 2013

Masako Fidler
Slavic Studies
Brown University

Dear Mako:

I would be very happy to serve in an advisory capacity on the NHM Project. I look forward to contributing to the progress of the project by advising on the development of the project for the duration of the grant. I anticipate that this will take place in face-to-face meetings and via video conference (Skype), once or twice monthly. I can also provide advice and feedback on the design and development of the research website, and look forward to seeing the results of your work at the Advisory Board meeting in 2015. As senior digital humanities librarian at the Brown University Library, I can provide technical advice and support on dissemination and preservation and on broader applications in digital humanities for the project's results.

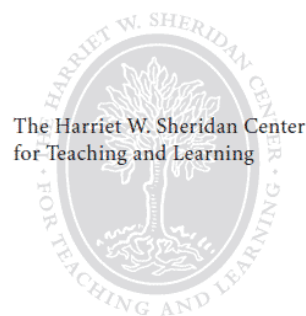
The results of this project, both software and methods, will be of great interest to a wide range of humanities and social science researchers. It forms part of a growing interest in textual analysis. Your implementation, which will be accompanied by documentation not only on how to use your software, but also how to answer research questions using it, should appeal to a broad audience.

With my best wishes for your success,

Elli Mylonas
Senior Digital Humanities Librarian
Research and Outreach Services and
Center for Digital Scholarship
Brown University



BROWN



August 18, 2013

Professor Masako Fidler
Department of Slavic Languages
Box E
Brown University

Dear Mako,

I am very pleased to provide this letter of commitment and support for your project entitled, “*The needle-in-a-haystack method: A new prototype corpus-based keyword analysis tool.*” Your project has immense significance for the humanities through the development and implementation of an innovative and objective methodology for the large-scale quantitative analysis of texts.

As a member of the project Advisory Board, I am fully committed to contributing my expertise and the resources of the Sheridan Center toward the application and dissemination of project outcomes toward professional development for students and faculty in digital literacy. Toward this end, I am prepared to participate in a full-day meeting during the grant period to plan a concrete approach toward the creation of a sustainable program in which digital humanities scholars can benefit from the exploration and application of significant innovative approaches to expand their scholarship in depth and breadth. I am committed toward creating opportunities to allow you to present your work in a featured program series on digital humanities at Brown. Furthermore, I am prepared to advise you on potential opportunities to disseminate your work on a large scale through my leadership of Brown’s online pedagogy initiatives (including Massive Open Online Course, or MOOCs).

The Sheridan Center has longstanding expertise in collaborating across the disciplines at Brown as well as nationally and internationally in providing recognized programs and initiatives in the professional development of faculty, graduate students and postdoctoral fellows. As one of the oldest Centers in the country, we have had the benefit of responding to the continually changing needs of the academic mission of the university for 25 years. Through this experience, we have built a solid foundation through which we create sustainable partnerships and programs that ultimately have had continued long-term impact. As such, I am committed toward enhancing the dissemination of your project and creating venues for further dialogue that will continue to inform future refinement or enhancement of your methodology. I invite you to present your work as a case study on digital humanities at the Center’s consistently well-attended (often fully-booked with waitlisted registrations) featured programs series. In collaboration with the Brown

Library, I believe the Sheridan Center will be able to provide multiple opportunities for training graduate students and faculty in the digital humanities, and nucleate cross-disciplinary dialogues.

Thank you very much for this opportunity to support your project, and I look forward to working with you.

With very best wishes,

A handwritten signature in black ink, appearing to be 'Kathy M. Takayama', written in a cursive style.

Kathy M. Takayama, PhD
Executive Director



BROWN

DEAN OF THE COLLEGE - SCIENCE CENTER
Brown University
Box 1922
Providence, RI 02912
PHONE 401 863-6890
FAX 401 863-6046
sciencecenter@brown.edu

Dear Mako

It is my pleasure to write this letter of detailing our support for your NEH Digital Humanities Start-up Grant (Level II) proposal, entitled "A Needle-in-a-Haystack Method: A New Prototype Corpus-Based Keyword Analysis Tool."

As Director of Brown University's Science Center, I welcome all opportunities to work closely with faculty and students from the social sciences, arts, and humanities. The proposed project falls within the scope of the Science Center's mission to integrate activities in these fields with those of the life and physical sciences. For example, as part of our experimental work with the Center for Computation and Visualization (CCV), we have supported an undergraduate "science center Fellow" concentrating in computational biology. He spent last semester developing modules to introduce students in archeology to high performance computing techniques. We plan to support such interdisciplinary collaborations in the future, especially as they relate to computation and the use of CCV's supercomputer. We envision support of computationally-capable undergraduates working with your project to develop tools that could benefit faculty and students in linguistics.

Finally, I plan to commit to serving on the advisory board, I bring to your board familiarity with high performance computing (as it relates to our "citizen science" sky imagery distributed analysis program in astrophysics), including various visualization methods used for natural science that could be of interest to your project. Furthermore, my expertise is in the area of undergraduate science communication and education, and I would bring this perspective to the task of dissemination and integration of project ideas into existing and new courses.

The specific commitment of my time to the selection and supervision of a computationally capable Science Fellow is variable but an estimation of one full day would be in line with current work on that project. My commitment to the advisory board would include a one-day session listed in the section Evaluation and Future Planning.

Best wishes,

David Targan, PhD

Director, Science Center
Adjunct Associate Professor of Physics
Associate Dean of the College for Science
Brown University
Providence, RI 02912
[401-863-2314](tel:401-863-2314) (University Hall)
[401-863-6890](tel:401-863-6890) (Science Center)
[401-225-8871](tel:401-225-8871) (Mobile)
<http://brown.edu/academics/science-center/>
<http://www.brown.edu/Departments/Physics/Ladd/>

9. APPENDICES

Appendix 1: Institutional support

The Institute of the Czech National Corpus (ICNC) (ucnk.ff.cuni.cz/english/index.php) at Charles University in Prague in the Czech Republic, is the depository of all major language data in Czech in the world, and includes, among other texts, periodical texts from the socialist period from the 1950s to 1989, which are crucial for our research.

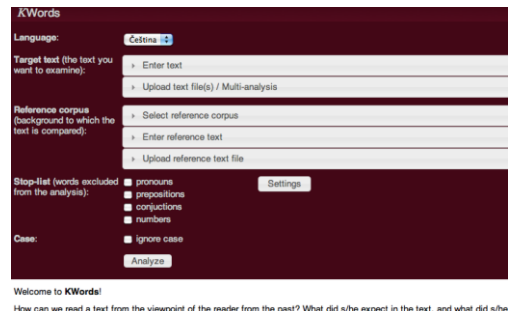
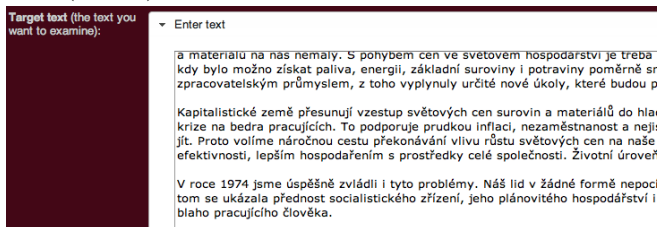
Brown University is known for its continued strengths in digital humanities. The Brown Corpus, which is in use internationally to this day, was a collaborative project by Henry Kučera and Nelson Francis during the 1960s at Brown; it was the first language database in the world and a basic source for the development of *The American Heritage Dictionary*. Other scholars have since contributed to digital humanities projects: George Landow's *The Victorian Web*, Massimo Riva's *The Decameron Web* (supported by the NEH) and *the Virtual Humanities Lab* (supported by the NEH) and, until recently, the Women Writers Project directed by Julia Flanders (supported by NEH, among others). Eugene Charniak, in Computer Science, has been involved in projects on computational analysis of grammar and meaning with a goal in machine translation. Currently he is working on a computer analysis of variety in Igbo, an African language.

The Center for Computing and Visualization at Brown manages large volumes of data and supports projects for digital humanities. The Center for Digital Scholarship in the Brown Library has been helping faculty develop digital humanities projects. The Brown Library Digital Repository is committed to preserving and making available scholarship produced by Brown researchers. The project team has been in communication with the members of all these units.

Appendix 2: Screen shots of web-based keyword analysis application for Czech

The current NHM (a testing phase beta version) opens with the main page (<http://kwords.korpus.cz/>) (right).

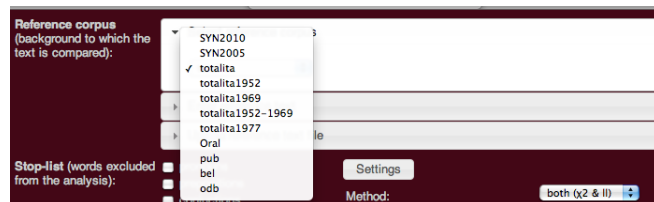
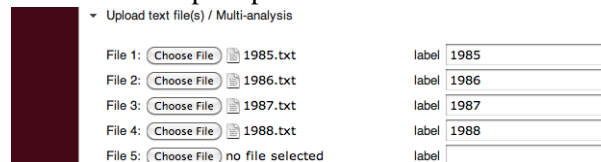
There are two ways to input T-txts. The user may enter a single T-txt s/he wishes to analyze by pasting it in (just as in Microsoft Word) or by uploading a plain text file in Unicode (below):

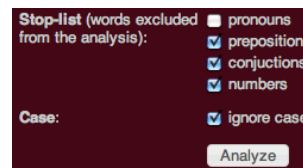
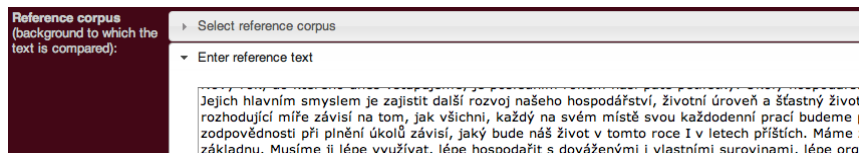


Alternatively, the user may input more than one text file in Unicode. The screenshot below shows how the user input four texts simultaneously. NHM allows the user to input up to 20 T-txts.

The user can use preloaded RefCs as well as user-defined corpora. She can choose from eleven subcorpora in the CNC: they are categorized in time (1950s, 1960s, 1970s, 1950s-1970s, the entire totalitarian era, and two contemporary subcorpora of contemporary Czech) as well as genre (transcribed oral speech, journalism, belle lettres, and science) (bottom right).

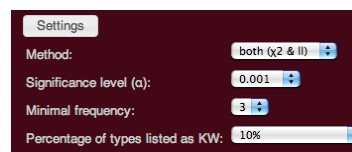
A T-txt can also be contrasted with a user-defined RefC (where the user can input his/her own corpus under "Enter reference text") (next page, top):





The user can set some important features for her KWA. "Case" turns on and off case-sensitivity; "Stop-list" allows the user to exclude pronouns, prepositions or/and conjunctions from the list of KWs with the stoplist, as these "function" words may not be central to the analysis (top right):

NHM allows selection in the method of KWA and ranking: chi-square or log-likelihood for identifying KWs; chi-square, log-likelihood and DICE for ranking KWs. The user can choose the significance level and minimal frequency; s/he can choose to view the top 5%, 10% and all KWs (right, second picture from the top).



Upon clicking the "Analyze" button for multiple texts, analysis of each text appears under different

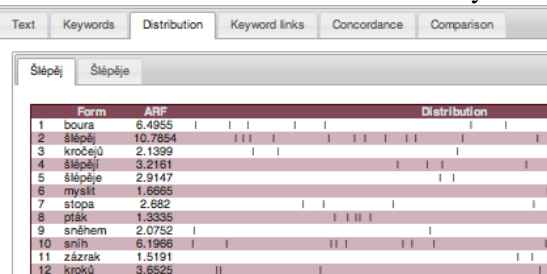
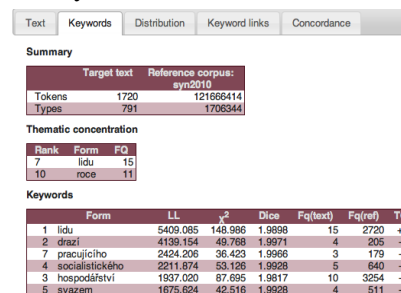


Vážené soudružky, vážení soudruzi, **drázi** spoluobčané!
 Vstoupili jsme do nového **roku** 1975. Dovolet mi, abych vás všechny na jeho prahu **jménem** ústředního výboru **Komunistické strany** československé vlády i nemocného prezidenta republiky a **jménem** svým srdečně pozdravil a popřál vám v novém **roce** hodně zdr
 úspěchů.
 V životě **člověka** i v životě **společnosti** a **státu** je to příležitost k zamyšlení, co se nám v **minulém roce** zдалo a co ne, co nás oče
 Můžeme **řici**, že uplynulý **rok** 1974 byl ve všech **oblastech** pro Československo dobrým **rokem**. Důstojně se přiřadil k několika před
 celé **společnosti**. Dobře pracoval náš průmysl, stavebnictví, zemědělství, příznivé jsou **výsledky** v ostatních **oblastech**. Úspěšně j
 Období, kterým jsme prošli od XIV. sjezdu **Komunistické strany** Československa, je charakterizováno příznivým vnitropolitickým v
socialistické demokracie, politickou angažovaností i pracovní aktivitou dělnické třídy, rolnictva, inteligence, žen, mládeže i prohlui

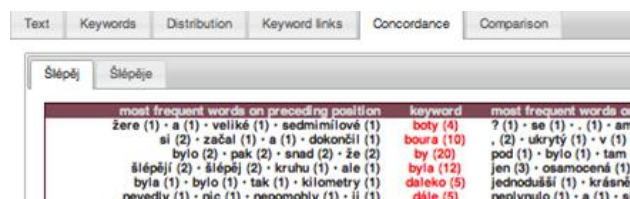
Tabs (left, Tabs for texts 1975 and 1976). The T-txt appears in the window with the keywords marked in red (left). When some KWs are particularly prominent, they will also be listed as KWs with "Thematic concentration" (marked in yellow).

Besides "Text", there are five more tabs: "Keywords", "Distribution," "Keyword links," "Concordance," and "Comparison". "Keywords" shows the list and ranking of KWs and overall statistics of the target text (right):

"Distribution" displays the positions of KWs within the target text. It also shows the value of ARF (average relative frequency): the extent to which each KW is distributed evenly throughout the text. (Below the central character *Boura* evenly occurs in the story).

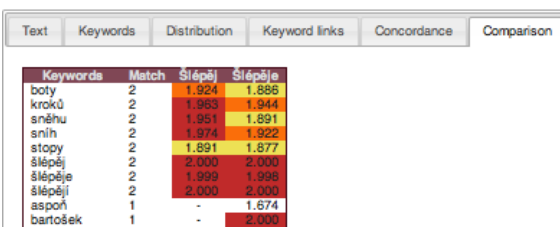


"Concordance" (right) shows collocation patterns for each KW (words occurring in close proximity to each KW)¹. It informs of the meaning and use of each KW (e.g., "brother" in "the Big brother" (=the USSR) would be different from the same



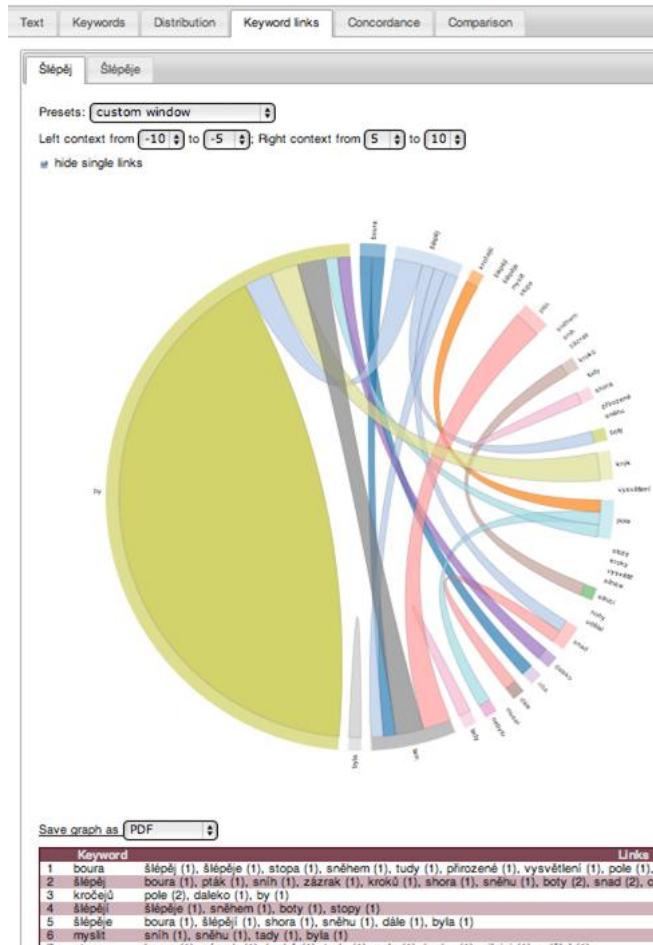
word used in "my brother and sister"). A click on the KW (in red) reveals all the contexts in which the KW occurs.

"Comparison" (left) compares KWs and their degrees of importance in more than one text. This table shows KWs from two texts. NHM currently enables simultaneous analysis of up to twenty texts.



¹ These cooccurring words may not be KWs; Concordance is therefore different from keyword links.

"Keyword links" (right) contains multiple features. It shows both a table at the bottom and a D3.js visualization (the "wheel" at the top) The user can adjust the range of context (left, right and both sides) where the links occur (two presets and custom adjustments). The wheel shows KW ranking from the 12 o'clock position in clockwise direction. When the user places the cursor on one KW, only the links to that particular KW appear and the number and the type of links. The wheel here shows the results when the contexts for KWlinks are custom adjusted.



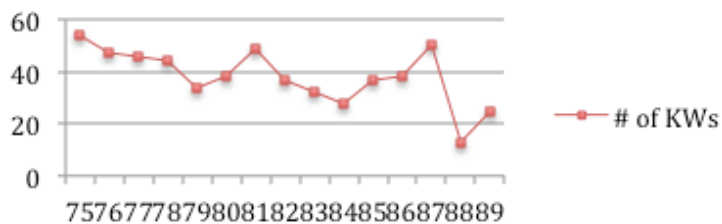
Appendix 3: Potential value of NHM to the humanities (Sample illustrations)

NHM has the potential to probe a variety of texts. Below are some examples of issues that NHM could help address.²

1a. "Ritualistic" censored political texts reflect societal shifts

The raw numbers of KWs in the New Year's Address by Husák have important informational value about the degree of confidence within the leadership of the socialist Czechoslovakia. The decline in the number of KWs often precedes unfavorable events in East Europe and the USSR and reflects Czechoslovak concern as one of the Warsaw Pact nations. The table below shows the total

number of different KWs (types) over the 15-year period (1975-89)³. The decrease in the number of KWs in January 1976 foreshadows the emergence of the Czech human rights movement Charter 77 in 1977, and the escalating demonstrations in the neighboring Poland. The number drops even further in January 1979 right before the spread of unrest in Poland in 1980-81. The number of KWs goes up slightly in January 1981 before the Marshal law is enforced and the demonstrations are suppressed by force in Poland in December 1981. However, the number then continues to decline; this closely precedes the intensifying instability in the Soviet leadership and the breakdown of the arms-control negotiations in 1983. The number of KWs then starts going up in January 1985, slightly before Gorbachev comes to power with his reform plans and starts a new round of disarmament negotiations with Reagan. The number, however, takes a deep plunge in 1988 and doesn't quite recover. The slight increase in the number of KWs in January 1989 precedes the demonstrations in Czechoslovakia in



² The results from Husák's speeches (1 and 2) will be further compared with other political speeches from the same period.

³Totalita was used as the RefC. Although the number of different KWs and length of texts are said to correlate frequently (Scott 2006) , the relationship might not be as strong as it is claimed.

1989 and the government's harsh reaction to them—the last attempt at oppression, followed by the fall of the socialist regime in November of 1989.

1b. "Isolate" word forms point to societal changes.

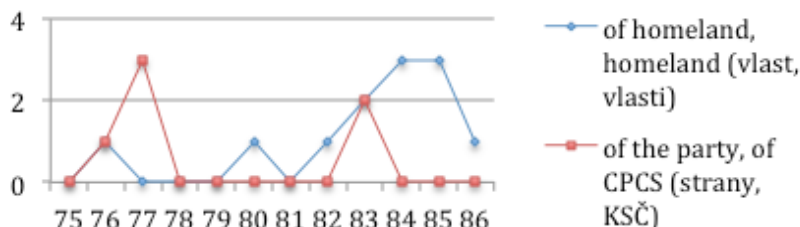
("Isolate KWs") are word forms that appear in the KW lists from Husák's speeches only occasionally during the 15-year period. They tend to coincide and/or even precede important events that are pertinent to the leadership's concern for maintaining its power in the country. Below is a non-exhaustive list of such words for illustration:

Speech & date	KWs	Events
January 1975	"natural resources," "prices," "crisis," "humankind"	The energy crisis (1975)
January 1977	"social structure" and (social) "layers"	The Helsinki Accord on human rights signed by Czechoslovakia (1976)
January 1981 and 1982	"peaceful" and "of peaceful (noun)" (1981) "of peace" (1982)	Marshall Law in Poland (December, 1981-83), breakdown of the disarmament negotiations between USA and the USSR (1983)
January 1985	"disarmament"	Restart of the disarmament negotiations between the USA and USSR (August, 1985)

1c. KWlinks of recurring KWs suggest societal changes.

Some word forms can repeatedly appear in the KW lists over time, but they may be linked to different KWs at different times. Such shifts in KWlinks can also reflect political change. For example, the word form "of (the) masses (*lidu*)" is consistently in the KW lists from 1975 to 1986. They are, however, linked to two different KWs roughly in the mid-70s and in the mid-80s (below): to the communist party (*strany*, *KSČ* [the Communist Party of Czechoslovakia]) in the mid-70s and to a less political term "homeland" (*vlast*, *vlasti*) in the mid-80s.

The different types of association correspond to the time when the old-guard communist leaders were losing grip on the nation from the late 70s to the mid 80s.



2. KWs and the longevity of a literary text.

Two short stories—*Courtship (Námluvy)* by Karolina Světlá and *A Picture from a Village (Obraz vesnický)* by Božena Němcová—were analyzed against the background of two RefCs: the corpus from the socialist period (Totalita) and the corpus of contemporary Czech (SYN2010). Světlá and Němcová represent Czech women writers of the second half of the nineteenth century. Both stories deal with a love relationship, which is complicated by the community that defines marriage as a socio-economic institution. Světlá's story ends in an unexpected happy ending, whereas Němcová's story ends in one protagonist's tragic death. While Němcová's texts are still read and studied, Světlá's texts are often considered as outdated and too melodramatic for today's reader.

NHM produces an interesting set of KWs for both texts. In contrast to the analyses of Němcová's story with the two corpora, the analysis of Světlá's story with SYN2010 produces different results from the analysis with Totalita. The former includes among the top 5% KWs (ranked by DICE) the adverb *snad* "perhaps", a word expressing the speaker's wishful thinking in Czech. *Snad* spawns the largest number of KWlinks (96 links) in the analysis based on SYN2010. In contrast, the analysis based on Totalita does not rank this word among the top 5% KWs. Today's readers thus seem to notice *snad* (and the implied wishful thinking) as much more pervasive in the story than the readers from the socialist period. This analysis may provide a clue to possible reasons why Světlá is quickly becoming less appealing than Němcová. At the same time, this analysis may help us understand what type of worldview was pervasive in the totalitarian period. It is highly possible that the totalitarian worldview was one with more pathos to the point of sentimentality, given the widespread narratives in media and literature about self-sacrifice for the

party, dedication and overcoming difficulties for the good of the society.

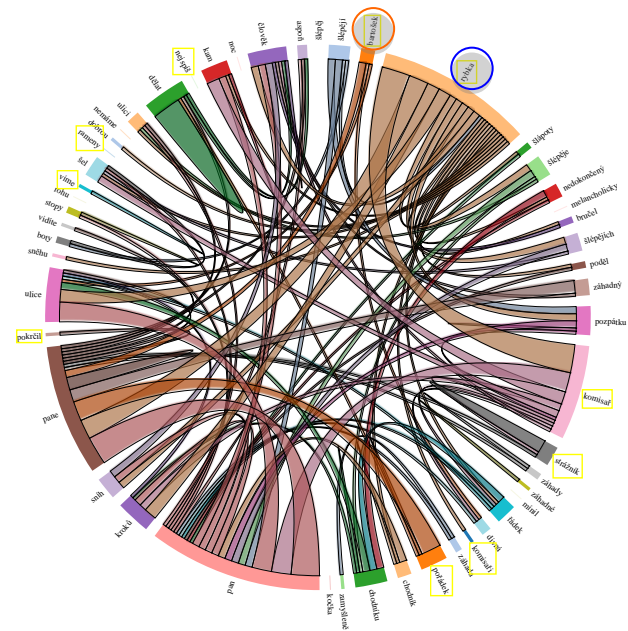
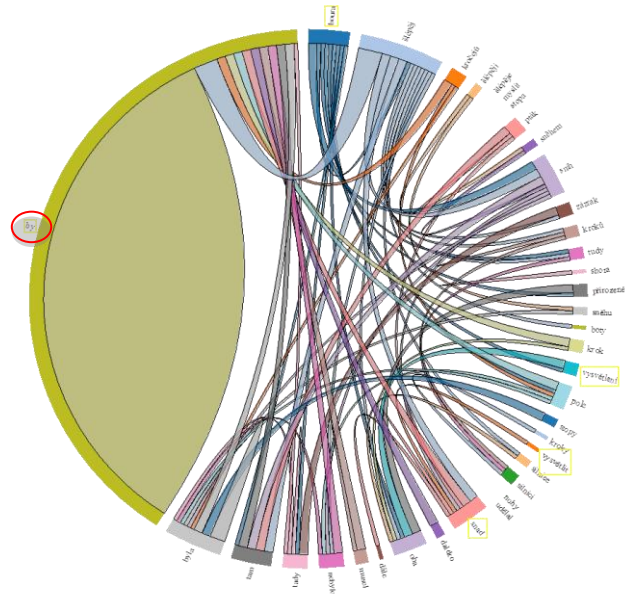
3. Differences between texts that utilize similar plots and motifs: A potential for NHM to explore intertextuality.

NHM has a potential to explore intertextuality based on the preliminary data from two short stories *A Footprint* (1917) and *Footprints* (1929) by Karel Čapek. Both texts deal with mysterious footprint(s) left on the snow. However, *A Footprint* "questions the existence of absolute universal ideologies, religions, and philosophies as well as any other notion for the masses and collective ways of relating to the world," whereas *Footprints* is about denial of such questions and the "pursuit of petty happiness within an enclosed noetic ghetto" (Bílek (brown.edu/Research/CZECH/Aud/Slepeje /); the main protagonist wants a police officer to erase the footprints, thereby "solving" the mystery.

The wheels on the right represent KW rankings as well as KWlinks⁴. The top 5% KWs are ranked by DICE from the 12 o'clock position to the right; the lines indicate links among the KWs⁵.

NHM offers some convincing evidence for this view. When contrasted with the same RefC, these texts show very different rankings of the top 5% of the KWs. In *A Footprint* (above right) the conditional form *by* ('would' marked with a red circle) that signals hypothetical statements is among the top 5% KWs and is connected with a number of top KWs (67 links, including the other occurrences of *by*). The highest ranked KW is *Boura*, the main protagonist who found the mysterious footprint. There are no other KWs referring to other characters, although they appear in the story. A search for explanation seems central to this text, as seen in word forms *vysvětlení* ('explanation') and *vysvětlit* ('to explain') among the KWs.

In *Footprints*, in contrast, conditional forms are not among the top 5% KWs (below right). Furthermore, police commissioner *Bartoušek's* name (marked with an orange circle) is ranked higher than the main protagonist *Rybka* (marked with a blue circle). Other terms referring to the policeman are also



⁴ SYN2010 was used as the RefC.

⁵ The areas that are not connected with other KWs in a wheel indicate that the particular KW is connected to other occurrences of the same KW.

highly ranked (*strážník* 'officer, nom. sg.', *komisař* 'commissioner, nom. sg.', *komisaři* 'commissioner, voc. sg. '), and so is the main concern of the police (*pořádek*, 'order'); word forms used in categorical statements (*víme*, 'we know') and in dismissing behavior (*pokrčil*, 'he shrugged'; *rameny*, 'shoulders, instr. pl. ') are also among the top 5% KWs.

Note that the results are not raw frequencies. NHM does not replace manual counting of word forms in a text. The KWs are obtained in comparison with the same RefC.

Application of NHM to texts that manipulate intertextuality, using similar plots and motifs, therefore seems to be productive. KWs seem to reveal different properties of apparently similar texts.

Appendix 4a: Month-by-month timeline for the Work Plan

The timing of the activities in the Work Plan is presented in the table below.

Abbreviations:

CL- preparation for and participation in Corpus Linguistics Conference, UK

AA CL- preparation for and participation in American Association of Corpus Linguistics conference

M- meeting

P-Posting on various lists

Mo /yr	5 /14	6	7	8	9	10	11	12	1 /15	2	3	4	5	6	7	8	9	10	
Activities	Further testing of NHM with NYAs, comparing data from the qualitative analysis and NHM data																		
	Exploratory application of NHM (information to be included in the Best Practice Guidelines)																		
	Writing the first draft of the Best Practice Guidelines					Writing the final draft of the Best Practice Guidelines													
		Discussion on added functionality to NHM																	
		Coding to improve the robustness of the software, additional functions to NHM for Czech (Public domain release 6/2015)																	
		Coding of the English version based on the Czech prototype (Public domain release 8/2015)																	
		Building the research website																	
							AA CL									CL			
		Journal article 1												Journal article 2					
		*									M 1			M 2					M3
									P								P		

*The meeting with the NEH will take place during the first year of the grant period.

Appendix 4b: A detailed description of the Work Plan

(1) Testing the software and qualitative analysis as a cyclical process.

At different stages of software development we have extracted keywords from each New Year's Address by Husák (NYA) with the beta version of our software, which used the chi-square and/or log-likelihood tests and other techniques ("thematic concentration") (Oakes 1998, Biber et al. 1998, Scott 1999, Mačutek et al. 2007, Popescu et al. 2007, Baayen 2008, Gries 2009, and Popescu 2009). We have also recently harvested word forms using the KW ranking with DICE. Testing of the software will continue as more

enhancements are made. To make the testing process consistent, we will compare the data from the qualitative analysis of the same set of texts and the data from NHM before and after the enhancements are added. Clearly, application of qualitative and quantitative analysis will be cyclical.

Testing of quantitative data with qualitative analysis is a reasonable and widely accepted procedure (cf. Benoit and Lowe 2012), but we will also examine how these two approaches might strengthen each other in order to obtain a more sophisticated and detailed interpretation of texts (Gupta 2013).

As we complete the Czech NHM prototype and the English beta version, we will also apply other texts for two different purposes. We will use different types of texts (short stories, pre-1989 and post-1989 presidential speeches). We will also choose initial testing materials in English (e.g., US presidential speeches) in consultation with the Consultants in corpus linguistics. Results of quantitative and qualitative analysis will be in the Best Practice Guidelines and used for disseminating the effectiveness of NHM.

(2) Exploratory application of NHM to become part of the Best Practice Guidelines

As the Best Practice Guidelines should demonstrate the connection between the raw data and the interpretation, we will include sample results from our tests and sample applications of NHM for various types of text analysis in the humanities (for examples, [Appendix 3](#)). The texts and the data will be uploaded on the research website so interested visitors can replicate the data.

(3) Best Practice Guidelines

The Guidelines will explain the connection between raw data and interpretation of texts. They will promote awareness that analysis does not start or end solely with the extraction of raw data.

Sample demonstrations of text analysis using NHM will also be uploaded to show that the NHM can be used for analyzing various aspects of texts. In order to promote responsible and informed use of NHM, "bad" practice samples will also be presented with explanations of why and how they are bad. This part of the Guidelines is important, as more and more users in the humanities are employing quantitative analysis, yet there is little comprehensive illustration of not only the merits but also the limits of quantitative analysis for a wide audience.

(4) Software development

The Czech NHM has undergone some revisions since the beginning of 2012. Completion of the prototype, however, requires minimally the following.

a. improving the robustness and stability of the application

The speed and the capacity to process larger T-txts (novel-length texts) must be improved. This is the most time-consuming task and likely require new code.

b. developing the User Manual

This feature concerns the mechanics of the software and is separate from the Best Practice Guidelines, which will concern the best procedures for analysis.

c. enhancements to the D3.js visualization of KW links

The D3.js visualization is one of the strengths of NHM and we plan to enhance it even further. We will make revisions to represent KWs more clearly when there is a larger number of KWs to represent.

d. a ruler in the distribution plot

Within the text, we will add a sliding ruler with adjustable width to help follow which word forms tend to occur near the word we examine.

e. collocation lists

An added function of systematically assessing collocation patterns will be necessary as NHM begins to deal with larger texts.

f. function for users to create subcorpora

NHM currently offers a choice of up to eleven RefCs. We wish to add more options to enable the user to slice out the necessary portion of the CNC as his/her RefC and T-txt for analysis (e.g., in terms of a specific author, genre, language⁶).

We are likely to add more features as we continue testing and exploring the potential of NHM in consultation with corpus linguistics Consultants.

⁶ *Intercorp* within CNC includes parallel foreign language corpora (translations of Czech texts).

Appendix 6: Bibliography

- Andrew, D. P. S., P. M. Pedersen, and C. D. McEvoy. 2011. *Research methods and design in sport management*. Human Kinetics.
- Baker, P. 2006. *Using corpora in discourse analysis*. London: Continuum.
- _____. 2009. 'The question is, how cruel is it?' Keywords, Fox Hunting and the House of Commons". *What's in a Word-list? Investigating word frequency and keyword extraction*, ed. in Dawn Archer. Farnham: Ashgates, 125-136.
- _____. 2010. *Sociolinguistics and corpus linguistics*. Edinburgh: Edinburgh UP.
- _____. 2011. Times may change but we'll always have money: a corpus driven examination of vocabulary change in four diachronic corpora. *Journal of English linguistics* 39: 65-88.
- Baker, P. and Ellece, S. 2011. *Key Terms in discourse analysis*. London: Continuum.
- Baker, P., Gabrielatos, C. and McEnery A. 2013. *Discourse analysis and media attitudes: The representation of Islam in the British Press*. Cambridge: Cambridge UP.
- Baayen, H. R. 2008. *Analyzing linguistic data*. Cambridge: Cambridge UP.
- Benoit, K. and W. Lowe. 2012. Qualitative validation of quantitative text scaling. Paper prepared for presentation at the 70th Annual Conference of the Midwest Political Science Association. Palmer House Hotel, Chicago, 12-15 April 2012. Version: 10 April 2012. (www.kenbenoit.net/).
- Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge UP.
- Blei, D. M. and J. D. Lafferty. 2009. Topic models (www.cs.princeton.edu/~7Eblei/papers/BleiLafferty2009.pdf)
- Block, S. 2010. Doing more with digitization: An introduction to topic modeling of early American sources (www.common-place.org/vol-06/no-02/tales/)
- Brown On-line Czech Literary Anthology (brown.edu/Research/CZECH)
- Bybee J. L. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge UP.
- Bybee, J. L. & Perkins, R. & Pagliuca, W. 1994. *The evolution of grammar: Tense, aspect and modality in the languages of the world*. Chicago: The U. of Chicago P.
- Cvrček, V. 2013. A tool for keyword analysis in Czech. Paper presented at the National Convention of the American Association of Teachers of Slavic and East European Languages. Boston.
- Cvrček, V. and M. Fidler. 2012. Analysis of Keywords in Czech Political Texts. A Needle in a Haystack Method. Paper presented at the CADS International Conference, Bologna, Italy.
- _____. 2013. Not all keywords are created equal: How can we measure keyness? Corpus Linguistics Conference, Lancaster, UK.
- Čermák, F. and V. Cvrček. 2009. *Slovník Bohumila Hrabala*. [Vocabulary of Bohumil Hrabal], Prague: NLN.
- _____. 2010. Author dictionaries revisited: Dictionary of Bohumil Hrabal. *Proceedings of the XIV. euralex international congress*, ed. A. Dykstra and T. Schoonheim. Ljouwet: Fryske Akademy Afûk, 592-8.
- Čermák, F., V. Cvrček, and V. Schmiedtová, ed. 2010. *Slovník komunistické totality* [Dictionary of the totalitarian regime]. Prague: NLN.
- Däubler, T, K. Benoit, S. Mikhaylov, and M. Laver. 2012, forthcoming. Natural sentences as valid units for coded political texts. *British journal of political science*. (www.kenbenoit.net/cv/publications/)
- Duguid, A., 2010. Newspaper discourse informalisation: a diachronic comparison from keywords. *Corpora*, 2(10): 109-138.
- Ensslin, A. and W. Slocombe. 2011. Training humanities doctoral students in collaborative and digital multimedia. *Arts and humanities in higher education* 11: 140-156.
- Fidelius, P. 1998. *Řeč komunistické moci* [The speech of communist power]. Prague: Triada.
- Fidler, M. and V. Cvrček. 2012. A keyword analysis of totalitarian texts: A case study. Paper presented at the 7th Annual Meeting of the Slavic Linguistics Society. Kansas. (Cf. Folder "Fidler-Cvrcek" at https://documents.ku.edu/xythoswfs/webview/fileManager?stk=&entryName=%2Fusers2%2Fmlg%2FSLS_7_handouts&msgStatus=)

- _____. 2013a. Keyword analysis with a usage-based perspective: A preliminary study in Czech. Paper presented at the National convention of the American Association of the Teachers of Slavic and East European Languages. Boston.
- _____. 2013b. Usage-based approach to discourse through keyword analysis. Paper presented at the International Conference of the Cognitive Linguistics Association. Edmonton, Canada. (www.ualberta.ca/~iclc2013/PRESENTATIONS/2013-07-01-19-34-49-fidlerandcvcrc%CC%8Cek_masakoandva%CC%81clav.pdf)
- Gabrielatos, C. and A. Marchi. 2012. Keyness: appropriate metrics and practical issues. Paper presented at the CADS International Conference, Gologna, Italy (www.gabrielatos.com/Presentations.htm).
- Gries, S. T. 2009. *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.
- Gupta, K. 2013. A triangulated approach to media representations of the British women's suffrage movement. Paper presented at the 7th International Corpus Linguistics Conference, Lancaster, UK.
- Heine, B. and E. C. Traugott, ed. 1991. *Approaches to grammaticalization. Typological studies in language*, 19. Amsterdam: John Benjamins.
- Hoey, M. 2005. *Lexical priming*. New York: Routledge.
- Hopper, P. and E. C. Traugott, *Grammaticalization*. 2003. 2nd ed. Cambridge, UK: Cambridge UP.
- Kilgarriff, A. 2009. Simple maths for keywords proc. Corpus Linguistics, Liverpool, UK (ucrel.lancs.ac.uk/publications/cl2009/171_FullPaper.doc).
- Mačutek, J., I. I. Popescu, and G. Altmann. 2007. Confidence intervals and tests for the h-point and related text characteristics. *Glottometrics* 15: 42–52.
- MacCallum, A. et al. MALLET project (mallet.cs.umass.edu/about.php)
- Mujis, D. 2010. *Doing quantitative research in education with SPSS*. Sage.
- Oakes, M. P. 1998. *Statistics for corpus linguistics*. Cambridge UP.
- Phiip, G. A defence of semantic preference. Paper presented at the 7th International Corpus Linguistics Conference, Lancaster, UK.
- Popescu, I. I. ed. 2009. *Word frequency studies*. Mouton de Gruyter.
- Popescu, I. I., K.H. Best, and G. Altmann. 2007. On the dynamics of word classes in Texts. *Glottometrics* 14: 58–71.
- Rosenfeld, B. and S. D. Penrod. 2011. *Research methods in forensic psychology*. John Wiley and Sons.
- Scott, M. 1997. PC analysis of keywords – and key key words. *System* 25 (2): 233-245.
- _____. 1999. *WordSmith tools help manual, Version 3.0*. Mike Scott and Oxford UP.
- _____. 2008. *WordSmith tools, Version 5.0*. Liverpool: Lexical Analysis Software.
- Scott, M. and C. Tribble. 2006. *Textual patterns: Keyword and corpus analysis in language education*. Benjamins.
- Steyvers, M. and T. Griffs, 2007. Probabilistic topic models. *Latent semantic analysis: A road to meaning*, ed. T. Landauer, D. McNamara, S. Dennis, and W. Kintsch. Laurence Erlbaum.
- Underhill, J. W. 2011. *Creating worldviews: Metaphor, ideology and language*. Edinburgh UP.

Corpora cited

- The Bank of English (www.mycobuild.com/about-collins-corpus.aspx)
- The British National Corpus (www.natcorp.ox.ac.uk)
- The Chinese Internet Corpus contains 280 million words (corpus.leeds.ac.uk/query-zh.html).
- The Corpus of Contemporary American English (<http://corpus.byu.edu/coca/>)
- The Czech National Corpus (<https://korpus.cz/english/struktura.php>)
- The German Reference Corpus (5400, www.ids-mannheim.de/kl/projekte/korpora/).
- The National Corpus of Polish (1500, nkjp.pl/index.php?page=0&lang=1);
- The Russian National Corpus (www.ruscorpora.ru/en/corpora-stat.html)