Content analysis and natural language processing

CS-4750 ‘Introduction to Natural Language Processing’, October 12, 3-3:50PM, EN-1001

Key points

- Objectivist and subjectivists approaches
- Content analysis: a third way?
- Content analysis and its types (qualitative, quantitative & mixed methods)
- Assessing the validity and reliability of the outcomes: mixed methods content analysis
- *ThinkMate* online platform and possible topics for course projects

Objectivist approach

- Words are viewed as having 'proper senses'
- One can understand an object entirely in terms of a *set* of its *inherent* properties; the 'truth' is objective
- For objectivists the word *digest* would have two different and distinct literal (objective) meanings – $digest_1$ for food and $digest_2$ for ideas
- The analysis of other words with which the word digest co-occurs is indicative of whether it has to be interpreted as $digest_1$ or $digest_2$

Subjectivist approach

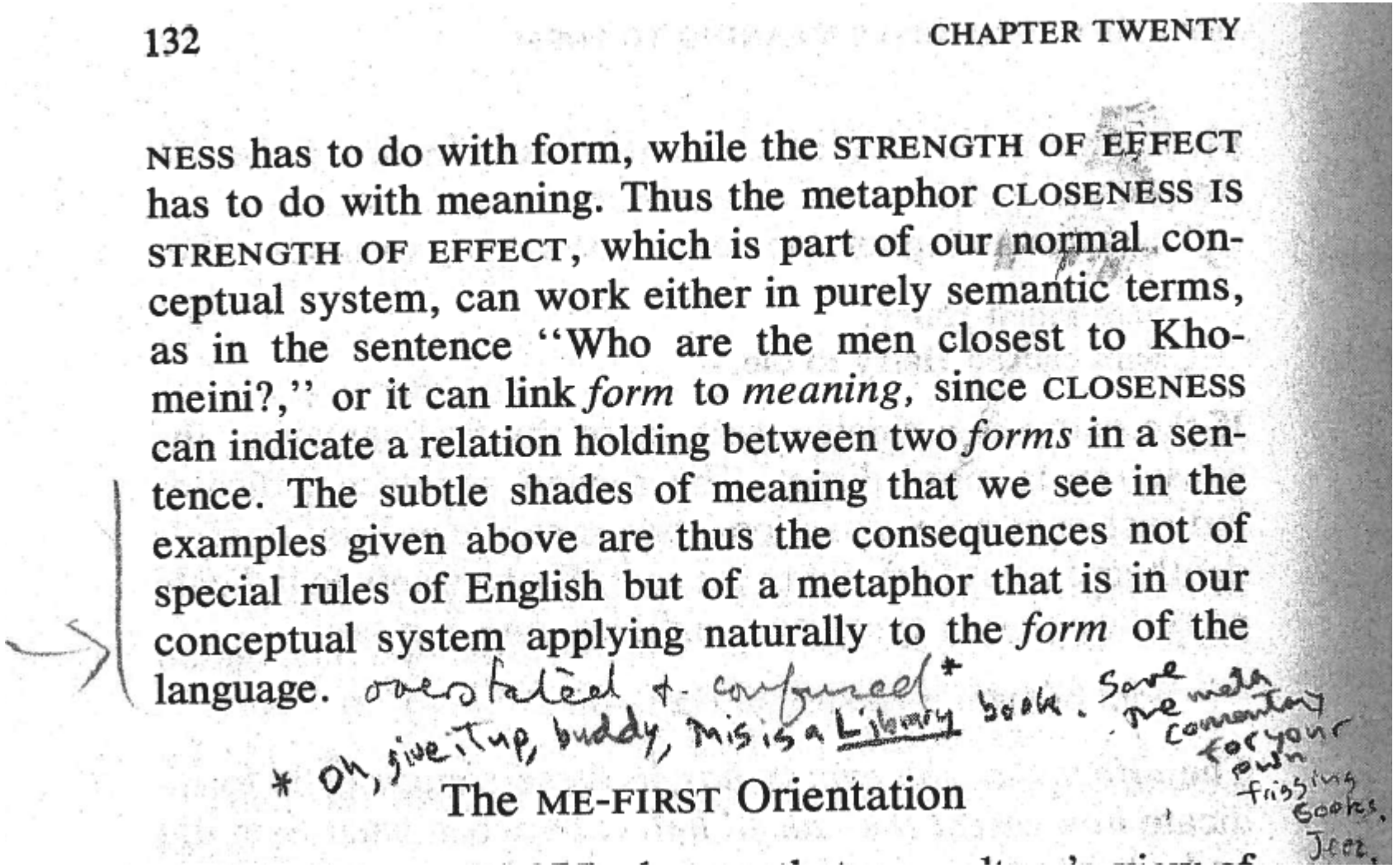
- Words are viewed as having senses attributed to them by the reader or listener
- Understanding of an object is always subjective; any 'truth' is relative and subjective
- For subjectivists, the word *digest* would have as many meanings as there are readers or listeners, $digest_i$
- It is impossible to discriminate between various meanings

Searching for a ‘third way’: Experientalism

- Lakoff, George & Johnson, Mark (1980), *Metaphors We Live By*. Chicago and London: The University of Chicago Press
- We understand experience metaphorically when we use concepts from one domain of experience (most often, our interaction with the physical environment) to structure experience in another domain. E.g., The concept ‘in’ emerges from spatial experience. The same goes for ‘at’ etc.
- From the experientalist point of view, being objective is always relative to a conceptual system

Searching for a ‘third way’: Content analysis

- **Content analysis**: the analysis of texts or images (but today I’ll focus on texts only) as a source of information about social action. It consists in ‘mining’ the text/image for relevant data. ‘Content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use’ (Krippendorff, Klaus, 2004, *Content Analysis*, Sage, p.18). My own definition of content analysis: a study of how readers interpret and comprehend texts/images
- Content analysis is a relatively ‘neutral’ research method that can be adapted to the needs of discourse analysis, critical discourse analysis, visual research methods and some other theories and research methods, including natural language processing (especially **quantitative** content analysis)


132 CHAPTER TWENTY

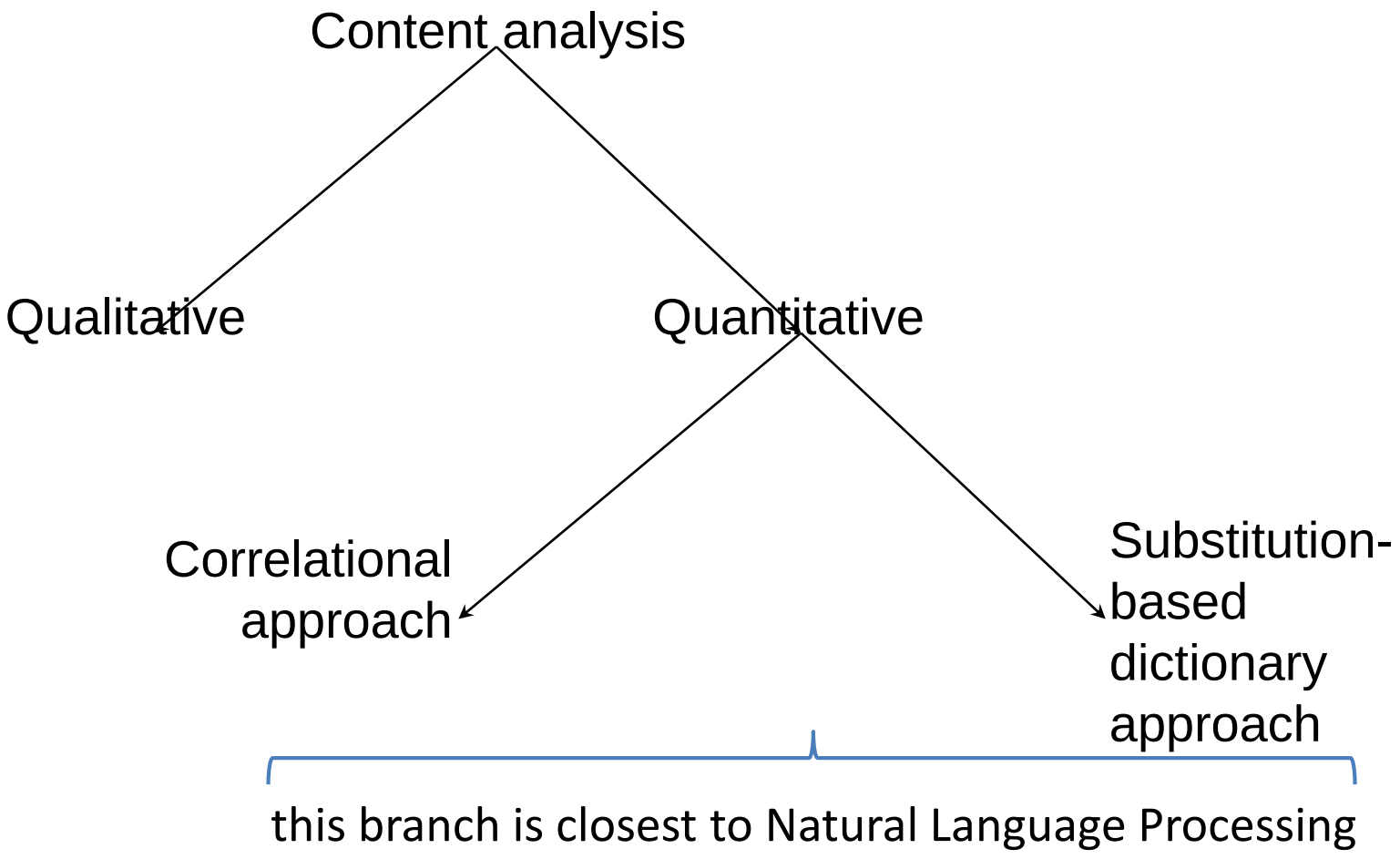
NESS has to do with form, while the STRENGTH OF EFFECT has to do with meaning. Thus the metaphor CLOSENESS IS STRENGTH OF EFFECT, which is part of our normal conceptual system, can work either in purely semantic terms, as in the sentence "Who are the men closest to Khomeini?," or it can link *form* to *meaning*, since CLOSENESS can indicate a relation holding between two *forms* in a sentence. The subtle shades of meaning that we see in the examples given above are thus the consequences not of special rules of English but of a metaphor that is in our conceptual system applying naturally to the *form* of the language. overstated + confused*

* Oh, give it up, buddy, this is a Library book. Save the meta commentary for your own frigging books. Jeez.

The ME-FIRST Orientation

Comments on the QEII copy of Lakoff & Johnson's book (at p. 132):
- The sentence 'the subtle shades…' is deemed to be 'overstated & confused' by one reader
- The other reader rebukes: 'oh, give it up, buddy, this is a library book. Save the meta commentary for your own frigging books'

Types of content-analysis



Qualitative content analysis

- **In vivo** coding: identifying relevant (to a particular research question) and simply interesting fragments of a text, image 'from scratch', i.e. the list of codes is an outcome of the coding
- Coding as manual attribution of particular meanings to fragments of a text with the help of a **codebook**. This type of qualitative content analysis requires
  - Clearly defined concepts (**Codes**) and rules for their application
  - Reliance on the coder's subjective judgment as to how to apply these rules (to reduce its subjectivity, coding is sometimes done in teams → the coefficient of inter-coder agreement, such as Krippendorff's alpha)

Quantitative content analysis

- **Correlational analysis**: the analysis of co-occurrences of words. It helps identify words that tend to co-occur in the text and form a cluster to which a particular meaning can eventually be attached (e.g., coffee and donuts)
- **Dictionary based on substitution**: a code is automatically attributed to a particular fragment if it contains a specified combination of words or phrases (e.g., code 'Food' is attributed if the fragment contains any of the following words: Bread, Tea, Broth, Breakfast, Lunch etc.

Dictionary based on substitution: an example

Human dignity (latent code)

- Politeness
- Loyalty
- Pride
- Goodness
- Love
- Courage
- Male dignity
- Responsibility
- Understanding
- Decency
- Human rights
- Self-respect
- Will power
- Conscience
- Justice
- Rank
- Respect
- Intelligence
- Honour
- Honesty

The list was identified with the help of an open-ended question asked in an online survey

Preconditions for quantitative content analysis

- **Parsing**: parts-of-speech (also known as POS, word classes, or syntactic categories) are useful because of the large amount of information they give about a word and its neighbors. Knowing whether a word is a noun or a verb tells us a lot about likely neighboring words (nouns are preceded by determiners and adjectives, verbs by nouns) and about the syntactic structure around the word (nouns are generally part of noun phrases), which makes part-of-speech tagging an important component of syntactic parsing
- **Tokenization**: segmenting running text into words and sentences
- Sentence tokenization methods work by building a binary classifier (based on a sequence of rules, or on machine learning) which decides if a period is part of the word or is a sentence boundary marker. In making this decision, it helps to know if the period is attached to a commonly used abbreviation; thus an abbreviation dictionary is useful

Vectorization

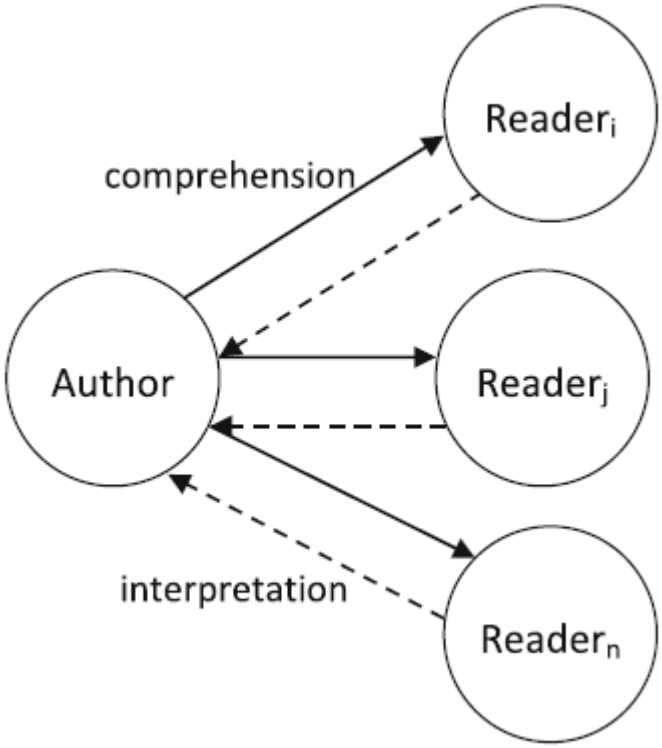
With the help of **tokenization**, the text is divided in sentences and then transformed into a table. These operations pave the way to **vectorization**. Sentences are rows in a table, variables (specific words or qualitative codes) are columns in the table

	$Word_1/Code_1$	$Word_2/Code_2$	…	$Word_i/Code_i$
$Sentence_1$				
$Sentence_2$				
$Sentence_3$				
…				
$Sentence_i$				

Types of texts

Rhetorical texts (novel, poem, diary, or essay) have a loose structure. Metaphors and analogies abound in such texts. The content analysis of rhetorical texts calls for prioritizing **interpretation**: they aim at provoking free associations and creative thinking (cf. **subjectivist** approach)

Stylistic texts (scholarly article, textbook, or scientific letter) have a clear, often rigid structure. Arguments in stylistic texts must meet high logical standards: exhaustiveness and mutual exclusiveness of categories, transitivity and consistency in their rank ordering and so forth. **Comprehension** seems to be more appropriate for the content analysis of stylistic texts as a result of their orientation toward conveying a message in the least ambiguous manner (cf. **objectivist** approach)



Triangulation: definition and forms

- ‘The combination of methodologies in the study of the same phenomenon’ (Jick, 1979)
- **Data** triangulation: the use of a variety of data sources
- **Investigator** triangulation: the use of more than one researcher
- **Theory** triangulation: using multiple perspective to interpret a single data set
- **Methodological** triangulation: the use of multiple research methods, including various forms of content analysis, to study one problem or document. **Mixed methods research** as practical manifestation

Validity and reliability in content analysis

- Mixed methods content analysis [a form of investigator and methodological triangulation] allows assessing the reliability of coding
- A conventional approach to assessing the reliability of a content analysis is to calculate coefficients of inter-coder agreement: Krippendorff's alpha, Cohen's kappa, Bennett, Alpert and Goldstein's S and some others
- Mixed methods content analysis allows adding correlation coefficients (Pearson's *r* between the qualitative content analysis outcomes and the correlational analysis outcomes, for instance) and to this list. Cf. 'In content analysis [the use of correlation coefficients] is seriously misleading' (Krippendorff, 2004, p. 245)
- The basic table takes a different shape: coders are in its rows, codes – in its columns
- The choice of a proper measure depends on particularities of the text

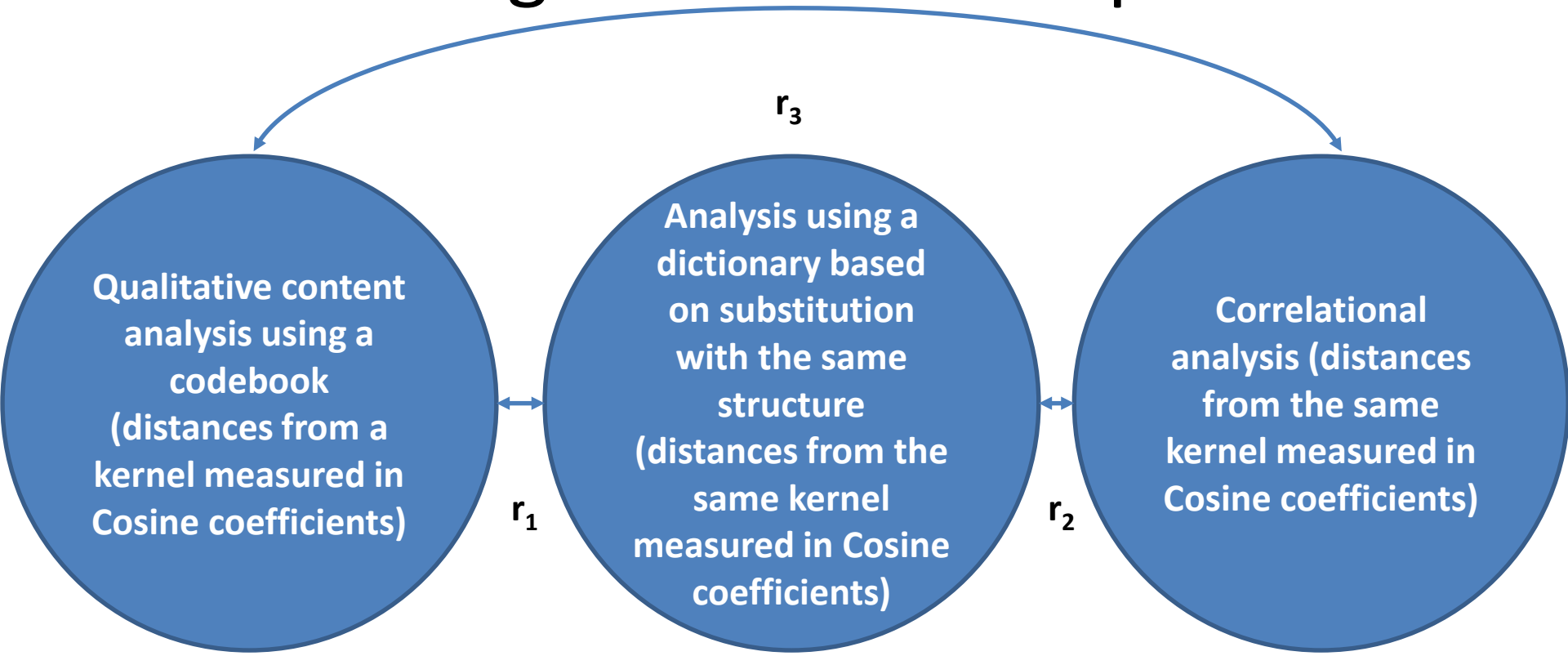
Choice of reliability measures and the type of text

Table 1 Reliability measures appropriate for particular types of text and research tasks

Type of reading	Text format	
	Stylistic	Rhetorical
Comprehension	Scott's π, Krippendorff's α, Pearson's r (between word co-interpretation occurrence and qualitative coding)	Perreault and Leigh's I_r
Interpretation	Perreault and Leigh's I_r	Bennett, Alpert and Goldstein's S, Cohen's κ, Pearson's r (between dictionary based on substitution and qualitative coding)

Pearson's *r* is calculated for distances between vectors (i.e., documents) expressed as **Cosine coefficients**

Using different types of content analysis when working on a dataset: a path-model

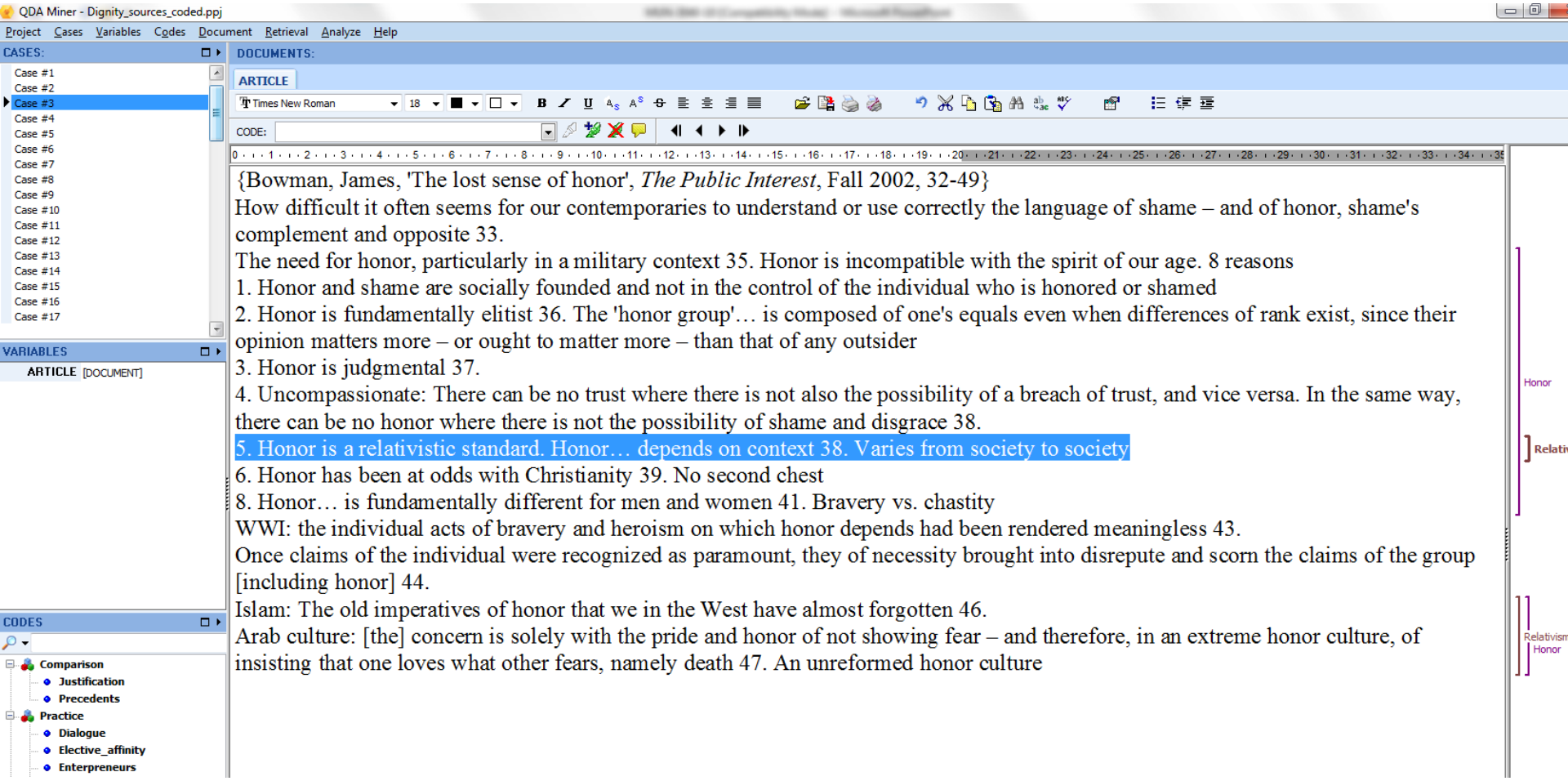


Regression analysis is also potentially possible, e.g., Results of the qualitative content analysis as a outcome variable (DV), Results of correlational analysis, Results of the analysis using a dictionary based on substitution and some controls (e.g., Time, Author's gender etc.) as predictor variables

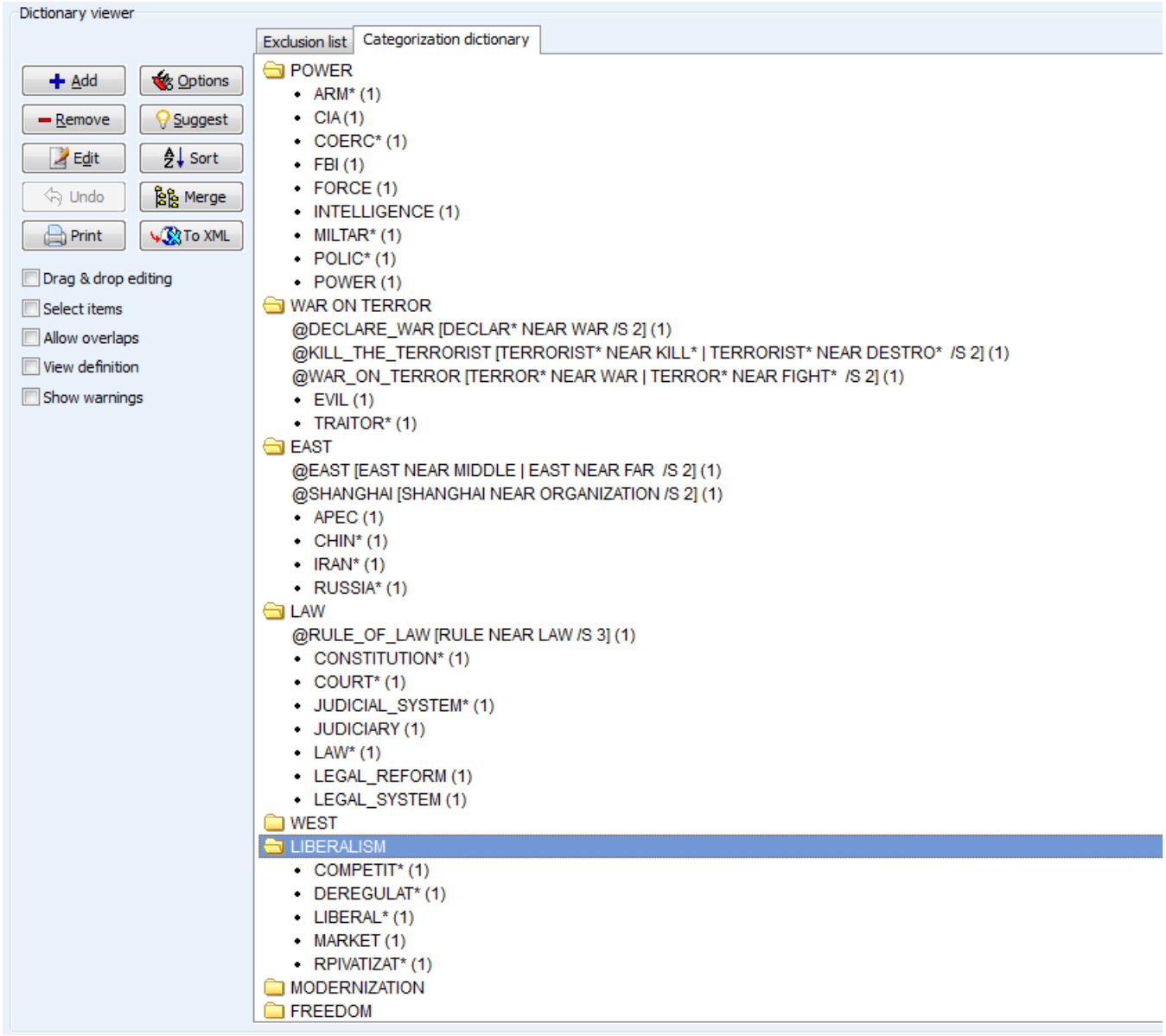
Computer programs for content analysis

- *QSR International*, Australia **http://www.qsrinternational.com/**: **NVivo**, **N6** (**NUD*IST**), **XSight**
- *Provalis Research*, Canada (Montreal) **http://www.provalisresearch.com/**: **QDA Miner** (module for qualitative content analysis) and **WordStat** (module for the analysis of co-occurrences and the use of dictionaries based on substitution). These two modules used in conjunction open multiple opportunities for mixed methods research when content analyzing texts
- **ThinkMate**: an on-line platform that I am currently working on. It has some similar features with **QDA Miner**, but also offer some new options, such as the search of similarly minded people on the basis of reading (and coding) particular texts, i.e., a novel type of social network

Qualitative coding using *QDA Miner*



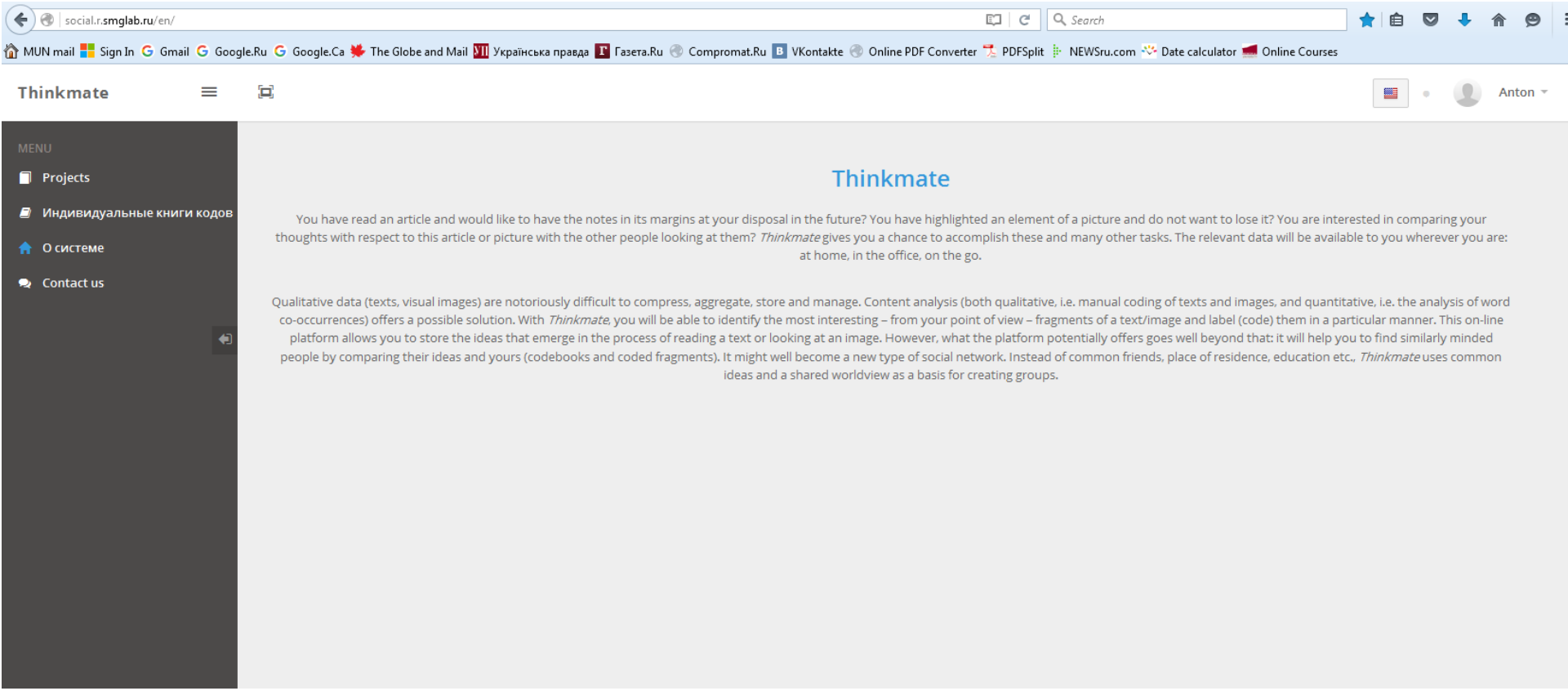
Dictionary based on substitution in *WordStat*



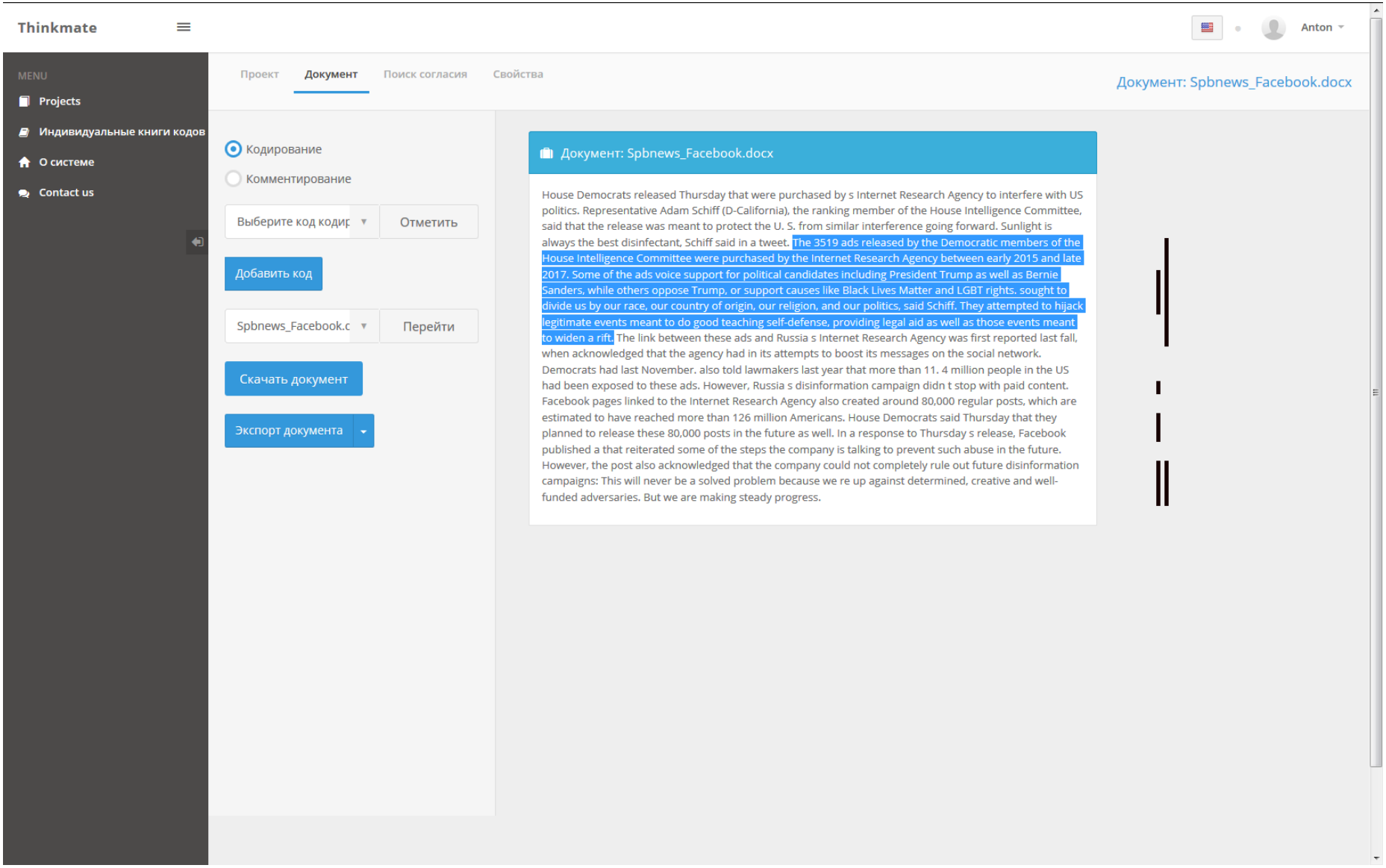
25 most frequent words used by Lord Byron (analysis of co-occurences using *WordStat*)

	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF • IDF
LOVE	891	0.80%	0.60%	0.30%	115	46.40%	297.4
EYE	719	0.60%	0.50%	0.20%	98	39.50%	289.9
MAN	718	0.60%	0.50%	0.20%	65	26.20%	417.5
HEART	668	0.60%	0.50%	0.20%	117	47.20%	217.9
DAY	529	0.50%	0.40%	0.20%	87	35.10%	240.7
TIME	461	0.40%	0.30%	0.10%	71	28.60%	250.4
LONG	460	0.40%	0.30%	0.10%	70	28.20%	252.7
LIFE	438	0.40%	0.30%	0.10%	66	26.60%	251.8
THING	422	0.40%	0.30%	0.10%	36	14.50%	353.7
SOUL	379	0.30%	0.30%	0.10%	91	36.70%	165
MAKE	368	0.30%	0.30%	0.10%	43	17.30%	280
JUAN	366	0.30%	0.30%	0.10%	2	0.80%	766.2
LEAVE	365	0.30%	0.30%	0.10%	57	23.00%	233.1
HAND	359	0.30%	0.30%	0.10%	50	20.20%	249.7
FALL	330	0.30%	0.20%	0.10%	58	23.40%	208.2
GREAT	325	0.30%	0.20%	0.10%	33	13.30%	284.7
LIE	322	0.30%	0.20%	0.10%	61	24.60%	196.1
EARTH	317	0.30%	0.20%	0.10%	57	23.00%	202.4
DIE	314	0.30%	0.20%	0.10%	54	21.80%	207.9
FEEL	305	0.30%	0.20%	0.10%	60	24.20%	188
HOUR	305	0.30%	0.20%	0.10%	73	29.40%	162
HIGH	302	0.30%	0.20%	0.10%	44	17.70%	226.8
HATH	293	0.30%	0.20%	0.10%	44	17.70%	220
HEAR	291	0.30%	0.20%	0.10%	55	22.20%	190.3

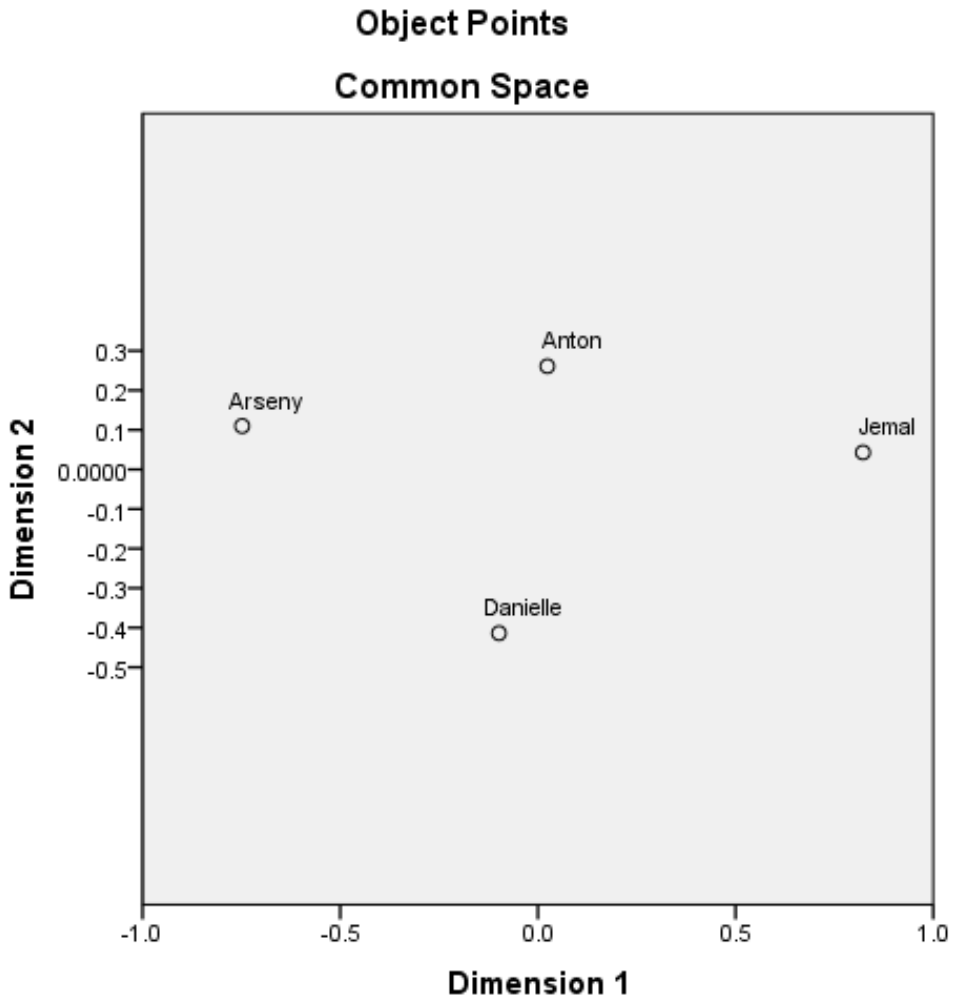
ThinkMate (beta-version)



http://social.r.smglab.ru/project/view/120/



Map of close-mindedness



Further development of *ThinkMate* (possible CS-4750 course projects)

- The **tokenization** algorithm needs improving. As of today, the algorithm is rather simplistic:
  - «dot + space + uppercase letter» OR «exclamation mark + space + uppercase letter» OR «question mark + space + uppercase letter»
- The content analysis of **images** could be done along the same lines, but relevant algorithms have to be adapted and adjusted accordingly
- The **multidimensional scaling** is not incorporated yet. Without the MDS one cannot really identify his or her similarly minded fellows
- The algorithm for **comparing the structure of the codebooks** also needs developing and so forth

Additional sources

- Oleinik, A., ‘What neural networks can’t do? On artificial creativity’ // *Big Data & Society* (under review)
- Oleinik, A., Popova, I., Kirdina, S., Shatalova T. (2016), ‘On academic reading: citation patterns and beyond’ // *Scientometrics*, 113(1), 417-435
- Oleinik, A. (2015), ‘The language of power: a content analysis of presidential addresses in North America and the Former Soviet Union, 1993–2012’ // *International Journal of the Sociology of Language*, 236, 181–204
- Oleinik, A. (2015), ‘On content analysis of images of mass protests: a case of data triangulation’ // *Quality & Quantity*, 49(5), 2203-220
- Oleinik, A., Popova, I., Kirdina, S., Shatalova T. (2014), ‘On the choice of measures of reliability and validity in the content-analysis of texts’ // *Quality & Quantity*, 48(5), 2703-2718
- Oleinik, A. (2011), ‘Mixing quantitative and qualitative content analysis: triangulation at work’ // *Quality & Quantity*, 45(4), 859-873