SCLC 2009
Corpus Workshop
Divjak & Clancy & Arppe
Introduction to Corpus Linguistics

We will cover WHY

- A bit of history: early corpus linguistics and the “Chomskyan revolution”
  - Why (not) corpus linguistics
- What is a corpus and what is in it?
  - Sampling & representativeness
  - Annotation
- Some examples
  - Lexicon: near-synonymy (RUS)
NOT how to

- for the technical details: come to one of our masterclasses or consult (among others)


1 Early Corpus Linguistics and the Chomskyan Revolution
Corpus Linguistics

- corpus (pl. corpora)
  - may be any body of text
  - but corpus today tends to mean
    “a finite collection of machine-readable text, sampled to be maximally representative of a language or variety”

  ➔ corpus linguistics = the study of language based on authentic, natural language use
Corpus Linguistics is …

- **not**
  - a branch of linguistics such as syntax, semantics, sociolinguistics, etc.
  - an aspect of language to be explained or described

**BUT**

- **about what you do with the data**
  - corpus linguistics as “nothing but a methodology”

? how to understand corpus methods with an overall view of linguistic methodology and theory
Pre-Chomsky and Post-Chomsky

- corpus linguistics recovering from marginalization due to Chomskyan lack of interest in empirical data
  - “More people live in New York than in Dayton Ohio”
  - “A corpus can’t describe a natural language entirely”
  - “Natural language is infinite”
On the nature of the data

debate between rationalists and empiricists on the nature of the data used to inform the theory? should we rely on naturally occurring observation or artificially elicited observations?

- rationalist approach
  - based on artificial behavioral data, conscious introspective judgments by native speakers or others making theoretical claims based on their reflections;
  - cognitive plausibility is a fundamental goal
  - the theory should have something to say about how the mental processing is undertaken

- empiricist approach
  - mainly concerned with naturally occurring data: find data for grammaticality of x in language y by gathering evidence from a corpus
Do we only have Chomsky to blame?

- early corpus linguistic methods, limitations, and the views of early corpus linguists are somewhat to blame for the purchase that Chomsky’s criticisms found
  - view that sentences of a natural language are finite
  - all sentences of a natural language could be collected and enumerated
  - could a corpus be the sole source of evidence in the formation of a linguistic theory; if so, would allow for the complete observation of a language by observation of the corpus

! not all early corpus linguists made such claims

  would be attractive if true since linguistic description would be a matter of objective fact and not a matter of subjective speculation
Do we only have Chomsky to blame?

- language can’t be finite because of recursion (repeating a rule indefinitely as in chaining prepositional phrases, genitives, etc.)
  - The official news agency carried excerpts from a speech by Brezhnev at a Kremlin dinner for visiting Cambodian leader Heng Samrin. (McEnery and Wilson)
  - The quilt is upstairs in the bedroom in the closet on the top shelf behind the boxes. (Langacker)
  - Фидель Кастро на пятом съезде своей компартии говорил без умолку шесть часов и сорок три минуты, что достойно рекордов Книги Гиннесса.

[Fidel Castro-NOM at fifth congress-LOC own communist-party-GEN spoke without pause-GEN six-ACC hours-GEN and forty-three minutes-ACC, that worthy records-GEN Book-GEN Guinness-GEN.]

At the fifth congress of his communist party, Fidel Castro spoke without pause for six hours and forty-three minutes, an accomplishment worthy of the Guinness Book of Records. (Janda and Clancy)
Chomsky’s objections

- could go on indefinitely → corpus could never be the sole explicandum of natural language ~ understanding grammar as a finite set of rules giving rise to an infinite number of sentences
- corpora, by nature, are incomplete and can never completely represent a language
- corpora are skewed with some sentences owing their presence in the corpus to their frequency of use in the language
- corpora are partial in the sense of their being incomplete
- even if language were finite, why wait for the finite number of sentences to enumerate themselves into a corpus when we can look at our own linguistic competence?
Fillmore’s view

Fillmore (1992)

“He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus the second word of a sentence.”
Was Chomsky right?

- Chomsky in a 1962 interview viewing the native speaker as the sole explicandum of linguistics
  
  Chomsky: The verb *perform* cannot be used with mass word objects: once can *perform a task* but one cannot *perform labor*.  
  Hatcher: How do you know, if you don’t use a corpus and have not studied the verb *perform*?  
  Chomsky: How do I know? Because I am a native speaker of the English language.

- but *perform magic* is in the BNC
  
  native speaker intuition allowed Chomsky to be wrong with an air of absolute certainty

! corpora are a means of checking up on such introspective judgments and are the only reliable sources of frequency-based information
Was Chomsky right?

- Summary of Chomskyan criticism:
  1. Corpora model the wrong thing (performance, not competence); goals of linguistics are not enumeration and description of performance, but introspection and explanation of linguistic competence
  2. Natural languages are not finite and thus cannot be fully enumerated and described
  3. Should use introspection

- Chomsky not the only problem for early corpus linguistics
  ~ Data processing was in its infancy in the pre-computer age

  E.g., Käding (1897) had searched his 11 million word corpus of German by hand with a team of human beings: slow, expensive, prone to error
Despite Chomsky

- despite Chomskyan criticism, some corpus work continued
  - phonetics relied heavily on naturally observed data
  - language acquisition research required observed data since it was not possible to inquire about metalinguistic issues with children
  - historical linguistics continued to rely on corpora
  - study of language variation had no place in the Chomskyan world

= corpus methods continued because certain topics could not be effectively studied in the artificial world of well-formedness judgments and idealized speaker-hearers
Corpus Linguists strike back

- the rationalist approach may well allow us to gather the data we want, when we want, and to gather data directly related to the system under study, the mind
- but
- corpus data renders public the point of view used to support a theory
- artificial data remains artificial
  - type of sentence typically analyzed by the formalist is far removed from the type of sentences we find in corpora
- corpora are sources of quantitative information beyond compare
  - what we do or do not find in corpora reveals important information about frequency and human beings have only the vaguest notion of frequency
- introspection may not be at all systematic, is subject to the whim of imagination
Fillmore’s view

Fillmore (1992) again

“He sits in a deep soft armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, ‘Wow, what a neat fact!', grabs his pencil, and writes something down ... having come still no closer to knowing what language is really like.”
Corpus Linguists strike back

- Corpora provide the basis of a much more systematic approach to the analysis of language
  - a more powerful methodology from the point of view of the scientific method
- Chomsky (1965): naturally occurring data is of “degenerate quality”
  - Labov (1969) claims Chomsky once stated that 95% of utterances in natural language are ungrammatical
  - Labov countered that the “great majority of utterances in all contexts are complete sentences”; the problem of ungrammatical sentences in corpora may not be so bad.
- Is quantitative data of any use?
  - all successful modern approaches to POS analysis rely on quantitative, corpus data
  - some of Chomsky’s criticisms have been useful to corpus linguistics
- the computer has turned corpus linguistics into a viable methodology
Corpus Linguistics Today

- corpora today almost always mean machine readable corpora
  - source may be written text or transcribed speech
- computers achieve total accuracy in text processing, millions of words can be analyzed with speed and accuracy
  - armies of human analysts cost money, work slowly, and make mistakes
- with appropriate annotation, can move beyond individual word forms, e.g. word and part of speech
  - yet there are plenty of tasks that cannot be undertaken due to lack of additional annotation in the text
Corpus Linguistics Today

- Computers now allow for:
  - searching for
  - retrieving
  - calculating number of occurrences
  - sorting the data in some way
  - producing a concordance, list of examples of a word KWIC (key word in context) report

- why not use a combination of empirical corpus-based methods and introspective methods?
  - elicited data and natural data can go hand in hand, provide converging evidence
A last word from Fillmore

- Fillmore (1992) yet again
  
  “I don’t think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore ... [but] every corpus I have had the chance to examine, however small, has taught me facts I couldn’t imagine finding out any other way. My conclusion is that the two types of linguists need one another.”
2 What is a corpus and what is in it?
Definition

a corpus may be any body of text, but corpus today tends to mean “a finite collection of machine-readable text, sampled to be maximally representative of a language or variety”
Sampling and Representativeness

- interest in a variety of a language, rather than an individual text or author
- construct a smaller sample of the variety, rather than an attempt at exhausting the variety
- some utterances may be excluded because of rarity, some excluded by chance, some rare utterances included by chance
- use a broad range of different authors and genres to
  - balance out the data
  - provide a general picture of the language
Finite Size Collections

- **finite corpora**
  - 1,000,000 running words of text, for instance (Brown, LOB)
  - 100,000,000 in BNC

- **monitor corpus**: an open-ended collection of texts, regularly growing (Cobuild, BYU Corpus of American English)
  - changing size and less rigorous sampling mean they are not the best sources of quantitative as opposed to qualitative data
Machine-readable Standard references

- a corpus is a standard reference for a language
  - readily available to researchers
  - continuous base of data used across studies
- may be searched and manipulated in ways that would not be possible in other formats
  - may be enriched with additional information (annotation)
Document annotation

- title
- author
- text type
- age of author
- sex of author
- date text was published or written
- variety of language
  - (US Engl, British Engl, New Zealand Engl, etc.)
- genre
- broad subject domain
  - (science, religion, detective fiction)
Element annotation: POS (1)

- tagging words in corpus with special codes (tags) to provide linguistic information
  = POS tagging (most basic type of annotation), grammatical tagging, morphosyntactic tagging
- provides fundamental basis for
  - increasing the specificity of data retrieval from corpora
  - further forms of analysis: syntactic parsing, semantic field annotation
  = basic step in the disambiguation of homographs
    - RUS est’ ‘there is’ or ‘to eat’
    - Eng, can distinguish boot as verb and noun, though not verbs meaning ‘kick’ and ‘start up computer’
Parts of Speech tagging (2)

- POS tagging has been around for a while and is highly accurate; CLAWS is 95% accurate, based on probabilistic methods.
  - In automatic tagging, conflict in annotation between providing fine distinctions that will be useful to the researcher vs. removing such distinctions to make automatic tagging more reliable.

- EAGLES has three levels of features:
  - **Obligatory**: POS: noun, verb, adjective, pronoun/determiner, article, adverb, adposition, conjunction, numeral, interjection, unique (negative *not*, INF *to*, foreign words, mathematical formulae), residual, punctuation.
  - **Recommended**: number, gender, case, type (common, proper).
  - **Optional**: generic (e.g., count/mass) or language-specific (definiteness).
Leech’s (1993) maxims for text annotation

- should be possible to remove the annotation and revert to the raw corpus
  Claire_NP1 collects_VVZ shoes_NN2._PUN
  Claire collects shoes.

- should be possible to extract the annotations for storage elsewhere (relational database, interlinear format); allow for maximum flexibility by the user

- annotation scheme should
  be based on guidelines available to the user (no need for guesswork)
  mention how and by whom the annotation was carried out
  annotation schemes should be based on widely agreed and theory-neutral principles

- end user should be made aware that the annotation is informative and potentially useful rather than infallible
  no annotation scheme has the right to be considered the standard over others
Parsing

- move to higher levels of syntactic relationships
  - parsed corpora aka treebanks: Penn Treebank project
- majority of parsing schemes “based on a form of context-free phrase structure grammar”
  - full parsing: as detailed as possible
  - skeleton parsing: less detailed, uses more coarse-grained syntactic constituent types

Also
- dependency grammars, functional grammars
- constraint grammar: marks grammatical functions rather than hierarchies of constituent phrase types

- automated parsing much less accurate than POS tagging, yet different human parsers also introduce inconsistencies
Lemmatization

- reduce all words in the corpus to a citation form as in a dictionary
  - KICK: kicked, kicks, kicking, kick
  - the variants form the lemma of KICK
  - GO: go, goes, going, gone, went
  - [this could range from straightforward, a-theoretical information to the necessity of theoretical decisions]

- lemmatization allows for easier, more comprehensive searching without need of entering all possible variants in a paradigm

- can be carried out automatically with a high degree of accuracy, but lemmatization has not been applied to a wide range of corpora thus far
Semantic annotation

- only beginning to be added to corpora
  - semantic relationships: agent, patient
  - semantic features: word senses
- subset of Brown corpus, WordNet is word-sense-tagged
  - one system used is top-level categories with subdivisions
    - *eg crown* as
      - Man:Bodily Being:General Human Needs:Headgear
- [also a matter of marking items on the grammatical/lexical continuum]
Discourse and text linguistic annotation

- discourse tags: discourse management rather than propositional content
  - apologies (*sorry, excuse me*)
  - hedges (*kind of, sort of*)
  - greetings (*hello, good evening*)
  - politeness (*please*)
  - general purpose responses (*really, that’s right*)

- identifying these items can be controversial as they are context-embedded and different annotators make different judgments
Other

- Phonetic transcription
  - narrow (precise phonetic details) or broad (phonemic level)
  - narrow requires skilled annotators; broad may be done automatically

- Prosodic information to corpus annotation
  - mostly information on stress, intonation, and pauses

- Anaphoric annotation
  - at a minimum: the what and who of pronoun reference
  - cohesion- means of connecting elements through use of pronouns, repetition, substitution, etc.
  - bootstrapping problem: human annotators must create training data for later computer programs
Multilingual Corpora

- parallel corpora
  - same text in multiple languages
  - an old idea: polyglot bibles with Hebrew, Greek, Latin, vernacular languages
  - parallel aligned corpus is aligned for sentence-level content across the languages
  - useful for research in machine translation, 2nd-language teaching, contrastive linguistics
  - sentence alignment can be carried out automatically with high degree of accuracy; also some attempts at word or multi-word unit alignment in addition to sentence alignment
Multilingual Corpora

- translation corpora (comparable corpora)
  - not the same text in translation, but sample texts from same domains and genres
  - provides authentic L1 texts in both languages, rather than artificial and error-prone L2 texts often found in parallel corpora
  - [How about a massive, cross-linguistic Harry Potter 1-7 corpus?]
More to be found here (1)

- Gateway to corpus linguistics
  http://www.corpus-linguistics.com/
- Bookmarks for corpus linguists
  http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm
- corpora discussion list
  http://torvald.aksis.uib.no/corpora/
international peer-reviewed journals dedicated to corpus linguistics, for example,

- **Corpora**,  
- **Corpus Linguistics and Linguistic Theory**,  
- **ICAME Journal** and the  
- **International Journal of Corpus Linguistics**.
Old Russian/Slavic corpora

- XIth Century Slavic texts (http://www.hf.ntnu.no/SofiaTrondheimCorpus)
- Old South Slavic (http://www.rcf.usc.edu/~pancheva/ParsedCorpus.html)
More to be found here (4)

- **Russian Corpora**
  - Amsterdam Russian Corpus of A.A. Barentsen (write to A.A.Barentsen@uva.nl for more information)
  - Russian National Corpus [http://www.ruscorpora.ru](http://www.ruscorpora.ru) (including spoken Russian)
  - Serge Sharoff’s corpora ([http://corpus.leeds.ac.uk/ruscorpora.html](http://corpus.leeds.ac.uk/ruscorpora.html))
  - Tuebingen Russian Corpora ([http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html](http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html))
  - Uppsala Corpus [http://www.slaviska.uu.se/korpus.htm](http://www.slaviska.uu.se/korpus.htm)
More to be found here (5)

- **Other Slavic Corpora**
  - Bosnian: Oslo Corpus of Bosnian Texts ([http://www.tekstlab.uio.no/Bosnian/Corpus.html](http://www.tekstlab.uio.no/Bosnian/Corpus.html))
  - Bulgarian: Bulgarian National Corpus ([http://www.ibl.bas.bg/BGNC_bg.htm](http://www.ibl.bas.bg/BGNC_bg.htm))
  - Croatian: Croatian National Corpus ([http://www.hnk.ffzg.hr/](http://www.hnk.ffzg.hr/))
  - Polish:
    - Polish National Corpus ([http://nkjp.pl/](http://nkjp.pl/))
    - IPI PAN Polish Corpus ([http://korpus.pl/](http://korpus.pl/))
    - PELCRA corpora ([http://pelcra.ia.uni.lodz.pl/corpora_pl.php](http://pelcra.ia.uni.lodz.pl/corpora_pl.php))
  - Serbian ([http://www.serbian-corpus.edu.rs/ns/preview/preview.htm](http://www.serbian-corpus.edu.rs/ns/preview/preview.htm)) ! Preview only!
More to be found here (6)

- Parallel Corpus of Slavic and Other Languages ([http://www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Slavistik/RPC/](http://www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Slavistik/RPC/))
More to be found here (7)

- Some Repositories of (old/new) Russian Electronic Texts
  - the well-known text repository Библиотека Максима Мошкова at [http://www.lib.ru](http://www.lib.ru)
  - a virtual library at [http://www.rvb.ru/](http://www.rvb.ru/)
  - about 1mill words of Old Russian in the texts at [http://lib.pushkinskijdom.ru](http://lib.pushkinskijdom.ru) and [http://titus.uni-frankfurt.de/indexe.htm?/texte/texte2.htm#aruss](http://titus.uni-frankfurt.de/indexe.htm?/texte/texte2.htm#aruss)

- Anything else? Send it to d.divjak@sheffield.ac.uk
3 An example: lexical variation
A clustered model for near-synonymy
Don’t You Try ...

But Sirota was still trying/making efforts to say something, and again it was impossible to understand a word of what he was saying. Finally, Malinin could not take it any longer and put an end to this mutual torture: “Don’t you try/endeavor, Sirota, I can’t understand you anyway: your mouth got smashed .... There is only sound, no voice. You’ll be in hospital for a while – it will heal, but for now don’t try, don’t torture yourself” (...) [K. Simonov. Živye i mertvye]

Prague, 15 October 2009
Ways of Trying in Russian

- case study on 9 verbs that mean ‘try’:
  пробовать, пытаться, стараться, силиться, норовить, порываться, тщиться, пыжиться, тужиться

- aim
  ■ contribute to relatively **little studied area**
  ■ illustrate **general line of research**: investigate how language data of different degrees of abstractness, typically grammar and lexicon, interact in conveying meaning differences
  ■ demonstrate direct **relevance** of linguistic research for e.g. language teaching

Prague, 15 October 2009
Outline (ct’d)

- Problems of near-synonym research
  - **Delineation** of the category:
    - 9 verbs that exhibit similar syntactical behavior and express similar semantic content
  - **Internal structuring** of the category:
    - Is there an internal structure? What does it look like?
    - Which verbs are more/less similar and cluster together?
  - **Description** of elements in the category:
    - Is there a verifiable way to describe scales of variation?
    - Which properties discriminate best between semantically similar verbs?

Prague, 15 October 2009
Ways of Investigating Near-Synonymy

- traditional analyses
  - minimal pair approach: possible vs impossible
  - rely on elicited data (! introspection) ~ dictionaries
    (Apresjan 1999; also Černova 1996, Evgen’eva 2001)
    - do not list the same verbs
    - do not group verbs the same way
    - do not interpret verbs the same way

- corpus-based analysis
  - multivariate approach: more or less probable
  - uses non-elicited data (1585 ex, 100-250 ex per verb)
    - from Amsterdam Corpus, Russian National Corpus: 1950-2000 Russian literature, written
    - supplemented with data from Internet where needed
  → “quantify” meaning: distribution // meaning
Parameters

→ 85 **ID tags** (Atkins 1987): operational definitions, sentence level

- **Finite verb**: aspect, mode, tense
- **Subject**: case + type of subject (9)
  - animate (human being vs animal) vs inanimate (abstract vs concrete, man-made vs non-man made etc.)
- **Infinitive**: aspect + degree of control (low, medium, high) + type of action (15)
  - physical action, perception, communication, intellectual activity, emotions etc.
- **Collocates**: adverbs, particles and connectors, negation
- **Clause type**: main vs subclause, declarative vs imperative vs interrogative vs exclamative

= **Behavioral profile** (Hanks 1996) on denotational level (cf. Edmonds & Hirst 2002)

Prague, 15 October 2009
Солидные люди носят котелки. Чтобы замаскировать свои босоножки, он спускал штаны пониже и старался не двигаться.
<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2,040,816,327</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>1,619,433,198</td>
<td>1,612,903,226</td>
<td>1,652,892,562</td>
<td>0</td>
<td>5,660,377,358</td>
<td>1,020,408,163</td>
<td>0,840,336,134</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0,413,223,144</td>
<td>0</td>
<td>1,020,408,163</td>
<td>0</td>
<td>0,8</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>0,404,858,3</td>
<td>0,403,225,806</td>
<td>1,652,892,562</td>
<td>4,166,666,667</td>
<td>9,43,962,264</td>
<td>7,142,857,143</td>
<td>1,680,672,269</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>0,404,858,3</td>
<td>0</td>
<td>0</td>
<td>0,123,966,421</td>
<td>1,886,792,453</td>
<td>2,040,816,327</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1,886,792,453</td>
<td>0</td>
<td>0,840,336,134</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1,239,666,421</td>
<td>1,020,408,163</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5,555,555,556</td>
<td>3,773,584,906</td>
<td>4,081,632,653</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>0,803,165,999</td>
<td>0,403,225,806</td>
<td>0,413,223,144</td>
<td>4,166,666,667</td>
<td>5,660,377,358</td>
<td>6,122,489,88</td>
<td>0,840,336,134</td>
</tr>
<tr>
<td>17</td>
<td>16</td>
<td>0,809,716,399</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1,886,792,453</td>
<td>1,020,408,163</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1,020,408,163</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>3,220,060,398</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2,777,777,778</td>
<td>7,547,169,811</td>
<td>3,061,224,493</td>
</tr>
<tr>
<td>20</td>
<td>21</td>
<td>15,384,153</td>
<td>9,27,419,358</td>
<td>6,611,570,248</td>
<td>12,5</td>
<td>22,641,509,43</td>
<td>11,224,898</td>
<td>8,403,361,345</td>
</tr>
<tr>
<td>21</td>
<td>22</td>
<td>4,852,999,956</td>
<td>5,645,161,293</td>
<td>0,826,446,201</td>
<td>2,777,777,778</td>
<td>5,620,377,358</td>
<td>3,061,224,493</td>
<td>7,563,002,51</td>
</tr>
<tr>
<td>22</td>
<td>23</td>
<td>7,632,076,92</td>
<td>7,285,086,416</td>
<td>6,193,437,107</td>
<td>4,166,666,667</td>
<td>7,547,169,811</td>
<td>5,102,040,816</td>
<td>4,201,680,672</td>
</tr>
<tr>
<td>23</td>
<td>24</td>
<td>4,852,999,956</td>
<td>1,612,903,226</td>
<td>3,719,330,826</td>
<td>2,777,777,778</td>
<td>0</td>
<td>1,020,408,163</td>
<td>4,201,680,672</td>
</tr>
<tr>
<td>25</td>
<td>26</td>
<td>7,267,493,398</td>
<td>7,285,086,416</td>
<td>8,264,462,801</td>
<td>22,222,222,222</td>
<td>0</td>
<td>0</td>
<td>1,680,672,269</td>
</tr>
<tr>
<td>26</td>
<td>27</td>
<td>8,383,308,097</td>
<td>1,380,898,889</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3,613,445,384</td>
</tr>
<tr>
<td>27</td>
<td>28</td>
<td>10,124,157,49</td>
<td>6,048,367,097</td>
<td>3,719,089,014</td>
<td>9,444,444,444</td>
<td>1,886,792,453</td>
<td>7,142,857,143</td>
<td>2,521,008,403</td>
</tr>
<tr>
<td>28</td>
<td>29</td>
<td>0,809,716,999</td>
<td>4,032,256,085</td>
<td>1,239,666,421</td>
<td>2,777,777,778</td>
<td>0</td>
<td>0</td>
<td>2,521,008,403</td>
</tr>
<tr>
<td>29</td>
<td>30</td>
<td>0</td>
<td>2,016,129,032</td>
<td>2,479,338,843</td>
<td>3,888,888,889</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>31</td>
<td>2,834,080,897</td>
<td>7,285,086,416</td>
<td>5,371,900,826</td>
<td>3,888,888,889</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>32</td>
<td>14,170,040,49</td>
<td>8,870,967,742</td>
<td>38,016,528,93</td>
<td>16,666,667</td>
<td>9,183,673,496</td>
<td>15,994,339,62</td>
<td>2,521,008,403</td>
</tr>
<tr>
<td>32</td>
<td>33</td>
<td>0,404,858,3</td>
<td>0,403,225,006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0,840,336,134</td>
</tr>
<tr>
<td>33</td>
<td>34</td>
<td>13,765,102,19</td>
<td>12,096,774,19</td>
<td>12,396,842,1</td>
<td>12,5</td>
<td>13,207,451,77</td>
<td>23,469,387,77</td>
<td>22,680,975,63</td>
</tr>
<tr>
<td>34</td>
<td>37</td>
<td>88,663,967,61</td>
<td>84,67,749,135</td>
<td>46,694,248,088</td>
<td>73,611,611</td>
<td>86,792,452,83</td>
<td>88,775,510,2</td>
<td>97,478,991</td>
</tr>
<tr>
<td>35</td>
<td>38</td>
<td>8,097,165,992</td>
<td>10,887,096,77</td>
<td>16,115,072,48</td>
<td>22,222,222,222</td>
<td>7,547,169,811</td>
<td>8,163,265,306</td>
<td>0,840,336,134</td>
</tr>
<tr>
<td>36</td>
<td>39</td>
<td>3,238,862,327</td>
<td>4,435,453,871</td>
<td>37,190,002,624</td>
<td>4,166,666,667</td>
<td>5,660,377,358</td>
<td>3,061,224,49</td>
<td>2</td>
</tr>
<tr>
<td>37</td>
<td>42</td>
<td>2,429,149,798</td>
<td>1,612,903,226</td>
<td>0</td>
<td>1,388,888,889</td>
<td>1,886,792,453</td>
<td>5,102,040,816</td>
<td>1,680,672,269</td>
</tr>
</tbody>
</table>
Techniques (cf. Backhaus et al. 1996) (with St.Th.Gries)

- 9 verbs - 85 ID tags - 1585 examples
  - Co-occurrence table: +/-135 000 data points

→ Find structure: cluster analysis
  - Exploratory, hypothesis generating technique
    - hierarchical agglomerative cluster analysis
      - compares elements, groups similar elements together
      - use canberra metric as (dis)similarity measure, Ward’s rule as amalgamation strategy

Prague, 15 October 2009
Пробовать
Пытаться
Стрелять
Наблюдать
Писать
Считать
Пытаться
Тыкать
Проникать
Норовить
Techniques (ct’d)

- **Describe structure**: $t$-values
  - Variables that discriminate between clusters
    - high $t$-values: over-represented
    - low $t$-values: under-represented
  - contrast clusters on basis of over- vs under-represented variables
An inanimate subject (concrete or abstract)

A human being is exhorted to undertake an attempt to move or to make someone move (rather than to undertake mental activities); often, these activities are negated.

[you could succeed]

pytat'sja
starat'sja

[you can't succeed]

tscit'sja
pyzit'sja
tuzit'sja

[you won't succeed]

silit'sja
porvbat'sja
norovit'
Techniques (ct’d)

→ Visualize radial category structure
  ■ Translate BP into radial tree
    □ Manually?
    □ Phylogenetic clustering: outputs tree-like graph that resembles radial category
      ▪ uses canberra metric as (dis)similarity measure, Fitch-Margoliash as amalgamation strategy

Prague, 15 October 2009
Prague, 15 October 2009
Techniques (ct’d)

- Scales of variation: overlapping, yet distinct, lexical and conceptual systems

→ Describe elements in category: z-scores

- Variables that discriminate between verbs in clusters
  - high z-scores: over-represented
  - low z-scores: under-represented
→ contrast verbs on basis of over- vs under-represented variables

Prague, 15 October 2009
avoids controllable actions; passive perception, figurative motion “other”; negated infinitives (avoiding an action); repeated duration and reduced intensity

avoids weakly controllable actions; figurative motion, mental activities; repeated attempts

abstract concepts; non- or weakly controllable actions; perception, making someone move, mental activities; intense attempt undertaken in vain

each try presented as completed experiment; exhortative particles; (permissive) restriction; reasons for failure external or internal

[YOU COULD SUCCEED]

probovat'

[YOU CAN'T SUCCEED]

norovit'

tscit'sja

tuzat'sja

Pyžat'sja

Prague, 15 October 2009
But Sirota was still trying [intense attempt in vain] to say something, and again it was impossible to understand a word of what he was saying. Finally, Malinin could not take it any longer and put an end to this mutual torture: “Don’t you try [relatively intense, durative attempt that implies repetition], Sirota, I can’t understand you anyway: your mouth got smashed .... There is only sound, no voice. You’ll be in hospital for a while – it will heal, but for now don’t try [experimental, repeated attempt], don’t torture yourself” (...)

Prague, 15 October 2009
Conclusion: corpus-based solutions

- Corpus-based quantitative ‘behavioral profile’
  - objective & verifiable solution to problems of near-synonym research
    - internal structuring
      - identify patterns within sets of words
    - description of clusters
      - identify shared semantic properties
    - description of elements in clusters
      - identify individual semantic properties
  - provides cognitively realistic language description
Literature

Exploratory univariate analysis – why?

- Univariate analysis allows one to see
  - the individual effects of each studied property concerning the studied phenomenon
  - in isolation
- Specific benefits in comparison to multivariate analysis
  - Univariate methods are less constrained by restrictions in multivariate analysis concerning the number of variables one can scrutinize
  - Relatively easy to calculate
Univariate analysis – why not?

- However, linguistic phenomena are inherently influenced and determined by a multitude of variables working together at the same time → cry out for multivariate statistical methods
- individual features are often pervasively intercorrelated
  - researchers can be and have been tempted to reduce the phenomena that they study into monocausal theories
  - though such simple explanations are mostly inadequate (Bresnan et al. 2007; Gries 2003a: 32-36)

Prague, 15 October 2009
Univariate exploratory analysis – how?

- Observing the **homogeneity/independence** (or heterogeneity/dependence) of the distribution of one or more properties among various alternative outcomes/categories/classes
  - Fundamentally: (in)dependence of property **in relation to** outcomes/classes
- Assessing the degree of **association** (≈correlation) between one or more properties among various alternative outcomes/categories/classes
Some statistical concepts

- Population
- Sample
- Distribution
- Homogeneity vs. heterogeneity
- Independence vs. dependence
- Level of significance
- Degrees of freedom
Example case – can one TRY in different ways in Russian?

- Do Russian near-synonymous TRY verbs differ with respect to the observable properties that they are associated with?
- If this is the case generally, what are the individual preferences wrt to the studied properties among the studied verbs, and vice versa?
Population vs. sample

- **Population**
  - Problem: what exactly constitutes a(n entire) population in linguistic research
  - In principle: all the occurrences of TRY verbs in the Russian language (use)
  - In practice: all the occurrences of TRY verbs in available Russian (written) corpora

- **Sample**
  - the data (or its subset) we have (collected or otherwise available) the data (or its subset) we have (collected according to some criteria or which is otherwise available)
  - How **representative** is the sample of the population?
Starting point – distributions

- compile for each studied property a contingency table representing the distribution of the particular property among the studied verbs in the data
  - also called a cross-classification or crosstabulation of the studied property and the verbs (Agresti 2002: 36-38)
- A property-specific contingency table contrasts
  - the frequency of the studied property with each verb (in the first row)
  - against the occurrences of each verb without the studied property (in the second row)
FINITE + ASPECT:IMPERFECTIVE

- Inherent property of TRYing
- $\Sigma$(IMPERFECTIVE) = $\frac{1075}{1351}$

<table>
<thead>
<tr>
<th>Verb/Property</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPERFECTIVE</td>
<td>241</td>
<td>119</td>
<td>250</td>
<td>206</td>
<td>197</td>
<td>62</td>
</tr>
<tr>
<td>~IMPERFECTIVE (PERFECTIVE)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>50</td>
<td>184</td>
</tr>
</tbody>
</table>
FINITE +
ASPECT:IMPERFECTIVE

<table>
<thead>
<tr>
<th>Property/Proportion</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Property-wise (%)</td>
<td>22</td>
<td>11</td>
<td>23</td>
<td>19</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>Lexeme-wise (%)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>83</td>
<td>80</td>
<td>25</td>
</tr>
</tbody>
</table>
Collostructional analysis

  - Properties are studied here only to the extent that they occur with the selected verbs
  - Regardless of how often a property may occur with other verbs besides the selected ones, these occurrences will not be considered
Homogeneity/heterogeneity of distribution

- Whether a studied property is distributed evenly and uniformly among the studied verbs or not, and
- What is the magnitude of the possible overall and verb-specific deviations from evenness
- Are the overall and/or verb-specific deviations (statistically) significant?
How this is done?

- Calculate the expected values
  - Representing a homogeneous distribution
- Calculate cellwise the squared differences of expected and actually observed values divided by expected values
- Sum up these cellwise values
- Compare this value with a theoretical distribution
  - e.g. Chi-squared ($\chi^2$) distribution
  - which represents the level of significance, i.e. the probability of sampling the observed distribution from an underlying population which would in fact be homogeneous
Expected values

- \( E_{i,j} = \frac{(R_i \cdot C_j)}{N} \)
  - where \( i \) indicates the row and \( j \) the column indexes
  - \( R_i \): marginal row total of Row \( i \)
  - \( C_j \): marginal column total of Column \( j \)
  - \( N \): overall total frequency of the table

- \( R(\text{IMPERFECTIVE}) = 1075 \)
- \( C(\text{silitsja}) = 241 \)
- \( E(\text{IMPERFECTIVE, silitsja}) \)
  \( = 1075 \cdot 241 / 1351 = 191.8 \)
### Expected values

<table>
<thead>
<tr>
<th>Property/Verb</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPERFECTIVE</td>
<td>191.8</td>
<td>94.7</td>
<td>198.9</td>
<td>197.3</td>
<td>196.5</td>
<td>195.7</td>
<td>1075</td>
</tr>
<tr>
<td>PERFECTIVE</td>
<td>49.2</td>
<td>24.3</td>
<td>51.1</td>
<td>50.7</td>
<td>50.5</td>
<td>50.3</td>
<td>276</td>
</tr>
<tr>
<td>Column total</td>
<td>241</td>
<td>119</td>
<td>250</td>
<td>248</td>
<td>247</td>
<td>246</td>
<td>1351</td>
</tr>
</tbody>
</table>

- **N.B. NOT(IMPERFECTIVE) ↔ PERFECTIVE**
### Expected vs. Observed values

<table>
<thead>
<tr>
<th>Property/Verb</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPERFECTIVE</td>
<td>191.8</td>
<td>94.7</td>
<td>198.9</td>
<td>197.3</td>
<td>196.5</td>
<td>195.7</td>
<td>1075</td>
</tr>
<tr>
<td></td>
<td>241</td>
<td>119</td>
<td>250</td>
<td>206</td>
<td>197</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>PERFECTIVE</td>
<td>49.2</td>
<td>24.3</td>
<td>51.1</td>
<td>50.7</td>
<td>50.5</td>
<td>50.3</td>
<td>276</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>50</td>
<td>184</td>
<td></td>
</tr>
<tr>
<td>Column total</td>
<td>241</td>
<td>119</td>
<td>250</td>
<td>248</td>
<td>247</td>
<td>246</td>
<td>1351</td>
</tr>
</tbody>
</table>

- N.B. Among the 6 selected TRY verbs
Cellwise differences

<table>
<thead>
<tr>
<th>Property/Verb</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPERFECTIVE</td>
<td>+49</td>
<td>+24</td>
<td>+51</td>
<td>+9</td>
<td>0</td>
<td>-134</td>
</tr>
<tr>
<td>PERFECTIVE</td>
<td>-49</td>
<td>-24</td>
<td>-51</td>
<td>-9</td>
<td>0</td>
<td>+134</td>
</tr>
</tbody>
</table>

- Are these cellwise deviances on the whole significant?
- Are some cellwise deviances significant on their own?
Evaluation of overall significance of degree of deviation from homogeneity

- Pearson statistic: \( \chi^2 \)
- \( \chi^2 = \sum_i \sum_j [(O_{i,j} - E_{i,j})^2 / E_{i,j}] \)
- \( \chi^2(\text{PERFECTIVE, silitśja}) = (241 - 191.8)^2 / 191.8 = 12.62065 \)
- Compare the sum of these cellwise Pearson contributions \( (\chi^2) \) with the theoretical value \( (\chi^2) \)
  - With degrees of freedom
    - \( df = 5 \leftarrow (6-1) \cdot (2-1) \)
  - \( \chi^2(df=5) = 11.07 \)
Cellwise $X^2$ contributions

<table>
<thead>
<tr>
<th>Property/Verb</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPERFECTIVE</td>
<td>13</td>
<td>6</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>91</td>
</tr>
<tr>
<td>~IMPERFECTIVE</td>
<td>49</td>
<td>24</td>
<td>51</td>
<td>1</td>
<td>0</td>
<td>356</td>
</tr>
</tbody>
</table>

- $X^2 = 13+6+13+0+0+91+49+24+51+1+0+356 = 604$
- $604 >> 11.07 \rightarrow$ overall deviances significantly greater than the minimum set by the theoretical chi-squared distribution
- $P(\chi^2=604, df=5)=0 << 0.05 \rightarrow$ corresponding P-level clearly under critical value
Level of significance

- also known as the **P-value** or **alpha (α)**
- indicates the probability that the observed values in the contingency table could have been sampled by chance from the assumed underlying population
- In the human science disciplines, the **critical P-value** or **critical alpha**, that is required for an observation to be considered statistically significant, is conventionally (and N.B. quite arbitrarily) set at **P<0.05** (e.g., Howell 1999: 128-129)
What does $P < 0.05$ mean?

- This particular critical P-value entails that there is a 5% risk (or chance) that the observations in question could have been sampled from a known/homogeneous population by chance.
- In other words, 1 in every 20 samples of observations with this particular P-level is:
  - more likely to be the result of random sampling variation;
  - than representative of a real lack of independence (i.e. heterogeneity) in the assumed underlying population of which the observations are a sample.
The observed distribution of the IMPERFECTIVE aspect of the FINITE verb exhibits overall deviances which are substantially greater than the minimum $\chi^2$ value for the critical P-level ($P<0.05$).

Thus, the probability of observing by chance the distribution in the data sample – assuming that the underlying distribution were in fact homogeneous – is extremely low ($P \to 0$).

N.B. Provided that the sample is representative (again: of what?)
Therefore

- the observed distribution can be interpreted as **heterogenous**
- The property IMPERFECTIVE and the TRY verbs are *not independent* of each other

What about the significance of individual cells?

→ Followup tests
Cellwise $X^2$ contributions and deviations revisited

<table>
<thead>
<tr>
<th>Property/Verb</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPERFECTIVE</td>
<td>13 (+)</td>
<td>6</td>
<td>13 (+)</td>
<td>0</td>
<td>0</td>
<td>91 (-)</td>
</tr>
<tr>
<td>~IMPERFECTIVE</td>
<td>49 (-)</td>
<td>24 (-)</td>
<td>51 (-)</td>
<td>1</td>
<td>0</td>
<td>356 (+)</td>
</tr>
</tbody>
</table>

- We can consider individual cells as significantly deviant which on their own exceed the overall critical theoretical $\chi^2(P=0.05, df=5) = 11.07$
- However, there are also less stringent minimum required levels for significant cellwise deviance
Verb-specific significant preferences with respect to the IMPERFECTIVE

<table>
<thead>
<tr>
<th>Property/Verb</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPERFECTIVE</td>
<td>+</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PERFECTIVE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

We can extend this analysis to groups of individual properties belonging together, other than ASPECT, e.g. TENSE, semantic characterizations of the INFINITIVE, or degrees of CONTROL exhibited (N.B. In relation to each other, not verbs in general → thus exaggerates!)
Semantic characterizations of the INFINITIVE

<table>
<thead>
<tr>
<th>Property/Verb</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMMUNICATION</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EMOTION</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EXCHANGE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MENTAL</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>META_MOTION</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>META_MOTION_Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>META_PHYSICAL</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>META_PHYS_EXCHANGE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>META_PHYS_OTHER</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MOTION</td>
<td>+</td>
<td>+</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>MOTION_OTHER</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PERCEPTION</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>PHYSICAL</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PHYSICAL_OTHER</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
## Degrees of CONTROL concerning the INFINITIVE

<table>
<thead>
<tr>
<th>Property/Verb</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTROL_HIGH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>CONTROL_MEDIUM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CONTROL_LOW</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>+</td>
<td>0</td>
</tr>
</tbody>
</table>
Degrees of CONTROL concerning the INFINITIVE

<table>
<thead>
<tr>
<th>Property/Verb</th>
<th>silitsja</th>
<th>poryvatsja</th>
<th>norovit</th>
<th>staratsja</th>
<th>pytatsja</th>
<th>probovat</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTROL_HIGH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>CONTROL_MEDIUM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CONTROL_LOW</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>+</td>
<td>0</td>
</tr>
</tbody>
</table>
BE in Slavic

For updated information, see http://home.uchicago.edu/~sclancy
BE in Czech

**EXISTENCE**

**LOCATION/POSITION**

**PRESENCE/ABSENCE**

**AUXILIARY**

**IMPERSONAL**

**COPULA**

**BE**
BE in Polish
BE in Bulgarian

EXISTENCE

LOCATION/POSITION

PRESENCE/ABSENCE

AUXILIARY

IMPERSONAL

COPULA

BE
BE in Russian

- **Existence**
  - **Location/Position**
  - **Presence/Absence**

- **Auxiliary**
  - **Impersonal**

- **COPULA**

- **Other copulas**

- **SUŠČESTVOVAT**

- **NAXODIT’JA**

- **EST’/NE EST’**
BE in Slavic (MDS-OC)

For updated information, see http://home.uchicago.edu/~sclancy
# BEING-BECOMING-UNBECOMING

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>BECOMING</th>
<th>BEING</th>
<th>UNBECOMING</th>
</tr>
</thead>
<tbody>
<tr>
<td>existence</td>
<td></td>
<td>быть, есть, Ø</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>byt’, est’, Ø</td>
<td></td>
</tr>
<tr>
<td>possession</td>
<td></td>
<td>иметься</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>imet’sja</td>
<td></td>
</tr>
<tr>
<td>creation</td>
<td>твориться</td>
<td>существовать</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tворит’сja</td>
<td>suществоват’</td>
<td></td>
</tr>
<tr>
<td>life</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>visibility, presence</td>
<td></td>
<td>являться</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>явлются</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>представлять собой</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>представят’себя</td>
<td></td>
</tr>
<tr>
<td>accessibility</td>
<td>находиться</td>
<td>присутствовать</td>
<td></td>
</tr>
<tr>
<td></td>
<td>находит’сja</td>
<td>присутствоват’</td>
<td></td>
</tr>
<tr>
<td>motion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>process</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>position</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>manipulation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For updated information, see http://home.uchicago.edu/~sclancy
Ja ne beremenna i ne javljajus’ biologičeskoj ili priemnoj mater’ju rebenka.
[I not pregnant and not appear biological or adopted mother child.]
‘I am not pregnant and am not the biological or adopted mother of a child.’

I ešče — lager’ predstavljaet soboj dovol’no točnuju model’ gosudarstva.
[And still — camp presents self-INST enough precise model-ACC state-GEN.]
‘And another thing — the camp is quite a precise model of the state.’
A Wider Paradigm for BE in Russian

Po-moemu, i učastvovat’ na duèli, i prisutstvovat’ na nej, xotja by v kačestve vrača, prosto beznravstvenno.
[By-mine, and participate in duel, and be-present in it, even would in quality doctor, simply immoral.]
‘In my opinion, both to participate in a duel and to be present at one, even if only in the role of a doctor, are simply immoral.’

Paul’ Rudi naxoditsja v tjur’me.
[Paul Rudy is-located in prison.]
‘Paul Rudy is in prison.’

Mašina stoit na ulice.
[Car stands on street.]
‘The car is on the street’
Next steps

- MDS is only the beginning and provides only a coarse-grained image of the semantics of these BE functions.
- 16000 examples in the Uppsala Corpus of 17 Russian BE-like verbs and constructions.
- Tagging a data set retrieved from a corpus yields ID tags that can be used in a Behavioral Profiles analysis (Gries and Divjak 2009 on Engl run, verbs of trying in Russian, begin/start/načat’/stat’)
- The BE data for Russian are being automatically or semi-automatically tagged for aspect, person, number, gender, types (pres, past, participle, infinitive, verbal adverb), positive/negative.
- Will be manually tagged for remaining information in the previous categories plus information about X and Y in the X ‘is’ Y construction (nouns, adjectives, case, animacy, auxiliary functions) and other ID tags.

For updated information, see http://home.uchicago.edu/~sclancy
Dendrogram of agnes(x = bb, method = "ward")

For updated information, see http://home.uchicago.edu/~sclancy

bb
Agglomerative Coefficient = 0.78
Preliminary conclusions

- a full set of hand-annotated ID tags will provide a fine-grained measure of similarity of BE-like verbs and constructions to test the hypothesis that the paradigm of BE in Russian is expanding and developing through suppletization.
- methods of quantitative analysis and visualization (although involving a steep learning curve) allow for more objective analyses than purely introspective methods and provide for discovery of relationships that would not otherwise be discovered.

For updated information, see http://home.uchicago.edu/~sclancy