

Finding Constructions: A needle in the haystack, fish in a barrel, or forest for the trees?

Steven J. Clancy

University of Chicago

[Contact: sclancy@uchicago.edu; home.uchicago.edu/~sclancy]

Abstract

Qualitative and quantitative approaches to automatically analyze and annotate natural language data have proven successful on many fronts. Goldsmith (2001, 2006, 2009) and Goldsmith's *Linguistica* software have demonstrated success in identifying inflectional morphology across languages from raw corpus data. The Brown, et al (1992) algorithm (as developed in software by Liang (2005)) and *Linguistica* were used in the current study to identify semantic categories based on word neighbors. The word classes and morphemes thus identified and the signatures, i.e., a database of which stems combine with which morphemes, were used here to successfully identify the part of speech of words strongly associated with the case marking morphemes of the immediately following one or two words. For Russian, this procedure yielded a semantic category in which 33 out of the 35 most frequent words (94%) in the category were prepositions. A further multiple correspondence analysis (MCA) on the prepositions and the morphology of the following words reveals clusters that correspond to case governance.

Although the research thus far has demonstrated that it is possible to identify some case constructions (prepositions and the cases they combine with), the ultimate goal of this research is to identify structure and meaning across the board for case marking systems in order to accelerate the acquisition of case data with inputs into corpus annotation and to broaden the range of languages studied with inputs into typological research within cognitive linguistics. Previous studies in Slavic case semantics (Janda and Clancy 2002, 2006) form a construction and a gold standard of case marking constructions for Russian and Czech and provide templates for possible unsupervised approaches to replicating this construction. Successful replication will feed into further research both within the Slavic languages (with immediate extensions to Polish and Bosnian/Croatian/Serbian) and to other languages with case and/or adposition marking constructions. The current study aims to refine and automate the process of identifying case constructions so that the case marking systems (morphology and semantics) can be identified directly from corpora in an unsupervised fashion, enabling the expansion of these projects to include many more languages by accelerating the process of data collection and analysis.

However, finding structure in contiguous items such as prepositions and their noun phrases is fairly straightforward, but the challenges of Slavic word order and the variety of constructions pose many problems for unsupervised learning. For a particular verb, such as отвечать/ответить, one might propose the following construction:

[NOM отвечать/ответить (DAT) (на +ACC)]
'someone-NOM answer (someone-DAT) (на+ something-ACC)'

However, a manual inspection for examples in Dostoevsky's *Бесы* revealed a plethora of constructions with this verb, such that the proposed construction might be something more along the lines of the following:

[(не) отвечать/ответить (КОМУ) (на (КАКОЙ) вопрос) (, что) (QUOTE) (не...GEN) (КАК) (КОГДА)
(что/что-нибудь/что-то) (ГДЕ) (КУДА) (ПОЧЕМУ)]
'(not) answer someone/something/that/"QUOTE"/NEGATIVE/how/when/something/where/to where/why'

Unsupervised techniques also face the problems of word order and long distances, such as this example from Bulgakov's *Мастер и Маргарита* with the *столько, сколько* 'so much, as' construction spread out over three exchanges in dialogue:

- Но меня, конечно, **не столько** интересуют автобусы, телефоны и прочая...
- Аппаратура! - Подсказал клетчатый.

- Совершенно верно, благодарю, - медленно говорил маг тяжелым басом, - **сколько** гораздо более важный вопрос: изменились ли эти горожане внутренне?

“But I’m, of course, **not so much** interested in buses, telephones, and such...”

“Equipment!” prompted the checkered one.

“Absolutely, thank you,” slowly spoke the magician with a heavy bass.

“**as** in a much more important question: have these city-dwellers changed internally?”

Constructional depth and richness and long distance relations pose seemingly insurmountable challenges for unsupervised methods. This paper will present the current results from this ongoing research in computational construction grammar and will also discuss the problems with automatic identification of case constructions when combined with manual analyses, annotations, and data collection.

I will discuss the nature of computational construction grammar as applied to the problem of identifying case constructions: are we looking for a needle in a haystack, i.e., are constructions difficult to identify, or are we shooting fish in a barrel (i.e., constructions are present at every level and merely need to be gathered up and classified?). Goldberg’s assertion that all linguistic structure is “constructions all the way down” testifies to the ubiquity of constructions at every level of language. However, this situation all too easily yields a situation in which we cannot see the forest for the trees.

References

- Brown, Peter F., Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18:467-479.
- Goldberg, Adele. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, Oxford, UK.
- Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27:153-198.
- Goldsmith, John. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering* 12(3):353-371.
- Goldsmith, John. 2009. Morphological analogy: Only a Beginning. In James P. Blevins and Juliette Blevins (eds.), *Analogy in Grammar: Form and Acquisition*, pp. 137-163. Oxford: Oxford University Press.
- Goldsmith, John. The Linguistica Project [software and supporting materials]. <http://linguistica.uchicago.edu>
- Janda, Laura A., and Steven J. Clancy. 2002. *The Case Book for Russian*. Slavica Publishers, Bloomington, Indiana.
- Janda, Laura A., and Steven J. Clancy. 2006. *The Case Book for Czech*. Slavica Publishers, Bloomington, Indiana.
- Liang, Percy. 2005. *Semi-Supervised Learning for Natural Language*. MA Thesis. Department of Electrical Engineering and Computer Science, MIT. <http://www.cs.berkeley.edu/~плианг/papers/meng-thesis.pdf>
- Liang, Percy. Word Clustering [software]. <http://www.cs.berkeley.edu/~плианг/software/brown-cluster-1.2.zip>

Finding Constructions

“It’s constructions all the way down.”

— Adele Goldberg

If true, then there is much more in common between the lexical (contents, words) and grammatical (skeletal, structural, more abstract) systems of language.

What’s a Construction?

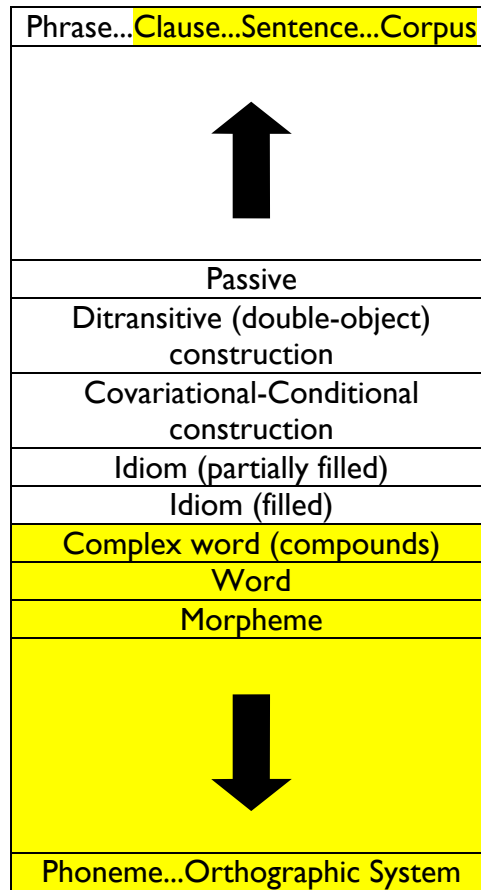
Typical sample constructions for English (based on Goldberg 2006)...

Construction	Form/Example	Function/Meaning
↑	↑	↑
Passive	Form: [Subj aux VPpp (PP _{by})] <i>The armadillo was hit by a car.</i>	Discourse function: to make undergoer topical and/or actor non-topical
Ditransitive (double-object) construction	Form: [Subj V Obj1 Obj2] <i>He gave her a Coke.</i> <i>He baked her a muffin.</i>	Meaning: transfer (intended or actual)
Covariational-Conditional construction	Form: [<i>the Xer the Yer</i>] <i>The more you think about it,</i> <i>the less you understand.</i>	Meaning: linked independent and dependent variables
Idiom (partially filled)	[jog X ^{ANIM} 's memory]	
Idiom (filled)	[going great guns]	
Complex word (compounds)	[daredevil], [shoo-in]	
Word	[avocado], [anaconda], [and]	
Morpheme	[anti-], [pre-], [-ing]	
↓	↓	↓

Everything’s a construction, but not all constructions are equally interesting

- are we looking for a needle in a haystack, i.e., are constructions difficult to identify
- or are we shooting fish in a barrel (i.e., constructions are present at every level and merely need to be gathered up and classified?)
- the ubiquity of constructions at every level of language yields a situation in which we cannot see the forest for the trees.

Constructions all the way down and all the way up



Computational methods have successfully identified constructions from the smallest levels up in an unsupervised manner, i.e., without knowing anything in particular about the specific language under analysis, using the empirical data we find in corpora.

Finding larger level constructions in a language should be the same as finding orthographic systems, phonemes, morphemes, words, and compound words, all of which have enjoyed a high degree of success in unsupervised learning approaches.

In the same way we can identify an orthographic system, a phonemic system, a lexicon, we should be able to build up a construction, i.e., a set of possible constructions in a given language.

Constructions from the top down

the corpus as construction

[...CORPUS...]

- trivial...the corpus as a whole is what it is...

the sentence as construction

[SENTENCE $\left[\begin{array}{c} . \\ ? \\ ! \end{array} \right]$]

- trivial...given the starting point of a punctuated corpus and using the punctuation to divide the corpus...

clauses and phrases as constructions

- depending on what kinds of phrases, we want to find, this could also be trivial using a punctuated corpus, but here we also begin to approach the difficulties involved...
- let's approach this phrase level from the bottom up first...

Constructions from the bottom up

something as simple as the information you can gather about the distribution of each unit (e.g., a letter) and it's right and left neighboring units can reveal much about the structure of language

letters as constructions

- identifying not only the orthographic symbols used (they're what the corpus consists of), but also recognizing digraphs and other spelling conventions
e.g., *th, sh, ch* in English; *rz, sz, cz, ś-si*, etc. in Polish

phonemes as constructions

- identifying the consonant and vowel categories in language

words as constructions

- tokens in the corpus; broken vs. unbroken corpora (e.g., Chinese)
- lemmas and lemmatizers

morphological constructions

- morphemes, prefixes, derivational and inflectional suffixes
- with words and morphemes, we can begin to approach the phrasal and clausal levels from below...

Two Problems:

Constructional depth and richness and long distance relations pose seemingly insurmountable challenges for unsupervised methods.

The problem of richness

Sample construction for отвечать/ответить

A construction entry for [отвечать/ответить] based on examples in Dostoevsky's *Бесы*

A first guess might be:

[NOM отвечать/ответить (DAT) (на +ACC)]
'someone-NOM answer (someone-DAT) (на+ something-ACC)'

However, a manual inspection for examples in Dostoevsky's *Бесы* revealed a plethora of constructions with this verb, such that the proposed construction might be something more along the lines of the following:

[NOM (не) отвечать/ответить (КОМУ) (на (КАКОЙ) вопрос) (, что) (QUOTE) (NOM не...GEN) (КАК) (КОГДА) (что/что-нибудь/что-то) (ГДЕ) (КУДА) (ПОЧЕМУ)]d
'Someone (not) answer someone/something/that/"QUOTE"/NEGATIVE/how/when/something/where/to where/why'

[отвечать/ответить]
[не отвечать/ответить]

[отвечать/ответить на ACC]
[на (КАКОЙ) вопрос-ACC]
[на мой поклон]
[на все это]
[на письмо]
[на мой повторительный стук и зов]
[на некоторые жалобы и запросы]
[на мой вопросительный взгляд]

[отвечать/ответить, что...]
[отвечать/ответить QUOTE]

[отвечать/ответить за ACC]
[за это]

[не отвечать/ответить GEN] GEN possible if negated
[ничего не отвечать/ответить]
[не отвечать/ответить ни слова]

[не отвечать/ответить GEN] ACC possible if indefinite (что/что-нибудь/что-то)

[отвечать/ответить КАК?]
1. [ADV]
2. other expressions
[с чем]
[с (КАКОЙ) улыбкой]
[почти с натуральным простодушием]
[письмом]
[(КАКИМ) голосом]
[гордым взглядом]
[просьбой ДЕЛАТЬ ЧТО]
[безо всякого удивления такому вопросу]
[скороговоркой]
[с пренебрежением]

[таким рыцарем]
[с тем же раздражением]
[чуть не в горячке]
[с самую ясную улыбкой]
[только смехом]
[вопросительным длинным взглядом , не слишком впрочем удивленным]
[взаимностью]
[по их манере, что, дескать, тонкого ума и со здравым суждением]
[мало] (АСС or КАК?)
[с точностью]
[безо всякого следа первоначального внезапного своего волнения и без малейшего смущения,
которое могло бы свидетельствовать о сознании хотя бы какой-нибудь за собою вины]

[отвечать/ответить КУДА?]
[сквозь двери]
[в глаза]

[отвечать/ответить ПОЧЕМУ?]
[зачем?]
[от бездарности]
[потому????]

[MODAL отвечать/ответить]
[не мочь отвечать]
[мочь ответить]
[не хотеть отвечать]
[хотеть ответить]
[не суметь ответить]
[не успеть ответить]
[предложить ответить]
[просить кого ответить]
[положено было систематически не отвечать]
[решиться не отвечать]
[перестать отвечать]
[не видеть надобности отвечать]
[с видимою готовностью отвечать]
[невозможно отвечать]
[неужто отвечать]
[приказать КОМУ отвечать]
□

[IMPERSONAL отвечать/ответить]
[трудно ответить]

The problem of distance and intervening material

- Unsupervised techniques also face the problems of word order and long distances
- Something as simple as the presence of adverbs
- parenthetical comments are ubiquitous
- extreme examples such as this example from Bulgakov's *Мастер и Маргарита* with the *столько, сколько* 'so much, as' construction spread out over three exchanges in dialogue:

- Но меня, конечно, **не столько** интересуют автобусы, телефоны и прочая...
- Аппаратура! - Подсказал клетчатый.

- Совершенно верно, благодарю, - медленно говорил маг тяжелым басом, - **сколько** гораздо более важный вопрос: изменились ли эти горожане внутренне?

“But I’m, of course, **not so much** interested in buses, telephones, and such...”

“Equipment!” prompted the checkered one.

“Absolutely, thank you,” slowly spoke the magician with a heavy bass.

“**as** in a much more important question: have these city-dwellers changed internally?”

Identifying Constructions: Procedure

- working with modified Uppsala Corpus (>1 million words)
- take most frequent multi-word collocates (e.g., 4-word, 3-word, 2-word) and mark the corpus for these, taking them out of the picture from the beginning
- corpus analyzed with Linguistica (Goldsmith, The Linguistica Project)
 - results in a morphologically tagged corpus
 - three aligned versions of the corpus can be produced using the results
 - 1) the original raw corpus
 - 2) the corpus with stems and endings
 - 3) the corpus with stems only
- perform the IBM-Brown algorithm on the corpus; use resulting categories to analyze certain words and surrounding morphology
- divide the corpus into phrases (based on punctuation); use resulting phrases to look at words and surrounding morphology
- use additional information on signatures (patterns of endings for a given stem) to identify paradigms or categories

Uppsala Corpus

Chosen for its availability (downloadable, raw text files) and size (> 1 million words; not too large for experimenting with various algorithms and procedures). Transliterated text returned to Cyrillic and entire corpus compiled as a single document with one-sentence per line.

In addition, I am using a randomly selected gold standard (>1500 words) comprised of:

<author=Виктор Пелевин, text=Омон Ра, name=Victor Pelevin, book=Omon Ra, length=590 words>

<author=Михаил Булгаков, text=Белая гвардия, name=Mikhail Bulgakov, book=The White Guard, length=543 words>

<author=Ф.М. Достоевский, text=Идиот, name=Fyodor Dostoevsky, book=Idiot, length=531 words>

crude method for identifying constructions from the top down

- take most frequent multi-word collocates
- for my corpus, 5-word (including punctuation) collocates

Дело в том , что	17	С другой стороны ,	7
Дело в том ,	14	в том , что в	7
в том числе и	12	И тем не менее	6
В самом деле ,	9	о том , что в	6
В то же время	9	Партийного Контроля	
на мой взгляд ,	8	при ЦК КПСС	6
В связи с этим	8	состоит в том , что	6
к сожалению , не	7		

Linguistica finds morphology from a raw corpus

John Goldsmith's research group (linguistica.uchicago.edu) at the University of Chicago has created a software package that explores structure in natural language, focusing on word structure/morphology. The program finds morpheme boundaries, identifies stems and suffixes, including case endings for languages with case.

Linguistica yields a broken, morphologically tagged corpus

Raw Corpus

Омон - имя не особо частое и , может , не самое лучшее , какое бывает .
Меня так назвал отец , который всю свою жизнь проработал в милиции и хотел ,
чтобы я тоже стал милиционером .
- Пойми , Омка , - часто говорил он мне , выпив , - пойдешь в милицию - так с
таким именем , да еще если в партию вступишь ...

Morphologically-Tagged Corpus

Омон - имя не особ+о част+ое и , мож+ет , не сам+ое лучш+ее , как+ое быв+ает
.
Меня+NULL так+NULL назв+ал отец , котор+ый всю свою жизнь проработ+ал в
милиции и хот+ел , чтобы я тоже стал милиционер+ом .
- Пойми , Омка , - част+о говори+л он мне+NULL , выпив+в , - пойд+ешь в
милицию - так+NULL с+NULL так+им имен+ем , да+NULL еще если в парт+ию
вступ+ишь ...

- marks the endings on analyzed words
- outputs a lexicon of words and signatures of morphemes they combine with
котор: ая ого ое ой ом ому ою ую ые ый ым ыми ых
4394 hits; 13 endings
еха: л ла ли ть
119 occurrences, 4 endings
двер: ей ец и ь ьми ью ям ями ях
321 occurrences, 9 endings
- all of which can be used to tag the original raw corpus
- stems and endings are not necessarily correct or what a linguist familiar with the language would use if annotating manually
 - *Linguistica* coded my toy corpus 68% of the time the same way I did

Examples:

by hand		<i>Linguistica</i>
бы+ла	vs.	был+а
девушк+у	vs.	девуш+ку
мн+е	vs.	мне+NULL
него	vs.	нег+о
он+NULL	vs.	он

он+а	vs.	она
он+и	vs.	они
с	vs.	с+NULL
теб+е	vs.	те+бе
уех+а	vs.	уех+а

What can you do with a broken corpus?

- goal is to find case structure automatically (i.e., in an unsupervised way) from a raw corpus
- where to begin?

The case of prepositions

- adpositions have a close connection with their contiguous neighbors
- can we find them in an unsupervised fashion by comparing words and their right hand neighbors?

Brown et al 1992: Algorithm for Finding Semantic Categories

- Brown et al 1992 identified semantic categories based on mutual information of words and their neighbors
 - Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays Saturdays
 - June March July April January December October November September August
 - people guys folks fellows CEOs chaps doubters commies unfortunates blokes
 - down backwards ashore sideways southward northward overboard aloft downwards adrift
 - water gas coal liquid acid sand carbon steam shale iron
 - great big vast sudden mere sheer gigantic lifelong scant colossal
 - man woman boy girl lawyer doctor guy farmer teacher citizen
 - American Indian European Japanese German African Catholic Israeli Italian Arab
 - pressure temperature permeability density porosity stress velocity viscosity gravity tension
 - mother wife father son husband brother daughter sister boss uncle
 - machine device controller processor CPU printer spindle subsystem compiler plotter
 - John George James Bob Robert Paul William Jim David Mike
 - anyone someone anybody somebody
 - feet miles pounds degrees inches barrels tons acres meters bytes
 - director chief professor commissioner commander treasurer founder superintendent dean custodian
 - liberal conservative parliamentary royal progressive Tory provisional separatist federalist PQ
 - had hadn't hath would've could've should've must've might've
 - asking telling wondering instructing informing kidding reminding bothering thanking deposing
 - that tha theat
 - head body hands eyes voice arm seat eye hair mouth

(Brown et al 1992: 475)

- Using the implementation of the Brown algorithm found in Liang 2005, I asked for up to 10 categories for corpus data from a number of languages
- among the categories identified, one in particular stands out for its concentration of prepositions
- aside from Russian and Polish (see below), this phenomenon holds up when the algorithm is run for other languages as well (e.g., Czech, Latin, German, French, Dutch)

on the Uppsala Corpus with 20 categories in Russian

- requested up to 20 categories requested; data was raw Uppsala Corpus, broken into phrases by punctuation and made all lower-case
- returned 19 categories
- reveals more than one interesting category (but for prepositions, see below, it wasn't as good)
- Category 1
 - the comma ,
 - [+ low frequency items]
- Category 2
 - в, на, с, к, :, по, ?, из, от, за, у, !, для, о, (, до, при, под, без, со, после, ,, через, между, об, перед, над, среди, около, кроме, из-за, про, против, с_, возле, вместо, сквозь, ан, ради, из-под, вдоль, благодаря
 - [these items with frequency > 100; 86% are prepositions]
 - 36 prepositions
 - Punctuation : ? ! ; and the open parenthesis
 - the initial С., АН
- Category 3
 - the period .
 - [+ low frequency items]
- Category 4
 - а, но, однако, причем
 - among less frequent items, lot of verbal adverbs and participles
- Category 5
 - и, или
 - [less frequent items]
- Category 6
 - one with lots of GEN and possessive items
- Category 7
 - one with a high proportion of question words/relative conjunctions
- Category 8
 - another one with question words/relative conjunctions
- Category 9
 - one with forms of быть, lots of modal particles
- Category 10
 - one with names, kinship, personal pronouns
- Category 11
 - one with infinitives and бы/б
- Category 12
 - one with comparatives
- But all of these have to be taken with a grain of salt. Even when tendencies seem to jump out, they jump out from lots of junk. None of the other categories are so good as the preposition category and that one wasn't as good here with more categories as it is with fewer categories requested

the “preposition” category in Russian

- top 35 items in one category by frequency
- 33 out of top 35 hits in one category are prepositions (94%); Brown Algorithm on raw corpus broken into phrases, 10 categories requested
- improved over this algorithm with the raw 1-sentence-per-line corpus where 24 out of top 35 hits had found Russian prepositions (69%)

в	28576	для	2984	между	593	возле	141
на	15474	о	2982	об	589	сквозь	141
с	10537	до	1971	перед	572	АН	134
к	5417	при	1364	над	556	Совета	130
по	5227	под	1201	около	320	вместо	118
из	4150	без	1115	среди	300	из-под	106
от	3804	со	991	из-за	262	ради	106
за	3580	после	712	про	245	вдоль	104
у	3051	через	660	против	230		

Multiple Correspondence Analysis (MCA) of Prepositions

- preposition plus morphology of neighbor to the right; retain the top 25 most frequent signatures of preposition plus the morphology 1 or 2 of the right-hand neighbors, e.g., после, после, между...

после +а	после +ия	после +ки	после +ного	после +их
после +ого	после +и	после +я	после +ции	после +ого +ия
после +го	после того ,	после чего	после +ов	после +ы .
после +ы	после +ого +а	после +ых	после ,	после того +NULL
после того	после +ой	после +ы ,	после +а +а	после +а ,
между +ами	между +NULL	между +ами ,	между +ными	между +ими +ами
между +ой	между +м	между +ой и	между ними .	между +ей
между +ом	между +ыми	между двумя	между +ами и	между +ой .
между ними	между +м ,	между +ами .	между +ями	между +ыми +ами
между +ом и	между +ими	между людьми	между ними ,	между +ем

- measure of similarity among the prepositions and the possible forms or case endings to the right; results in 2-D map of the prepositional space; conducted in R using the mca() function
- map accounts for multiple case usage, case ending syncretism, general polysemy of the prepositions

See Prezi for this talk for MCA graphs of these prepositions.

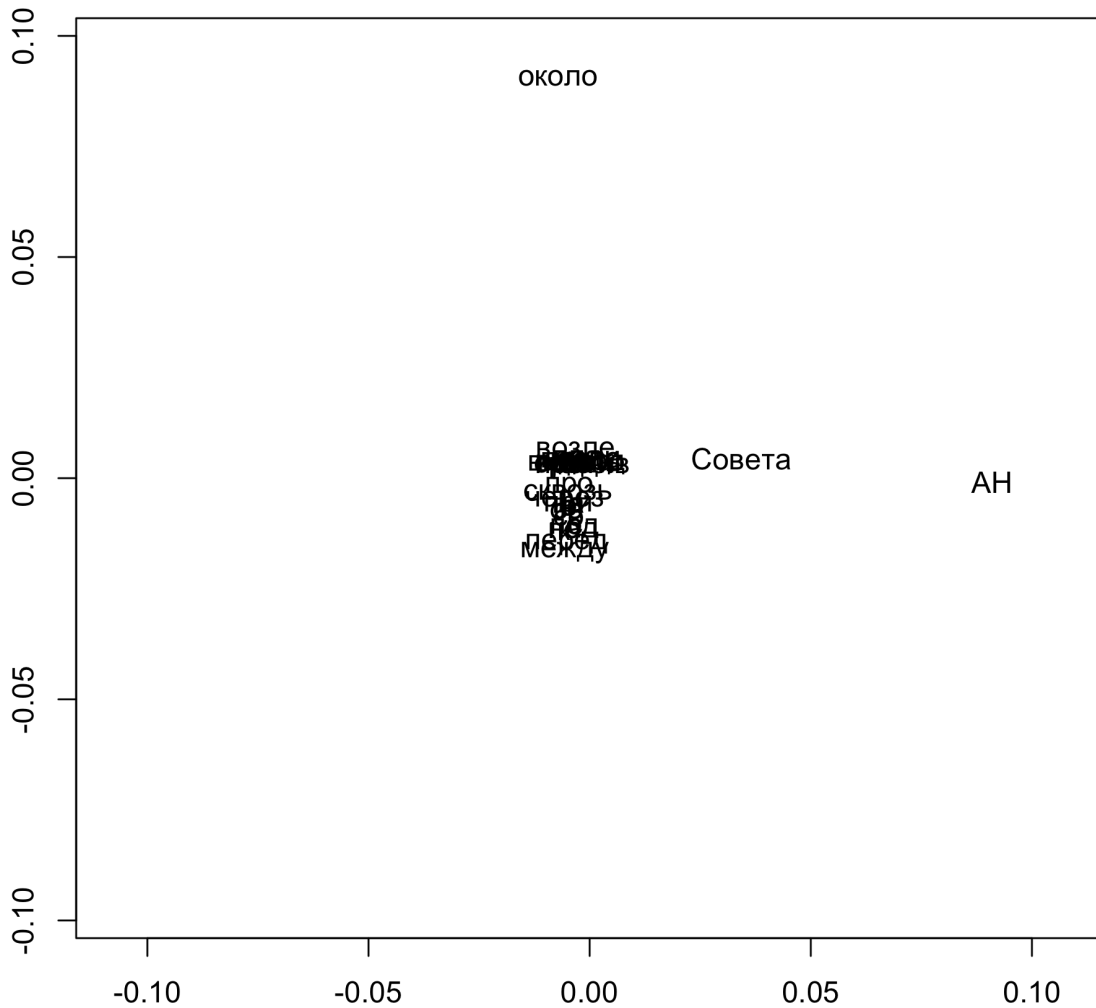


Figure 1:
MCA Analysis
32 Russian Prepositions (using top 25 morphemes from R1 and R1+R2) (forms selected by 35 most frequent items in one of the categories found by the Brown Algorithm)

Preliminary Conclusions and Future Directions

- the unsupervised morphological analysis provided by *Linguistica* makes it possible to partially annotate raw corpora and then align multiple versions of the corpus (raw, broken, stems, morphology)
- additional information provided by *linguistica* may make it possible to get to POS identification and identification of paradigms
- these additional discoveries can then be fed back into further annotations to the corpus
- the Brown et al 1992 approach makes it possible to find some semantic categories from raw corpora
- these algorithms are not language specific
- these tools provide us with corpora that can be employed in a Computational Construction Grammar approach in order to build up the construction of a language
- aside from the unsupervised approach, the morphologically annotated corpora from *Linguistica* can be used to explore other phenomena
- the resultant richly-annotated corpora will provide material for quantitative methods such as Behavioral Profile (BP) analysis (Divjak 2006, Divjak & Gries 2006), Multidimensional Scaling-Optimal Classification Method (MDS-OC) analysis (Clancy 2006, Feist 2008), and other techniques such as Multiple Correspondence Analysis (MCA)
- if multi-word collocations, prepositional phrases, and some other constructions can be identified, then the corpus can be collapsed in those areas and further combinations words and neighbors can be explored
- there is much room for experimentation with the methods, variables, and approaches used in this study, for instance word + R1 morphology vs. word + R1&R2 morphology
- the goal is to identify a construction that will approach the ~1000 constructions identified in studies of the Russian and Czech case systems (Janda and Clancy 2002, 2006)