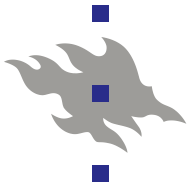


Exemplars, *prototypes,* or both?
What evidence do corpora contain for the representation of linguistic categories?

Antti Arppe
University of Helsinki
Finland

Dagmar Divjak
University of Sheffield
UK

The fundamental big issues

- How are linguistic *structures*?
  - Acquired
  - Represented in the human mind
  - Reflected in usage, e.g. corpora
  - Linked with *categorization*
- How are linguistic *categories* stored and represented as cognitive structures?
  - Prototypes
  - Exemplars

(Linguistic/Cognitive) categories – models of representation

- Prototype theory (e.g. Rosch 1978)
  - Storage of highly abstract representations
  - Consisting of aggregate properties
    - Characteristic of the category
    - Instead of strictly defining/delineating
- Exemplar theory (e.g. Hintzman, 1986; Nosofsky, 1986)
  - Storage of detailed memory traces of all the individually encountered examples
  - Little or no abstraction over instances

Models of representation

- Prototype and exemplar models
  - traditionally treated as mutually exclusive
- Alternatively (Verbeemen et al. 2007)
  - Opposite ends along a *gradient continuum*
  - Following a *usage-based* view, prototypes would emerge from repeated exposure to and abstraction over exemplars

Our objective

- Primarily: How can these various models of categorization be observed to be manifested in *corpus data*?
  - With the help of synonymy (six Russian near-synonymous verbs denoting TRY)
  - Based upon the results of *polytomous logistic regression* – a multivariate statistical method
- Secondarily: Is there support for a continuum between the two alternative models?

From extreme exemplars to most abstract prototypes

Exempar-route – *expected probabilities*:

Individual example sentences

→ ”Best” exemplars (manifesting distinct contextual property clusters + high probability ← many/strongly favorable odds)

Property-route – *odds*:

Individual contextual property combinations

→ Distinct verb+context types

→ Distinct context types (disregarding outcome verb)

→ Aggregates of properties with (strongly) favorable odds → (abstract) prototypes?

The model / statistical method

- Multivariate analysis: (polytomous) logistic regression
  - Looks at outcomes as proportions among all observations with the same context rather than individual *either-or* dichotomies of occurrence vs. non-occurrence
  - Thus: estimates ***probabilities of occurrence*** given a particular context
  - Estimates variable parameters for properties which can be interpreted naturally as ***odds*** (Harrell 2001)
    - How much does the *existence* of a variable (i.e. property) in the context increase (or decrease) the *chances* of a particular *outcome* (i.e. lexeme) to occur, with all the other explanatory variables being equal?
  - Optimizes the *fit* of the outcome proportions in the data with the probabilities yielded by the parameters

One-vs-rest classification

(Rifkin&Klautau, 2004)

- Distinguishes each member of the set without requiring some baseline category (as required in proper multinomial logistic regression)
- Provides directly lexeme-specific odds with respect to selected variables (representing linguistic properties)
- Selection rule: pick the verb receiving the highest probability

  $arg_{Verb}\ max[P(Verb|Context)]$

- Highest estimated probability not necessarily P>0.5 – can range from slightly over 1/6 → 1.0

The data

- Divjak: TRY in Russian
  - The 6 most frequent verbs that express TRY when combined with an infinitive
    - ***prob****ovat'*
    - *py**tat**'sja*
    - *sta**rat**'sja*
    - ***si****lit'sja*
    - *noro**vit'***
    - *pory**vat**'sja*
- Data extracted from corpora
  - Amsterdam Russian Corpus
  - Russian National Corpus
  - (selected) Internet pages
- In all 1351 occurrences
  - Roughly equal proportions for each TRY verb (*n*=119…250)

Getting the data into the model

- Original analysis contains 14 variables amounting to 87 variable categories (Divjak& Gries 2006)
  - have to be pruned, since (polytomous) logistic regression allows for maximally 1/10 variables per data points of least frequent outcome (i.e. 150/10=15)
- **Selection strategy**: variables with a broad dispersion among the 6 TRY verbs
  - focuses on the interaction of variables in determining the expected probability in context
  - in contrast to allowing individual distinctive variables, linked to only one of the verbs, to alone determine the choice
- Selection criteria
  - overall frequency in the data at least 45
  - occurrence at least twice (i.e. not just single chance) with all 6 TRY verbs
- Technical restrictions/requirements
  - exclusion of one variable for each fully complementary case (e.g. aspect of verb form)
  - exclusion of one of the variables when mutual pairwise association statistic Uncertainty Co-Efficient *UC*>0.5 (i.e. one variable accounts for more than ½ of the variance of the other)

Variable summary

- Altogether 18 variables [7 structural/11 semantic]
  - Clause type: Main (vs. subordinate) [1/2]
  - Sentence type: Declarative (vs. other rarer types) [1/4]
  - Finite verb: morphological properties [4/12]
  - Infinitive verb: morphological properties [1/2]
  - Infinitive verb: degree of control (semantic) [1/3]
  - **Infinitive verb**: semantic characterization (communication, exchange, motion, metaphorical motion, etc.) [9/14]
  - Syntactic subject: semantic characterization (Animate human vs. rarer other types) [1/9]

Model performance

Statistic	Value
Recall	699/1351 (51.7%)
R_L^2(TEACH)	0.31
R_L^2(TEST)	0.31
lambda (Menard)	0.41
tau (Menard)	0.42

Cf. these results with an over-fitting model with 26 variables (in which only complementary variables were excluded) which reaches a *Recall* of 53% (719/1351 correct choices)?

BEAR IN MIND: we are predicting a 6-way choice between near-synonyms. In a 4-way choice (between imposed, believed, requested and correlated), the average non-English US college applicant gets 64.5% correct (Landauer and Dumais 1997).

Fitted model – verb-specific property-wise odds

Property/Verb	Probovat'	Silit'sja	Starat'sja
(Intercept)	1/22	(1/5.8)	1/47
CLAUSE.MAIN	3.4	(1/1)	(1/1.1)
FINITE.ASPECT_PERFECTIVE	29	(1/49253205)	(1.1)
FINITE.MOOD_GERUND	1/8.3	7	2.2
FINITE.MOOD_INDICATIVE	1/2.8	(2.1)	(1.9)
FINITE.TENSE_PAST	(1/1)	2.1	1/2
INFINITIVE.ASPECT_IMPERFECTIVE	6.1	1/10	4
INFINITIVE.CONTROL_HIGH	**(1/1.2)**	**1/6.4**	**1.6**
INFINITIVE.SEM_COMMUNICATION	2.1	(1/1)	(1/1.6)
INFINITIVE.SEM_EXCHANGE	(1.4)	1/11	(1/1.5)
INFINITIVE.SEM_METAPHORICAL_MOTION	(1.5)	1/3.7	(1/1.5)
INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE	(1/1.3)	1/3	(1.8)
INFINITIVE.SEM_METAPHORICAL_PHYSICAL_OTHER	(1.3)	(1/1.3)	(1/1.1)
INFINITIVE.SEM_MOTION	(1.7)	1/4.5	1/3.2
INFINITIVE.SEM_MOTION_OTHER	(2.6)	(1/1.3)	1/3.6
INFINITIVE.SEM_PHYSICAL	3.9	(1.1)	(1/1.8)
INFINITIVE.SEM_PHYSICAL_OTHER	2.5	1/2.6	1/2.1
SENTENCE.DECLARATIVE	1/2.8	(3.2)	2.8
SUBJECT.SEM_ANIMATE_HUMAN	(1.5)	(1/1)	2.5

Fitted model – verb-specific property-wise odds – cont'd

Property/Verb	Norovit'	Poryvat'sja	Pytat'sja
(Intercept)	(1/2.2)	1/3380	1/12
CLAUSE.MAIN	(1/1.2)	(1/1)	1/1.6
FINITE.ASPECT_PERFECTIVE	(1/181239356)	(1/29548257)	(1.1)
FINITE.MOOD_GERUND	1/6	(2.8)	(1.2)
FINITE.MOOD_INDICATIVE	(1/1.2)	(1.8)	(1.3)
FINITE.TENSE_PAST	1/3.3	3.3	2.4
INFINITIVE.ASPECT_IMPERFECTIVE	1/2.9	(1)	1/2.7
INFINITIVE.CONTROL_HIGH	**2.6**	**4.7**	**3.1**
INFINITIVE.SEM_COMMUNICATION	(1.2)	8.4	1/1.9
INFINITIVE.SEM_EXCHANGE	7.7	9.1	(1/1.9)
INFINITIVE.SEM_METAPHORICAL_MOTION	6.1	(1.9)	(1)
INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE	4	(4)	1/2.6
INFINITIVE.SEM_METAPHORICAL_PHYSICAL_OTHER	2.7	(1.3)	(1/1.3)
INFINITIVE.SEM_MOTION	8.1	19	1/4.2
INFINITIVE.SEM_MOTION_OTHER	4.5	5.1	(1/1.5)
INFINITIVE.SEM_PHYSICAL	6	(1.6)	1/4.1
INFINITIVE.SEM_PHYSICAL_OTHER	6.1	3.1	(1/1.5)
SENTENCE.DECLARATIVE	(1)	(1.3)	(1/1.1)
SUBJECT.SEM_ANIMATE_HUMAN	1/4	4.1	(1.4)

Fitted model – Sentence-wise probability estimates

- ***Poprobuj*** *ujti otsjuda ili otkazat'sja ot obeshhannogo!*

  "Try to go away from here or to renounce what was promised!"

- The context with the third highest probability:

  {CLAUSE.MAIN,
  FINITE.ASPECT_PERFECTIVE,
  INFINITIVE.CONTROL_HIGH,
  INFINITIVE.SEM_MOTION,
  SUBJECT.SEM_ANIMATE_HUMAN}

Probability estimates

P(*probovat'*|CONTEXT) =

1:1.9 ~ *Intercept* +

3.4:1 ~ CLAUSE.MAIN +

29:1 ~ FINITE.ASPECT_PERFECTIVE +

1:1.2 ~ INFINITIVE.CONTROL_HIGH +

1.7:1 ~ INFINITIVE.SEM_MOTION +

1.5:1 ~ SUBJECT.SEM_ANIMATE_HUMAN

= 0.92

Probability estimates

- All individual sentences can be ranked in terms of the verb-specific context-based expected probability
- N.B. Estimated probability is the same for each verb in all similar contexts
  - 3 sentences with above context and property combination – all with *probovat'*
- In all 296 distinct property combination types (i.e. Contexts)
- Overall 660 verb+property combination types (verb-specific contexts) – 100 for *probovat'*

Probabilities vs. proportions

- E.g. most frequent property combination in the data
  - CLAUSE.MAIN+ FINITE.MOOD_INDICATIVE+ FINITE.TENSE_PAST+ INFINITIVE.CONTROL_HIGH+ INFINITIVE.SEM_COMMUNICATION+ SENTENCE.DECLARATIVE+ SUBJECT.SEM_ANIMATE_HUMAN
  - *N*=60
- Original frequencies in the data

norovit	poryvatsja	probovat	pytatsja	silitsja	staratsja
6	17	3	11	18	5

- Original proportions of occurrence in the data

norovit	poryvatsja	probovat	pytatsja	silitsja	staratsja
0.10	0.28	0.05	0.18	0.30	0.08

- Estimated probabilities according to the model

norovit	poryvatsja	probovat	pytatsja	silitsja	staratsja
0.07	0.26	0.05	0.25	0.25	0.11

Selecting best/most prototypical exemplars

- Are the sentences with the highest probability estimates per each verb their best/most prototypical exemplars?
- Picking only the most probable sentences yields ones with similar contexts and property combinations
  - 277 sentences for which P(*probovat'*|C)>0.5
  - Represent 85 distinct property combinations
  - What degree of overlap? (have to figure out)

Clustering the exemplars according to their properties

- Hierarchical Agglomerative Clustering
  - Internally similar but group-wise distinct clusters
  - Number of clusters can be arbitrarily defined and tried out: 40 (in this study after some trial)
  - Distance metric: *binary* ← binary property values
  - Clustering algorithm: *Ward* ← compact clusters
- Selecting from each (distinct property cluster) the verb (and exemplar sentence) with the highest probability estimate
  - Expectation as the best representative for the property cluster in question

Property clustering

- 40 clusters divided unequally among the 6 TRY verbs:
  - *Probovat'*: 13 exemplary sentences representing distinct property combinations
  - *Silit'sja*: 12
  - *Norovit'*: 10
  - *Starat'sja*: 3
  - *Pytat'sja*: 1
  - *Poryvat'sja*: 1
- Manifestations of verb-specific property-wise/contextual diversity?

Clustering: *probovat'*

Property/Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13
CLAUSE.MAIN	+	+	+	+	-	+	+	+	+	+	+	+	+
FINITE.ASPECT_PERFECTIVE	+	+	+	-	+	+	+	+	+	+	+	+	+
FINITE.MOOD_INDICATIVE	+	-	+	+	-	+	+	-	-	-	-	+	-
FINITE.TENSE_PAST	+	-	-	+	-	+	+	-	-	-	-	+	-
INFINITIVE.ASPECT_IMPERFECTIVE	-	-	-	+	-	+	+	-	-	-	-	+	-
INFINITIVE.CONTROL_HIGH	+	+	+	+	+	+	+	-	-	+	-	-	-
INFINITIVE.SEM_COMMUNICATION	**+**	**-**	**-**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
INFINITIVE.SEM_EXCHANGE	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**+**	**-**	**-**	**-**	**-**
INFINITIVE.SEM_MOTION	**-**	**-**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**+**	**-**	**+**	**-**
INFINITIVE.SEM_MOTION_OTHER	**-**	**+**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
INFINITIVE.SEM_PHYSICAL	**-**	**-**	**-**	**+**	**-**	**-**	**-**	**+**	**-**	**-**	**-**	**-**	**-**
INFINITIVE.SEM_PHYSICAL_OTHER	**-**	**-**	**+**	**-**	**-**	**-**	**+**	**-**	**-**	**-**	**+**	**-**	**-**
SENTENCE.DECLARATIVE	+	+	-	-	+	-	+	+	+	-	-	+	-
SUBJECT.SEM_ANIMATE_HUMAN	+	+	+	+	+	+	+	+	+	+	+	+	+

Corresponding exemplary exemplars – *probovat'*

- Cluster 3: *P*(probovat|{CLAUSE.MAIN + FINITE.ASPECT_PERFECTIVE + FINITE.MOOD_INFINITIVE + INFINITIVE.CONTROL_HIGH + INFINITIVE.SEM_PHYSICAL_OTHER})=0.771
- Вы меня на крыше подстрахуете, я спущусь в окно, попробую/poprobuju открыть сейф _ как? Ето легавых не касается. Давай фотоаппарат.
  - *Vy menja na kryshe podstrahuete, ja spushhus' v okno, poprobuju otkryt' sejf _kak? Jeto legavyh ne kasaetsja. Davaj fotoapparat.*
- "You cover me from the roof, I'll go down through the window, I'll try to open the safe. How? That's none a of police spy's business. Give me the camera."

Exemplars – *probovat'*

- Cluster 4: *P*(probovat|{CLAUSE.MAIN + FINITE.MOOD_INDICATIVE + FINITE.TENSE_PAST + INFITIVE.ASPECT_IMPERFECTIVE + INFINITIVE.CONTROL_HIGH + INFINITIVE.SEM_PHYSICAL})=0.63
- Кстати, у вас изумительный голос. Вы никогда не пробовали/probovali петь?
  - *Kstati, u vas izumitel'nyj golos. Vy nikogda ne probovali pet'?*
- "By the way, you have a wonderful voice. Have you ever tried singing?" [Kukarkin]

Exemplars – *probovat'*

- Cluster 6: *P*(probovat|{CLAUSE.MAIN + FINITE.ASPECT_PERFECTIVE + FINITE.MOOD_INDICATIVE + FINITE.TENSE_PAST + INFINITIVE.ASPECT_IMPERFECTIVE + INFINITIVE.CONTROL_HIGH + INFINITIVE.SEM_COMMUNICATION})=0.765
- Я бессильно опустился на сидение машины, потом снова попробовал/poproboval голосовать – бесполезно! Шоссе – дорога на аэродром – считается правительственной трассой, здесь и останавливаться нельзя, даже если
  кто-нибудь и мог бы отлить мне свой бензин, он не станет этого делать, зачем ему рисковать автомобильными правами?
  - *Ja bessil'no opustilsja na siden'e mashiny, potom snova poproboval golosovat' – bespolezno! Shosse – doroga na ajerodrom – schitaetsja pravitel'stvennoj trassoj, zdes' i ostanavlivat'sja …*
- "Powerlessly I sank back into the car seat, then I again tried to thumb a ride – without success. The highway – the road to the airport – is considered a governmental route, it is not even allowed to stop here, even if someone could give me some of his gas, he would not do it, why would he risk his driving license?" [Neznanskij, F.]

Fitted model revisited – from favorable odds to preferred/ing contextual properties

Property/Verb	Probovat’
(Intercept)	1/22
CLAUSE.MAIN	**3.4**
FINITE.ASPECT_PERFECTIVE	**29**
FINITE.MOOD_GERUND	1/8.3
FINITE.MOOD_INDICATIVE	1/2.8
FINITE.TENSE_PAST	(1/1)
INFINITIVE.ASPECT_IMPERFECTIVE	**6.1**
INFINITIVE.CONTROL_HIGH	(1/1.2)
INFINITIVE.SEM_COMMUNICATION	**2.1**
INFINITIVE.SEM_EXCHANGE	(1.4)
INFINITIVE.SEM_METAPHORICAL_MOTION	(1.5)
INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE	(1/1.3)
INFINITIVE.SEM_METAPHORICAL_PHYSICAL_OTHER	(1.3)
INFINITIVE.SEM_MOTION	(1.7)
INFINITIVE.SEM_MOTION_OTHER	(2.6)
INFINITIVE.SEM_PHYSICAL	**3.9**
INFINITIVE.SEM_PHYSICAL_OTHER	**2.5**
SENTENCE.DECLARATIVE	1/2.8
SUBJECT.SEM_ANIMATE_HUMAN	(1.5)

Practically possible (occurring) property combinations

- 35/48: CLAUSE.MAIN + FINITE.ASPECT_PERFECTIVE+ INFINITIVE.ASPECT_PERFECTIVE + INFINITIVE.SEM_COMMUNICATION
- 10/15: CLAUSE.MAIN + FINITE.ASPECT_PERFECTIVE + INFINITIVE.ASPECT_PERFECTIVE + INFINITIVE.SEM_PHYSICAL
- 21/31: CLAUSE.MAIN + FINITE.ASPECT_PERFECTIVE + INFINITIVE.ASPECT_PERFECTIVE + INFINITIVE.SEM_PHYSICAL_OTHER

Fitted model revisited – from favorable odds to preferred/ing contextual properties

Preferring properties for *probovat'*
CLAUSE.MAIN
FINITE.ASPECT_PERFECTIVE
INFINITIVE.ASPECT_IMPERFECTIVE
INFINITIVE.SEM_COMMUNICATION
INFINITIVE.SEM_PHYSICAL
INFINITIVE.SEM_PHYSICAL_OTHER

- Are these properties as a whole manifestations of (the core of) a prototype for *probovat'*?
  - you tell someone to *try* (PERFECTIVE) and carry out a *physical action*, to *manipulate* someone or something , or to *communicate* (IMPERFECTIVE – without insisting on it being taken to its natural end)
  - most frequently used TRY verb in mother-child interaction (Stoll corpus)
  - *experimental* character of the attempt

Preferred/ing contextual properties

Preferring properties for *pytat'sja*
FINITE.TENSE_PAST
INFINITIVE.CONTROL_HIGH

Preferring properties for *starat'sja*
FINITE.MOOD_GERUND
INFINITIVE.ASPECT_IMPERFECTIVE
INFINITIVE.CONTROL_HIGH
SUBJECT.SEM_ANIMATE_HUMAN
SENTENCE.DECLARATIVE

Preferring properties for *silit'sja*
FINITE.MOOD_GERUND
FINITE.TENSE_PAST

Preferred/ing contextual properties

Preferring properties for *poryvat'sja*
FINITE.TENSE_PAST
INFINITIVE.CONTROL_HIGH
SUBJECT.SEM_ANIMATE_HUMAN
[semantic types of inf] physical (other), motion (other), communication, exchange

Preferring properties for *norovit'*
INFINITIVE.CONTROL_HIGH
[semantic types of inf] (metaphorical) physical (other), (metaphorical) motion (other), (metaphorical) exchange

Characteristics of the aggregated prototype

- EITHER: triggers the selection when explicitly evident in the context of the verb
  - *Ne vypustit'? Da ja sama ne hochu ehat', ponjatno? Vy menja poprobuj te teper' ugovorit'! Mne izdergali nervy, u menja golos sel, a ja v Amerike'*
  - "Not let him/her out/leave? I myself don't want to go, do you get that? Now you try and convince me. I'm overworked, I've lost my voice, and I'm in America"
  - *P*(*probovat*|{CLAUSE.MAIN, FINITE.ASPECT_PERFECTIVE, INFINITIVE.SEM_COMMUNICATION, SUBJECT.SEM_ANIMATE_HUMAN})=0.927
- OR: Inherently invoked even when the properties with strongly favorable odds are not present in the context

Unexpected *probovat'*

- P(L|C): norovit' poryvat'sja probovat' pytatsja silit'sja **starat'sja**
  #393 0.024 0.050 0.154 0.222 0.039 **0.511**
- Николай сначал пробовал/proboval смеяться вместе со всеми, потом стал разочаровываться, скучнеть, затем страшно обиделся
  - *Nikolaj snachala proboval smejat'sja vmeste so vsemi, potom stal razocharovyvat'sja, skuchnet', zatem strashno obidelsja …*
- "First Nikolai tried laughing together with the others, then he started to get disappointed, bored, and then he took terrible offense …"
- *Probovat'* used here since laughing with the others is one of the (PHYSICAL/COMMUNICATIVE?) things the subject experiments with in order to handle the situation, but it does not work out for him (ends of own volition? ← PERFECTIVE) and gets disappointed, bored, then offended.

Unexpected *probovat'*

- P(L|C): norovit' poryvat'sja probovat' pytat'sja silit'sja **starat'sja**
  #416 0.024 0.050 0.154 0.222 0.039 **0.511**
- только Иван Иванович работал с тем же трагическим старанием, как и раньше. Савельев пробовал/proboval урезонить Ивана Ивановича во время одного из перекуров.
  - *Tol'ko Ivan Ivanovich rabotal s tem zhe tragicheskim staraniem, kak i ran'she. Savel'ev proboval urezonit' Ivana Ivanovicha vo vremja odnogo iz perekurov.*
- "Only Ivan Ivanovich worked with the same tragic diligence as before. Savel'ev tried to bring Ivan Ivanovich to reason during one of their smoke breaks."
- *Probovat'* not expected here, probably because bringing this particular person to reason is considered a difficult task that requires effort. The use of *probovat'* here seems to signal that this attempt is an experiment: it is not important whether it succeeds (they just give it a go during a smoke break), it is just one of the many things they tried.

From extreme exemplars to most abstract prototypes

Exempar-route – *expected probabilities*:

(1351) Individual example sentences

→ (40) ”Best” exemplars (manifesting distinct contextual property clusters + high probability ← many/strongly favorable odds)

Property-route – *odds*:

(1351) Individual contextual property combinations

→ (660) Distinct verb+context types

→ (296) Distinct context types (disregarding outcome verb)

→ (6) Aggregates of properties with (strongly) favorable odds → (abstract) prototypes?

Next steps

- Application of statistical/computational methods directly incorporating the different models of representation of categories
  - Exemplar: Memory/exemplar-based learning
  - Prototype: Self-Organizing Maps
- Experiments with native language users on whether their views of prototypicality correspond with the estimated probabilities?

Conclusions

- A systematic way for
  - Identifying key properties for a linguistic category (i.e. near-synonyms)
  - Extracting a subset of exemplars incorporating these properties
  - Constructing a prototype-core aggregating the properties in the most exemplary exemplars
    - Triggers the selection of the near-synonym when explicitly evident
    - Invoked implicitly when properties not evident
- Operationalization of both the prototype and exemplar views for the representation of (linguistic) categories based on corpus data

Thank you for your attention!

Questions?

The issue(s)

- Descriptive: the how & why of lexical alternations
  - In some contexts one alternative is clearly preferred
  - In other contexts multiple alternatives are allowed
    ? are all options equally suited/good/etc
- Methodological: **probabilistic** lexical semantics
  - not replace but explicate and facilitate sound lexicographical work
- Theoretical: exemplar & prototype semantics
  - Probabilistic interpretation of linguistic rules/regularities – outcomes in context
  - basis for language acquisition/representation research

Model revisited - verb-specific property-wise odds

Property/Verb	probovat	silitsja	staratsja
(Intercept)	(1/1.9)	**1/25**	**1/3.3**
CLAUSE.SUBORDINATE	**1/3.9**	(1.1)	(1.3)
FINITE.ASPECT_PERFECTIVE	**22**	(1/124907469)	(1)
FINITE.MOOD_INDICATIVE	**1/2.8**	**1/1.9**	(1/1)
FINITE.TENSE_PRESENT	**1/4.9**	(1/1.4)	**1.7**
INFINITIVE.ASPECT_PERFECTIVE	**1/5.4**	**8.1**	**1/3.7**
INFINITIVE.CONTROL_LOW	(1/3.8)	**13**	(1/1.8)
INFINITIVE.CONTROL_MEDIUM	(1.5)	**2.7**	(1.2)
INFINITIVE.SEM_COMMUNICATION	(1.6)	**1.9**	**1/1.7**
INFINITIVE.SEM_MENTAL	(1/1.2)	**3**	**1/2.1**
INFINITIVE.SEM_METAPHORICAL_MOTION	(1.1)	(1/1.6)	(1/1.7)
INFINITIVE.SEM_METAPHORICAL_PHYSICAL	(1/1.1)	(1/1.6)	(1/1.1)
INFINITIVE.SEM_MOTION	(1.2)	**1/2.4**	**1/3.5**
INFINITIVE.SEM_PHYSICAL	**3.7**	(1.9)	**1/1.9**
INFINITIVE.SEM_PHYSICAL_OTHER	**2**	(1/1.5)	**1/2.3**
SUBJECT.SEM_ANIMATE_HUMAN	(1.3)	(1)	**2.8**

The big issue(s)

- What is the nature of the relationship between naturally produced language and the posited underlying language system that governs such usage?
  1) use and choice among lexical and structural alternatives in language, and
  2) underlying explanatory factors, following some theory representing language as a system
- How can this be modeled using multivariate statistical methods
  - use of multiple linguistic variables from a range of categories, instead of only one or two
  - use of multivariate statistical methods
- How much of actual, real usage can accurately be modeled?

The data

- Russian and Finnish datasets
  - Arppe: THINK in Finnish
    - The 4 most frequent synonyms meaning think, reflect, ponder, consider
      - *ajatella, miettiä, pohtia, harkita*
  - Divjak: TRY in Russian
    - The 6 most frequent verbs that express try when combined with an infinitive
      - *probovat', pytat'sja, starat'sja, silit'sja, norovit', poryvat'sja*
- Data extracted from corpora
  - Arppe: 2 months of newspaper text (Helsingin Sanomat 1995) and 6 months of Internet newsgroup discussion (SFNET 2002-2003). In all 3404 occurrences
  - Divjak: Amsterdam Russian Corpus, Russian National Corpus, (selected) Internet pages, in all 1581 occurrences

Getting the data into the model

- Original analysis contains 14 variables amounting to 87 variable categories (Divjak& Gries 2006)
  - have to be pruned, since (polytomous) logistic regression allows for maximally 1/10 variables per data points of least frequent outcome (i.e. 150/10=15)
- **Selection strategy**: variables with a broad dispersion among the 6 TRY verbs
  - focuses on the interaction of variables in determining the expected probability in context
  - in contrast to allowing individual distinctive variables, linked to only one of the verbs, to alone determine the choice
- Selection criteria
  - overall frequency in the data at least 45
  - occurrence at least twice (i.e. not just single chance) with all 6 TRY verbs
- Technical restrictions/requirements
  - exclusion of one variable for each fully complementary case (e.g. aspect of verb form)
  - exclusion of one of the variables when mutual pairwise association statistic Uncertainty Co-Efficient *UC*>0.5 (i.e. one variable accounts for more than ½ of the variance of the other)

Variable summary

- Altogether 18 variables [9/9]
  - Clause type: Main (vs. subordinate) [1/2]
  - Sentence type: Declarative (vs. other rarer types) [1/4]
  - Finite verb: morphological properties [4/12]
  - Infinitive verb: morphological properties [1/2]
  - Infinitive verb: degree of control (semantic) [1/3]
  - **Infinitive verb**: semantic characterization (communication, exchange, motion, metaphorical motion, etc.) [9/14]
  - Syntactic subject: semantic characterization (Animate human vs. rarer other types) [1/9]

Model performance

Statistic	Value
Recall	699/1351 (51.7%)
R_L^2(TEACH)	0.31
R_L^2(TEST)	0.31
lambda (Menard)	0.41
tau (Menard)	0.42

Cf. these results with an over-fitting model with 26 variables (in which only complementary variables were excluded) which reaches a *Recall* of 53% (719/1351 correct choices)?

BEAR IN MIND: we are predicting a 6-way choice between near-synonyms. In a 4-way choice (between imposed, believed, requested and correlated), the average non-English US college applicant gets 64.5% correct (Landauer and Dumais 1997).

Think (carefully again)

En halua esittää mielipiteitä **miettimättä** tarkasti, mita oikeastaan **ajattelen**.

I do not want the present opinions without thinking carefully, what I really think [about them]

Don’t you try ...

Но **Сирота** все еще **силился/sililsja** **что-то сказать**, и снова невозможно было понять ни слова из того, что он говорил. Малинин наконец не выдержал и прекратил эту обоюдную муку+ **Ты не старайся/starajsja**, Сирота, все равно я не понимаю+ у тебя рот разбитый ... звук и только, а голоса нет. В госпитале полежишь - восстановится, а сейчас **не пробуй/probuj**, не мучь себя (…) [К. Симонов. Живые и мертвые]

But Sirota was still **trying** 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**, 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]

The model / statistical method

- Multivariate analysis: (polytomous) logistic regression
  - Looks at outcomes as proportions among all observations with the same context rather than individual *either-or* dichotomies of occurrence vs. non-occurrence
  - Thus: estimates ***probabilities of occurrence*** given a particular context
  - Estimates variable parameters which can be interpreted naturally as ***odds*** (Harrell 2001)
- How much does the existence of a variable (i.e. feature) in the context increase (or decrease) the *chances* of a particular outcome (i.e. lexeme) to occur, with all the other explanatory variables being equal?

One-vs-rest classification

(Rifkin&Klautau, 2004)

- Distinguishes each member of the set without requiring some baseline category (as required in proper multinomial logistic regression)
- Provides directly lexeme-specific odds with respect to selected variables (representing linguistic properties)
- Selection rule: pick the verb receiving the highest probability

  $arg_{Verb}\ max[P(Verb|Context)]$

- Highest estimated probability not necessarily P>0.5 – can range from slightly over 1/6 → 1.0

Getting the data into the model

- Original analysis contains 14 variables amounting to 87 variable categories (Divjak& Gries 2006)
  - have to be pruned, since (polytomous) logistic regression allows for maximally 1/10 variables per data points of least frequent outcome (i.e. 150/10=15)
- **Selection strategy**: variables with a broad dispersion among the 6 TRY verbs
  - focuses on the interaction of variables in determining the expected probability in context
  - in contrast to allowing individual distinctive variables, linked to only one of the verbs, to alone determine the choice
- Selection criteria
  - overall frequency in the data at least 45
  - occurrence at least twice (i.e. not just single chance) with all 6 TRY verbs
- Technical restrictions/requirements
  - exclusion of one variable for each fully complementary case (e.g. aspect of verb form)
  - exclusion of one of the variables when mutual pairwise association statistic Uncertainty Co-Efficient *UC*>0.5 (i.e. one variable accounts for more than ½ of the variance of the other)

Variable summary

- Altogether 18 variables [9/9]
  - Clause type: Main (vs. subordinate) [1/2]
  - Sentence type: Declarative (vs. other rarer types) [1/4]
  - Finite verb: morphological properties [4/12]
  - Infinitive verb: morphological properties [1/2]
  - Infinitive verb: degree of control (semantic) [1/3]
  - **Infinitive verb**: semantic characterization (communication, exchange, motion, metaphorical motion, etc.) [9/14]
  - Syntactic subject: semantic characterization (Animate human vs. rarer other types) [1/9]

Model performance

Statistic	Value
Recall	699/1351 (51.7%)
R_L^2(TEACH)	0.31
R_L^2(TEST)	0.31
lambda (Menard)	0.41
tau (Menard)	0.42

Cf. these results with an over-fitting model with 26 variables (in which only complementary variables were excluded) which reaches a *Recall* of 53% (719/1351 correct choices)?

BEAR IN MIND: we are predicting a 6-way choice between near-synonyms. In a 4-way choice (between imposed, believed, requested and correlated), the average non-English US college applicant gets 64.5% correct (Landauer and Dumais 1997).

Verb-specific odds

Property/Verb	probovat	silitsja	staratsja
(Intercept)	(1/1.9)	**1/25**	**1/3.3**
CLAUSE.SUBORDINATE	**1/3.9**	(1.1)	(1.3)
FINITE.ASPECT_PERFECTIVE	**22**	(1/124907469)	(1)
FINITE.MOOD_INDICATIVE	**1/2.8**	**1/1.9**	(1/1)
FINITE.TENSE_PRESENT	**1/4.9**	(1/1.4)	**1.7**
INFINITIVE.ASPECT_PERFECTIVE	**1/5.4**	**8.1**	**1/3.7**
INFINITIVE.CONTROL_LOW	(1/3.8)	**13**	(1/1.8)
INFINITIVE.CONTROL_MEDIUM	(1.5)	**2.7**	(1.2)
INFINITIVE.SEM_COMMUNICATION	(1.6)	**1.9**	**1/1.7**
INFINITIVE.SEM_MENTAL	(1/1.2)	**3**	**1/2.1**
INFINITIVE.SEM_METAPHORICAL_MOTION	(1.1)	(1/1.6)	(1/1.7)
INFINITIVE.SEM_METAPHORICAL_PHYSICAL	(1/1.1)	(1/1.6)	(1/1.1)
INFINITIVE.SEM_MOTION	(1.2)	**1/2.4**	**1/3.5**
INFINITIVE.SEM_PHYSICAL	**3.7**	(1.9)	**1/1.9**
INFINITIVE.SEM_PHYSICAL_OTHER	**2**	(1/1.5)	**1/2.3**
SUBJECT.SEM_ANIMATE_HUMAN	(1.3)	(1)	**2.8**

The odds & probabilities exemplified

Можно. Я вас даже покатаю на катере. Кстати, у вас изумительный голос. Вы никогда не пробовали/**probovali** петь?

P(*probovat*|CONTEXT)=0.63

Have you (n)ever tried to sing? main clause, human subj, indicative past, impf inf, high control, physical action

Страшно. Ну и что ... с ним? _ спросил, стараясь/**starajas'** не выдавать голосом своего волнения, некий невидимый из-за спин, торсов […]

P(*staratsja*|CONTEXT)=0.78

Насчет бабочки, Емма, какая-то ерунда, -- заговорил Никитин пасмурно, тщетно силясь/**siljas'** найти немецкие слова. - Не в этом дело. А, черт, язык! Ну, как же тебе объяснить? [Бондарев Ю. В.]

P(*silitsja*|CONTEXT)=0.67

Model performance revisited

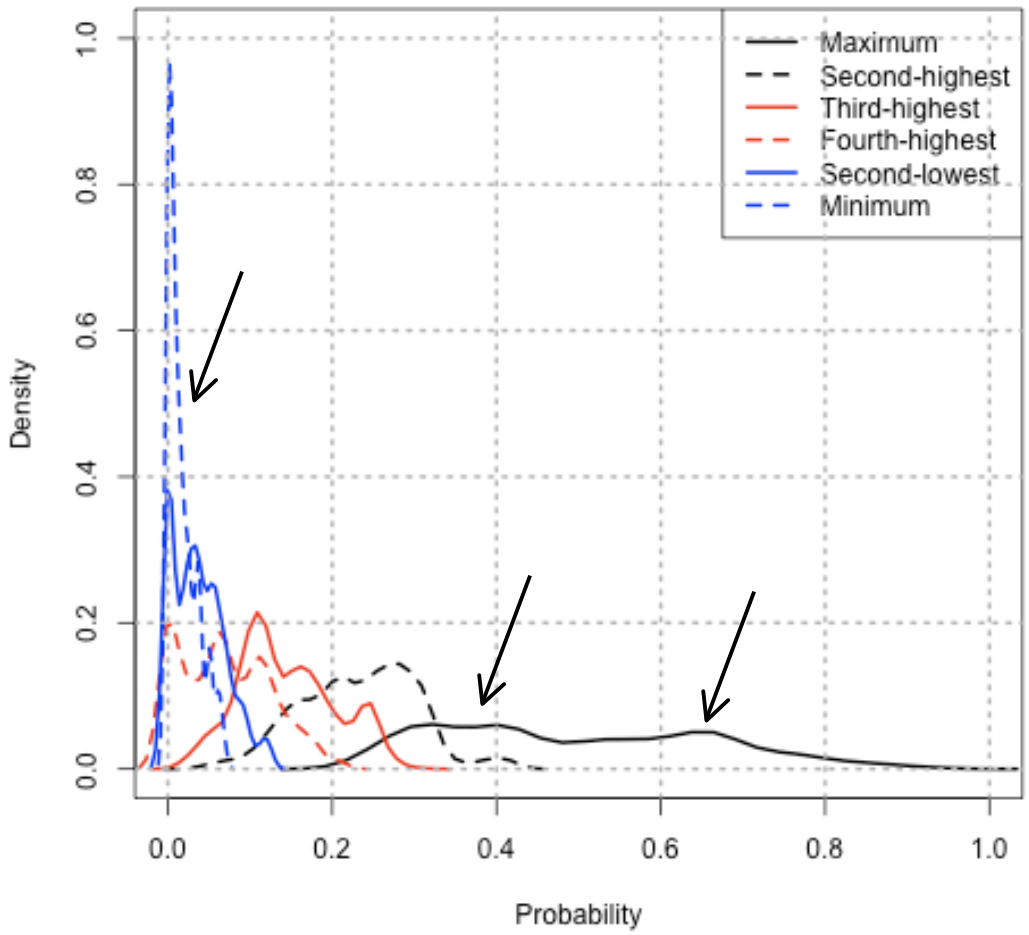
- Recall 699/1351 (51.7%)

? What happens when the verb with the highest probability estimate is not the one that is used in the data

  - (polytomous) logistic regression models overall occurrence proportions in the data, given a context, not individual selections of outcomes (i.e. verbs)
  - when the probability estimates are sustantially dispersed among two or more of the outcome verbs, that means that in the original data these outcomes should have corresponding proportions, given the same context

- This also means that the variables in question cannot alone distinguish/determine which of the outcomes should categorically occur
  - An extreme example: 6-way equiprobable case

Overall sentence-wise probability distributions



6-way synonymous contexts

Frequency	Variable/property set
60	CLAUSE.MAIN+ FINITE.MOOD_INDICATIVE+ FINITE.TENSE_PAST+ INFINITIVE.CONTROL_HIGH+ INFINITIVE.SEM_COMMUNICATION+ SENTENCE.DECLARATIVE+ SUBJECT.SEM_ANIMATE_HUMAN
17	same but INFINITIVE.SEM_MOTION_OTHER
11	same but INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE
9	CLAUSE.MAIN+ FINITE.MOOD_INDICATIVE+ INFINITIVE.CONTROL_HIGH+ INFINITIVE.SEM_COMMUNICATION+ SENTENCE.DECLARATIVE+ SUBJECT.SEM_ANIMATE_HUMAN
3	FINITE.MOOD_INDICATIVE+ INFINITIVE.CONTROL_HIGH+ INFINITIVE.SEM_COMMUNICATION+ SENTENCE.DECLARATIVE+ SUBJECT.SEM_ANIMATE_HUMAN

Probabilities vs. proportions

- Original frequencies in the data

norovit	poryvatsja	probovat	pytatsja	silitsja	staratsja
6	17	3	11	18	5

- Original proportions of occurrence in the data

norovit	poryvatsja	probovat	pytatsja	silitsja	staratsja
0.10	0.28	0.05	0.18	0.30	0.08

- Estimated probabilities according to the model

norovit	poryvatsja	probovat	pytatsja	silitsja	staratsja
0.07	0.26	0.05	0.25	0.25	0.11

- Three verbs [poryvat'sja, pytat'sja and silit'sja] virtually **interchangeable in this context**

  → Look at a few sentences that contain precisely this combination of properties

- Марвич не испытывал злобы к певцу и поэтому знал, что проиграет. Он силился/**sililsja** вызвать злобу, но она не появлялась. [Аксенов]
- Мы нетерпеливо ждали, храня молчание, и только Калаид силился/**sililsja** что-то сказать и шипел. [Стругацкие]
- Она явно силилась/**sililas'** что-то сказать, но вместо слов в сыром воздухе подземелья раздавалось лишь какое-то шипение. [Громов Вадим]
- Чуть не каждое утро она порывалась/**poryvalas'** мне позвонить — и все времени не хватало. [Михаил Бутов. Свобода // «Новый Мир», №.1–2, 1999]
- Нестор порывался/**poryvalsja** запевать, но его не поддерживали, на него вообще как-то не обращали внимания. [Валентин Распутин. Живи и помни (1974)]
- И азиат, живший в сталинской душе, пытался/**pytalsja** обмануть свободу, хитрил с ней, отчаявшись добить ее до конца. [Гроссман]
- Он смотрел на Кирьку опаляющим взглядом, пытался/**pytalsja** что-то сказать. Вошел доктор и бросился к больному. [Шукшин]
- Степана слушали с интересом, немножко удивлялись, говорили “хм”, “ты гляди”, пытались/**pytalis'** сами что-то рассказать, но другие задавали новые вопросы, и Степан снова рассказывал. [Шукшин]

Why do we get such similar estimates?

	norovit	poryvatsja	probovat	pytatsja	staratsja	silitsja
Main clause	(0.9)	(1)	**3**	**0.6**	(0.9)	(1)
Indicative TRY	(0.8)	(2)	**0.4**	(1)	(2)	(2)
Past TRY	**0.3**	**3**	(1)	**2**	0.5	**2**
High infinitive control	**3**	**5**	(0.8)	**3**	**2**	**0.2**
Infinitive communication	(1)	**8**	**2**	**0.5**	(0.6)	(1)
Declarative sentence	(1)	**(1)**	**0.4**	**(0.9)**	**3**	(3)
Human subject	**0.3**	**4**	(2)	(1)	**2**	(1)

Interim conclusion (1)

- The recall rate seems to reach a ceiling at around 52%, and appears indifferent to whether some individual variables are left out → pretty robust
- Did we miss something? Difficult to identify
  - any additional contextual properties or new property clusters
  - pertaining to current, conventional models of morphology, syntax and semantics, and
  - applicable within the immediate sentential context
  - that are not incorporated in the current analysis to some extent
- BUT need more data to model ALL properties reliably
  - include optional elements such as adverbs, etc.

“More” examples

- Он силился/**sililsja** что-то сказать, но у него ничего не получалось: лишь жалкое мычание доносилось сквозь стиснутые зубы да пузырилась в уголках рта пена.
- Ему еще не хватало воздуха, разинутый рот чернел дырой, алые губы обвисли, а он что-то силился/**sililsja** проговорит.
- Мы пришли когда уже давно сидели за столом. Бутылки на треть были опорожнены, мужчины сняли пиджаки и кто-то уже порывался/**poryvalsja** запеть. Но благолепие праздничного стола не было окончательно разрушено (...)
- Бумажку с номером телефона я долго хранил, несколько раз порывался/**poryvalsja** позвонить Рабину, но смущение не позволило мне, к счастью, этого сделать.
- Наверное, кончилось мыло? Дать вам чистое полотенце? Мать задавала наводящие вопросы. Настойчиво пыталась/**pytalas’** вынудить друга к гигиене.

Interim conclusion (2)

- Less is more and more is less
  - More context, less interchangeable
  - Less context, more interchangeable
    - The choices in equiprobable cases cannot be explained on the basis of observed property preferences
      - need to look at properties observed in entire data
    - Lexemes present a different window on the situation
      - the semantic differences between any of the TRY lexemes
        - » are embedded and manifested in the lexemes
        - » would not necessarily have/require explicit manifestation

    ~ prototypes
    ? what are they made up of

Exemplary exemplars (1)

Но такие ребяческие желания могли показаться серьезным людям смешными. Опасаясь этого, он всеми силами пытался/**pytalsja** не выдавать себя. Но это не совсем удавалось. Трудно было ему скрыть свое счастье _ горячий румянец отчетливо проступал на смуглых крепких щеках.]

Можно. Я вас даже покатаю на катере. Кстати, у вас изумительный голос. Вы никогда не пробовали/**probovali** петь?

Страшно. Ну и что ... с ним? _ спросил, стараясь/**starajas'** не выдавать голосом своего волнения, некий невидимий из-за спин, торсов […]

Насчет бабочки, Емма, какая-то ерунда, -- заговорил Никитин пасмурно, тщетно силясь/**siljas'** найти немецкие слова. - Не в этом дело. А, черт, язык! Ну, как же тебе объяснить?

Exemplary exemplars (2)

Хорошо. Он учится во вторую смену, с дженадцати до семи. К восьми вечера он будет здесь. Настя отправилась к себе, с трудом удерживая норовящие/**norovjashhie** выскользнуть из рук папки и с удивлением думая о том, почему это она так спокойно позволяет Заточному распорюжаться ее временем.]

Потом мне рассказывали что и раньше он несколько раз порывался/**poryvalsja** уйти из семьи. Но во-первых уйти было некуда а во-вторых жена приходила жаловаться.

Preferred/ing contextual properties

Preferring properties for *probovat'*
CLAUSE.MAIN
FINITE.ASPECT_PERFECTIVE
INFINITIVE.ASPECT_IMPERFECTIVE
INFINITIVE.SEM_COMMUNICATION
INFINITIVE.SEM_PHYSICAL
INFINITIVE.SEM_PHYSICAL_OTHER

- Are these properties as a whole manifestations of (the core of) a prototype for *probovat'*?
  - you tell someone to try (pf) and carry out a physical action, to manipulate someone or something , or to communicate (impf – without insisting on it being taken to its natural end)
  - most frequently used TRY verb in mother-child interaction (Stoll corpus)

Preferred/ing contextual properties

Preferring properties for *pytat'sja*
FINITE.TENSE_PAST
INFINITIVE.CONTROL_HIGH

Preferring properties for *starat'sja*
FINITE.MOOD_GERUND
INFINITIVE.ASPECT_IMPERFECTIVE
INFINITIVE.CONTROL_HIGH
SUBJECT.SEM_ANIMATE_HUMAN
SENTENCE.DECLARATIVE

Preferring properties for *silit'sja*
FINITE.MOOD_GERUND
FINITE.TENSE_PAST

Preferred/ing contextual properties

Preferring properties for *poryvat'sja*
FINITE.TENSE_PAST
INFINITIVE.CONTROL_HIGH
SUBJECT.SEM_ANIMATE_HUMAN
[semantic types of inf] physical (other), motion (other), communication, exchange

Preferring properties for *norovit'*
INFINITIVE.CONTROL_HIGH
[semantic types of inf] (metaphorical) physical (other), (metaphorical) motion (other), (metaphorical) exchange

Conclusion

- Analysis of exemplars
  - reveals key properties and the combinations in which they can occur together
  - lets us stipulate a prototype-core represented by the aggregate of the exemplars
- Creation of a prototype
  - one or more exemplary examples instantiating (parts of) an (idealized) property configuration occurring significantly more frequently with one lexeme than with others
  - property configuration triggers lexeme explicitly or accompanies it implicitly
  - BUT does not *directly* predict which verb is favored in less typical contexts: assessment of (dis)similarity needed
    - choice for less favored lexeme may be deliberate

? how to compute the cut-off level and model the way in which permissible changes are computed