UiT

NORGES
ARKTISKE
UNIVERSITET

Strategic targeting of rich inflectional morphology for linguistic analysis and L2 acquisition / TROLLing for linguistic data

Laura A. Janda, UiT The Arctic University of Norway

CLEAR

Cognitive Linguistics: Empirical Approaches to Russian

UNIVERSITETET I TROMSØ

Members of the SMARTool team


Radovan Bast


Tore Nesset


Francis Tyers


Mikhail Kopotev


Valentina Zhukova


Elizaveta Kibisova


Svetlana Sokolov


Evgeniia Sudarikova


Ekaterina Rakhilina


Olga Lyashevskaya


James McDonald

Financing and collaboration

Financed by the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education as a collaboration between UiT the Arctic University of Norway and the Higher School of Economics in Moscow

CLEAR

Cognitive Linguistics: Empirical Approaches to Russian

Diku

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Overview

- **Strategic targeting of rich inflectional morphology for linguistic analysis and L2 acquisition**
  - Challenges of rich inflectional morphology
  - Zipfian distribution
  - “Paradigm Cell Filling Problem”
  - A data-based model of rich morphology
  - A machine-learning experiment
  - The SMARTool
- **TROLLing for linguistic data**
  - A professionally-curated archive for linguists everywhere

Strategic targeting of rich inflectional morphology for linguistic analysis and L2 acquisition

- **What I used to believe:**
  - students need to memorize full paradigms in order to achieve fluency

- **What I now believe:**
  - full paradigms are a fiction and we can strategically target acquisition of morphology

Challenges of rich inflectional morphology

- Most languages have rich inflectional morphology
- Rich inflectional morphology can involve huge numbers of forms
- Inflectional morphology can be complex

Most languages have rich inflectional morphology: A sample from Europe

HIGH	**Slavic:** Bosnian, Croatian, Czech, Montenegran, Polish, Russian, Rusyn, Serbian, Slovak, Slovene, Sorbian **Other Indoeuropean:** Albanian, Greek, Irish, Latvian, Lithuanian, Romany, Scottish Gaelic, Welsh **Uralic :** Estonian, Finnish, Hungarian, Kven, Saami **Other:** Basque, Maltese, Turkish
MID	**Slavic:** Bulgarian, Macedonian **Other Indoeuropean:** Catalan, French, German, Italian, Luxembourgish, Portuguese, Spanish, Yiddish
LOW	**Germanic språk:** Danish, Dutch, English, Norwegian, Swedish

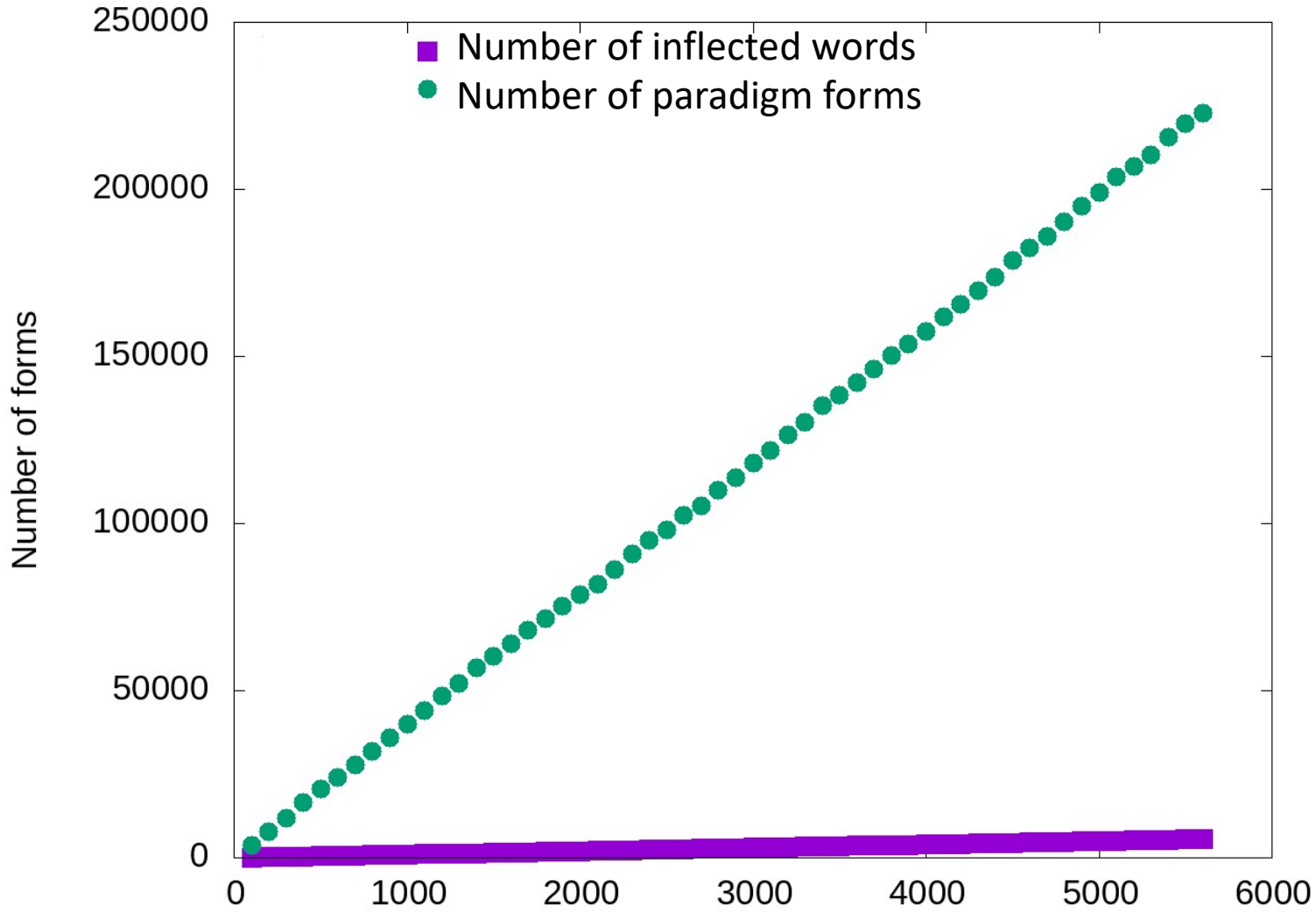
Rich inflectional morphology can involve huge numbers of forms

Kolik řečí znáš, tolikrát jsi člověkem
'How many languages you know – that many times you are a person'

Word	Gloss	Grammatical categories marked by morphology
Kolik	How many	Indefinite numeral in Accusative case
řečí	language	Feminine noun in Genitive case and Plural number
znáš	know	Imperfective verb in Present tense, Second person Singular
tolikrát	that many times	Adverb
jsi	be	Imperfective verb in Present tense, Second person Singular
člověkem	person	Masculine noun in Instrumental case and Singular number

Inflectional morphology in a Czech proverb

In a language with rich inflectional morphology, even a small basic vocabulary can entail **hundreds of thousands of paradigm forms.**


Number of inflected words
Number of paradigm forms
Number of forms
250000
200000
150000
100000
50000
0
0
1000
2000
3000
4000
5000
6000

Inflectional morphology can be complex

- Morphology is considered both to be essential to L2 acquisition and to be a “bottleneck”, more difficult than both syntax and semantics (Slabakova 2009 & 2014, Jensen et al. 2019)
- Morphophonemic alternations can complicate the picture – inflectional morphology is not always a matter of adding desinences to stems
  - consonant and vowel alternations
  - fleeting vowels (jers)
  - suprasegmental alternations

North Saami example

	Basic paradigm of *guoibmi* "partner"	
	NOM.SG	*guoibmi*
	GEN.SG=ACC.SG	*guoimmi*
	ILL.SG	*guoibmá-i*
	LOC.SG	*guoimmi-s*
	COM.SG=LOC.PL	*guimmi-in*
	NOM.PL	*guoimmi-t*
	GEN.PL=ACC.PL	*guimmi-id*
	ILL.PL	*guimmi-ide*
	COM.PL	*guimmi-iguin*
	ESS	*guoibmi-n*

Table 2. 81 additional paradigm forms required by NPx for *guoibmi* "partner"

NOM.SG:		GEN.SG=ACC.SG:		ILL.SG:	
1SG	*guoibmá-n*	1SG	*guoibmá-n*	1SG	*guoibmá-s-an*
2SG	*guoibmá-t*	2SG	*guoimmá-t*	2SG	*guoibmá-s-at*
3SG	*guoibmi-s*	3SG	*guoimmi-s*	3SG	*guoibmá-s-is*
1DU	*guoibmá-me*	1DU	*guoibmá-me*	1DU	*guoibmá-s-eame*
2DU	*guoibmá-de*	2DU	*guoimmá-de*	2DU	*guoibmá-s-eatte*
3DU	*guoibmi-ska*	3DU	*guoimmi-ska*	3DU	*guoibmá-s-easkka*
1PL	*guoibmá-met*	1PL	*guoibmá-met*	1PL	*guoibmá-s-eamet*
2PL	*guoibmá-det*	2PL	*guoimmá-det*	2PL	*guoibmá-s-eattet*
3PL	*guoibmi-set*	3PL	*guoimmi-set*	3PL	*guoibmá-s-easet*
LOC.SG:		COM.SG=LOC.PL:		GEN.PL=ACC.PL (=NOM.PL 1SG/DU/PL):	
1SG	*guoimmi-st-an*	1SG	*guimmi-in-an*	1SG	*guimmi-id-an*
2SG	*guoimmi-st-at*	2SG	*guimmi-in-at*	2SG	*guimmi-id-at*
3SG	*guoimmi-st-is*	3SG	*guimmi-in-is*	3SG	*guimmi-id-is*
1DU	*guoimmi-st-eame*	1DU	*guimmi-in-eame*	1DU	*guimmi-id-eame*
2DU	*guoimmi-st-eatte*	2DU	*guimmi-in-eatte*	2DU	*guimmi-id-eatte*
3DU	*guoimmi-st-easkka*	3DU	*guimmi-in-easkka*	3DU	*guimmi-id-easkka*
1PL	*guoimmi-st-eamet*	1PL	*guimmi-in-eamet*	1PL	*guimmi-id-eamet*
2PL	*guoimmi-st-eattet*	2PL	*guimmi-in-eattet*	2PL	*guimmi-id-eattet*
3PL	*guoimmi-st-easet*	3PL	*guimmi-in-easet*	3PL	*guimmi-id-easet*
ILL.PL:		COM.PL:		ESS:	
1SG	*guimmi-idas-an*	1SG	*guimmi-id-an-guin*	1SG	*guoibmi-n-an*
2SG	*guimmi-idas-at*	2SG	*guimmi-id-at-guin*	2SG	*guoibmi-n-at*
3SG	*guimmi-idas-as*	3SG	*guimmi-id-is-guin*	3SG	*guoibmi-n-is*
1DU	*guimmi-idas-ame*	1DU	*guimmi-id-eame-guin*	1DU	*guoibmi-n-eame*
2DU	*guimmi-idas-ade*	2DU	*guimmi-id-eatte-guin*	2DU	*guoibmi-n-eatte*
3DU	*guimmi-idas-aska*	3DU	*guimmi-id-easkka-guin*	3DU	*guoibmi-n-easkka*
1PL	*guimmi-idas-amet*	1PL	*guimmi-id-eamet-guin*	1PL	*guoibmi-n-eamet*
2PL	*guimmi-idas-adet*	2PL	*guimmi-id-eattet-guin*	2PL	*guoibmi-n-eattet*
3PL	*guimmi-idas-aset*	2PL	*guimmi-id-easet-guin*	3PL	*guoibmi-n-easet*

Zipfian distribution

Language & Corpus Name	Corpus Size	Paradigm Size	Total Lexemes	Lexemes with full Paradigm	% Lexemes with full Paradigm
English Web Treebank	254,830	2	6,369	1,524	23.92%
Norwegian Dependency Treebank	311,277	4	12,587	393	3.12%
Russian SynTagRus	**1,032,644**	**12**	**21,945**	**13**	**0.06%**
Czech Prague Dependency Treebank	1,509,242	14	17,904	3	0.02%
Estonian ArborEst	234,351	28	14,075	0	0%

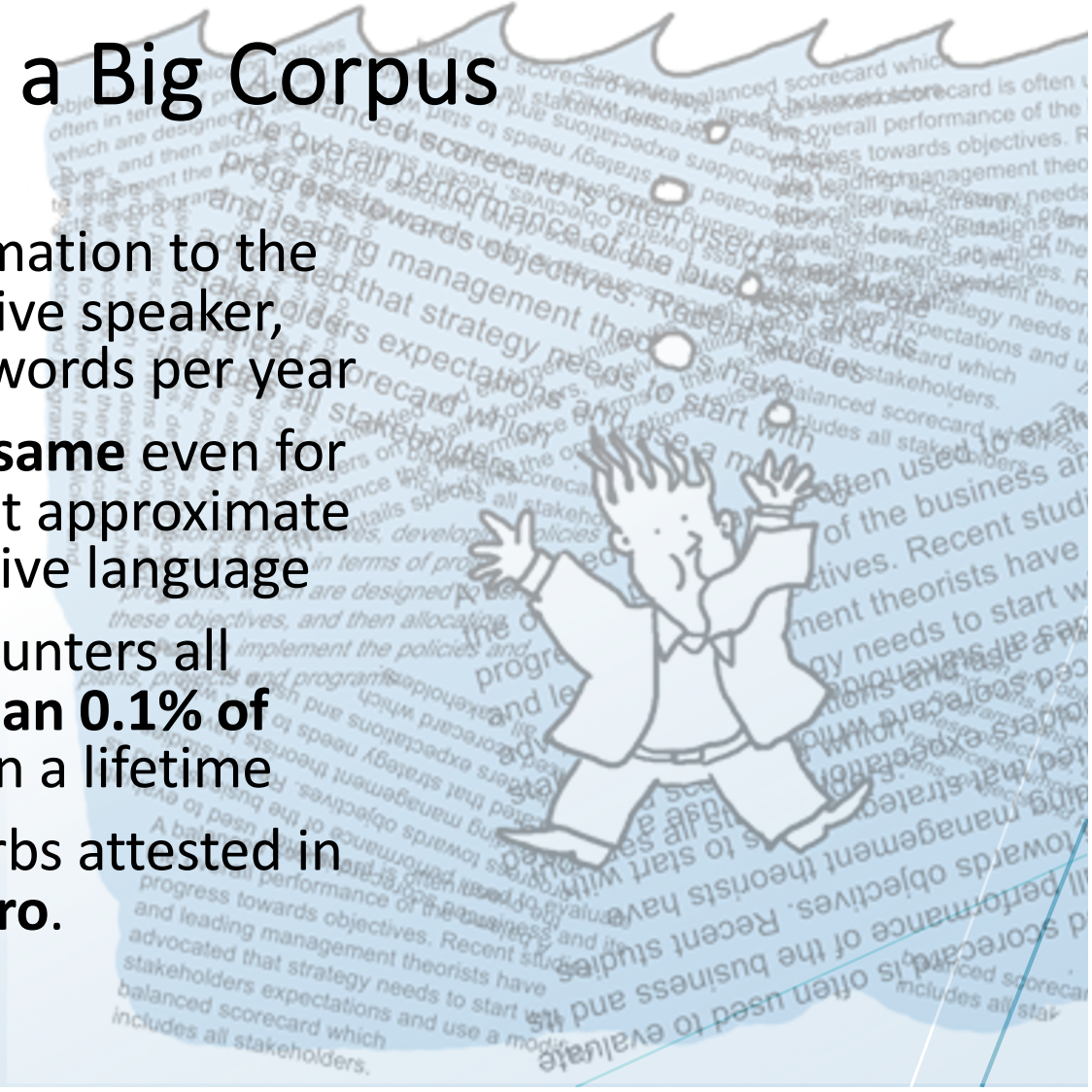
Zipfian distribution

Language & Corpus Nam	Corpus Size	Paradigm Size	Total Lexemes	Lexemes with full Paradigm	% Lexemes with full Paradigm
E				,524	23.92%
				93	3.12%
				13	**0.06%**
				3	0.02%
Estonian ArborEst	254,591	28	14,079	0	0%

Because Zipf's Law scales up, these numbers will never change substantially, no matter how large the corpus is

Language Exposure as a Big Corpus

- A large corpus is a close approximation to the **lifetime linguistic input** for a native speaker, estimated at about 5-10 million words per year
- Zipfian distributions **remain the same** even for very large corpora, like those that approximate a speaker's exposure to their native language
- A native speaker of Russian encounters all twelve paradigm forms of **less than 0.1% of nouns** that they are exposed to in a lifetime
- The portion of adjectives and verbs attested in all paradigm forms is **virtually zero**.



What this means for words

- At the word level, each lemma has a unique “grammatical profile” -- frequency distribution of its paradigm forms

Full paradigms

	‘fear’	*‘soldier’*	*‘department’*	*‘concept’*	*‘memory’*
sg.nom	страх	солдат	отделение	концепция	память
sg.gen	страха	солдата	отделения	концепции	памяти
sg.dat	страху	солдату	отделению	концепции	памяти
sg.acc	страх	солдата	отделение	концепцию	память
sg.ins	страхом	солдатом	отделением	концепцией	памятью
sg.loc	страхе	солдате	отделении	концепции	памяти
pl.nom	страхи	солдаты	отделения	концепции	памяти
pl.gen	страхов	солдат	отделений	концепций	памятей
p.dat	страхам	солдатам	отделениям	концепциям	памятям
pl.acc	страхи	солдат	отделения	концепции	памяти
pl.ins	страхами	солдатами	отделениями	концепциями	памятями
pl.loc	страхах	солдатах	отделениях	концепциях	памятях

Paradigm forms attested in 1M words

	'fear'	*'soldier'*	*'department'*	*'concept'*	*'memory'*
sg.nom	**страх**	**солдат**	отделение	**концепция**	память
sg.gen	**страха**	солдата	**отделения**	**концепции**	**памяти**
sg.dat	страху	солдату	отделению	концепции	памяти
sg.acc	страх	солдата	отделение	концепцию	память
sg.ins	страхом	солдатом	отделением	концепцией	памятью
sg.loc	страхе		отделении	концепции	памяти
pl.nom	страхи	**солдаты**	отделения		
pl.gen	страхов	**солдат**	отделений	концепций	
p.dat		солдатам			
pl.acc	страхи	солдат	отделения	концепции	
pl.ins	страхами		отделениями	концепциями	
pl.loc	страхах	солдатах	отделениях		

bold >20%, plain >10%, gray 1-9%, (blank) not attested

Full paradigms

	‘background’	*‘champion’*	*‘expanse’*	*‘frame’*	*‘problem’*
sg.nom	фон	чемпион	протяжение	рамка	трудность
sg.gen	фона	чемпиона	протяжения	рамки	трудности
sg.dat	фону	чемпиону	протяжению	рамке	трудности
sg.acc	фон	чемпиона	протяжение	рамку	трудность
sg.ins	фоном	чемпионом	протяжением	рамкой	трудностью
sg.loc	фоне	чемпионе	протяжении	рамке	трудности
pl.nom	фоны	чемпионы	протяжения	рамки	трудности
pl.gen	фонов	чемпионов	протяжений	рамок	трудностей
p.dat	фонам	чемпионам	протяжениям	рамкам	трудностям
pl.acc	фоны	чемпионов	протяжения	рамки	трудности
pl.ins	фонами	чемпионами	протяжениями	рамками	трудностями
pl.loc	фонах	чемпионах	протяжениях	рамках	трудностях

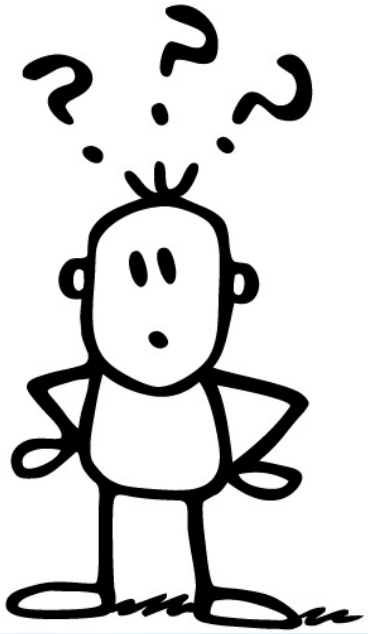
Paradigm forms attested in 1M words

	'background'	*'champion'*	*'expanse'*	*'frame'*	*'problem'*
sg.nom	фон	чемпион			трудность
sg.gen	фона	**чемпиона**			трудности
sg.dat		чемпиону			трудности
sg.acc		чемпиона			трудность
sg.ins		**чемпионом**			трудностью
sg.loc	**фоне**		**протяжении**		
pl.nom		чемпионы		рамки	**трудности**
pl.gen		чемпионов		рамок	**трудностей**
p.dat		чемпионам			
pl.acc		чемпионов		рамки	**трудности**
pl.ins		чемпионами		рамками	трудностями
pl.loc				**рамках**	трудностях

bold >20%, plain >10%, gray 1-9%, (blank) not attested

Paradigm Cell Filling Problem

(Ackerman et al. 2009)



Native **speakers** of languages with **complex inflectional morphology** routinely **recognize** and **produce** forms that they have **never encountered.**

Example:

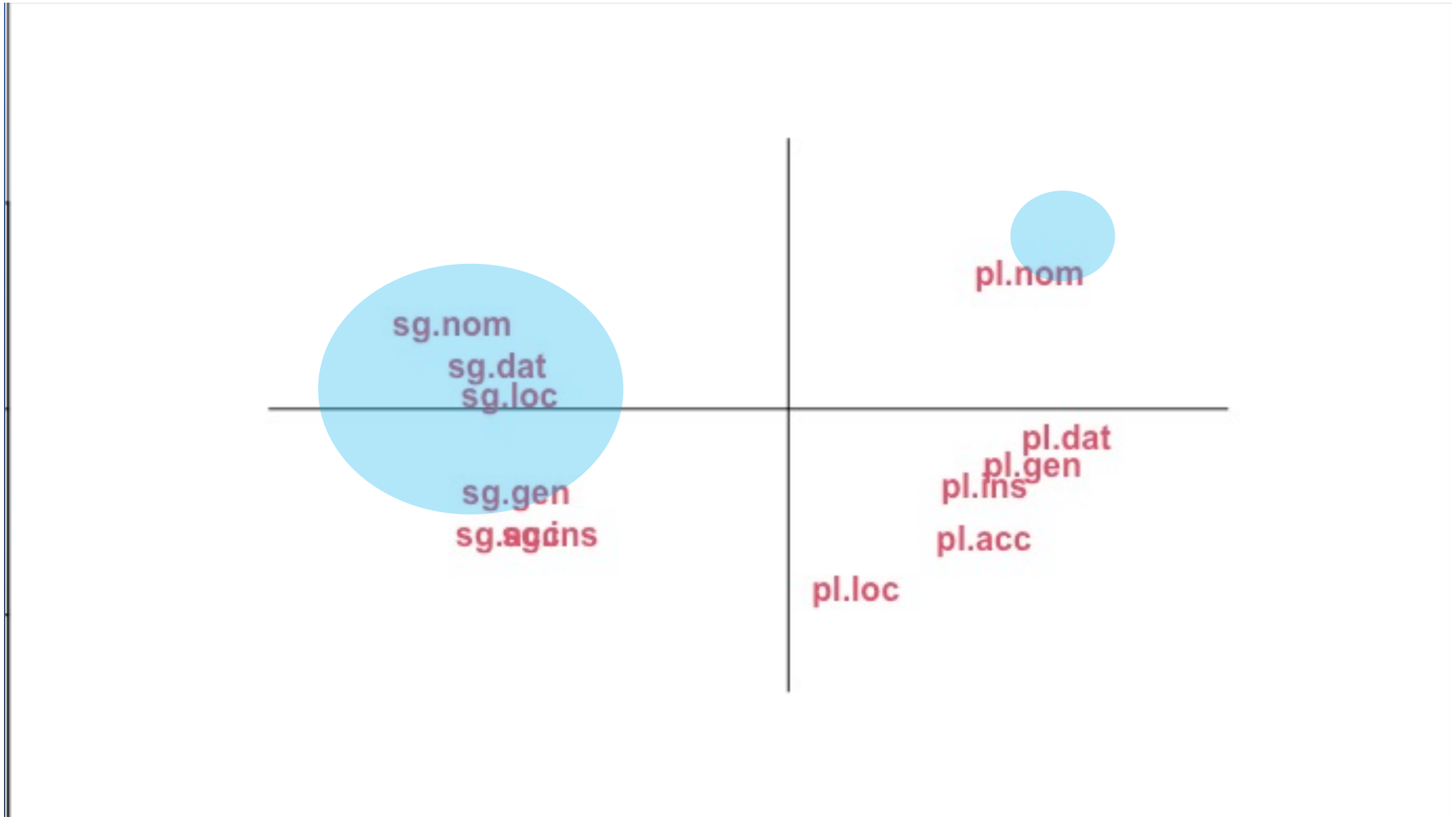
Russian gerund *nedokarmlivaja* ‘while underfeeding’ has no attestations in the Russian National Corpus (>360M words ≈ lifetime exposure), but all native speakers of Russian can be expected to readily understand and to produce these forms in appropriate contexts, as evidenced by rare occurrences that turn up in Google searches.

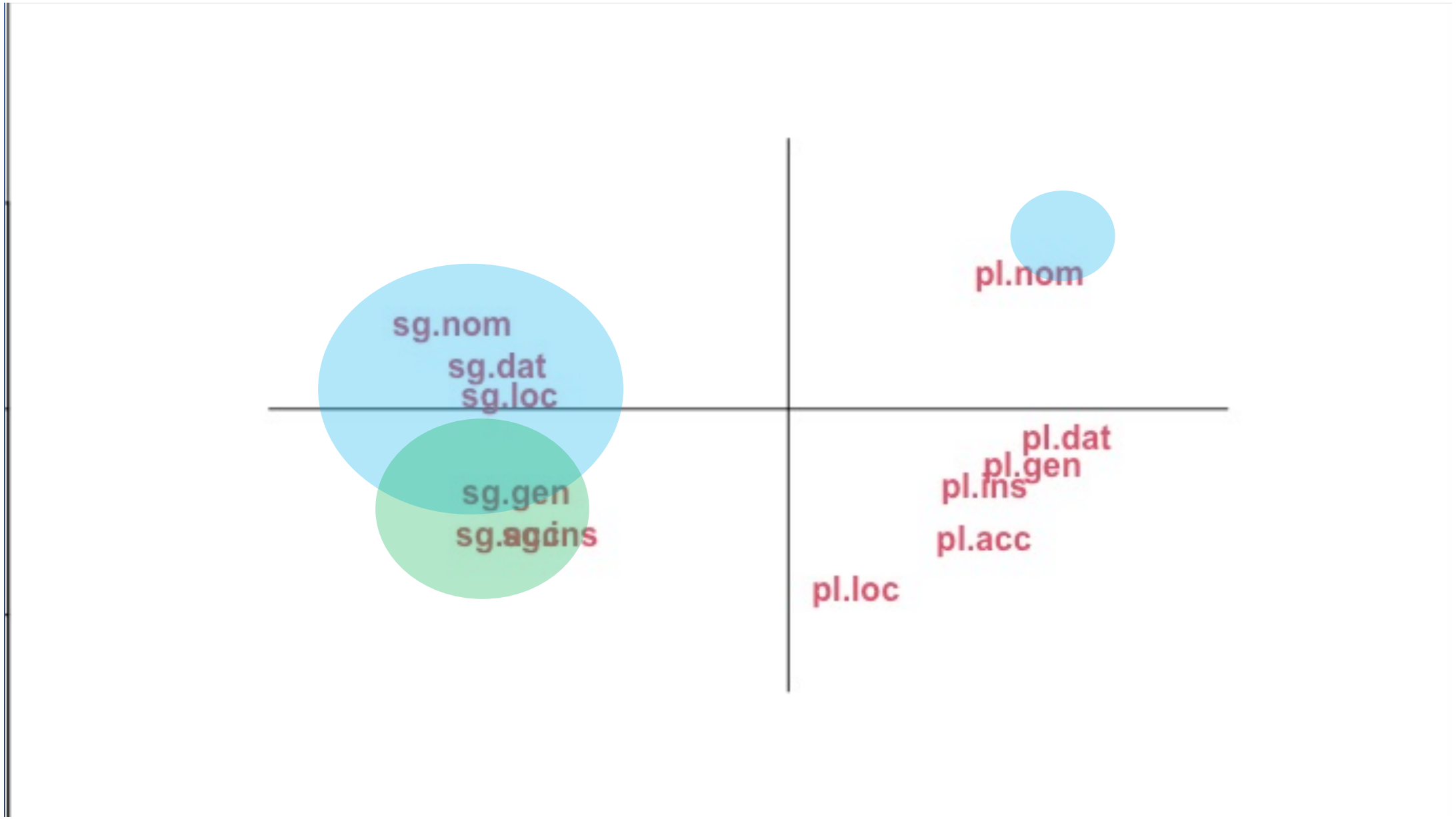
A data-based model of rich morphology

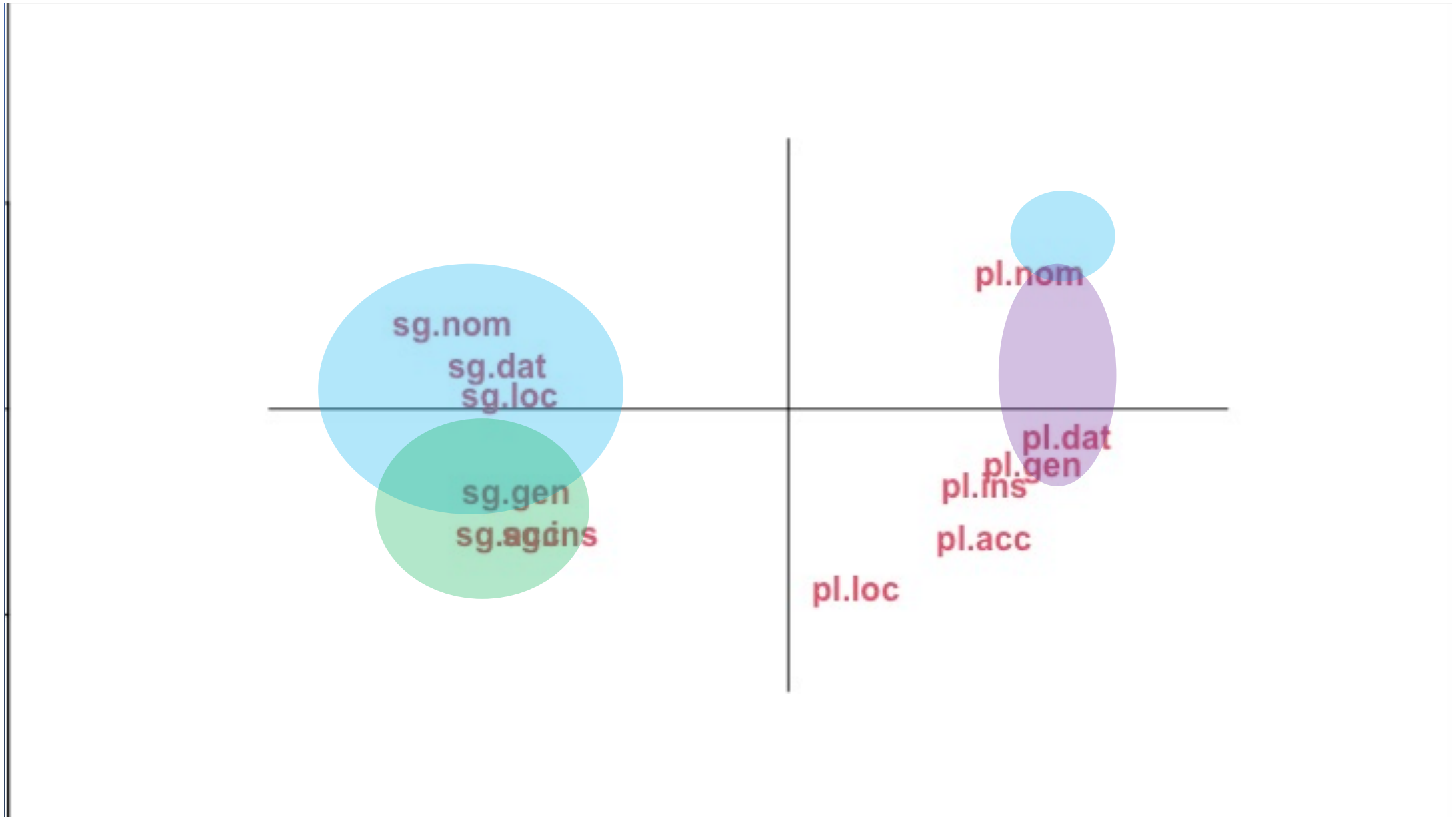
Instead of a table of equiprobable cells:

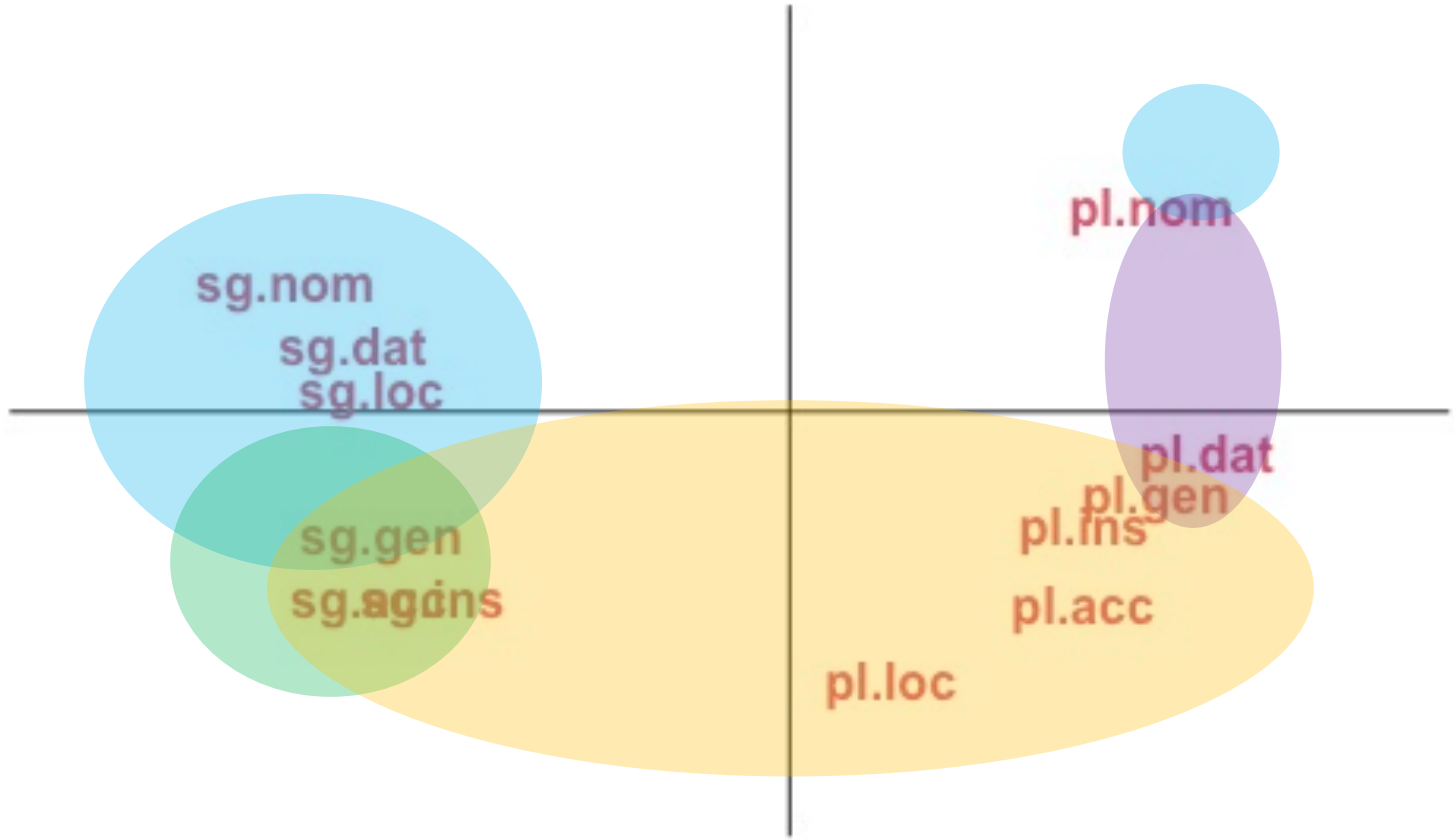
- A multidimensional space that is **partially populated in many different ways** by many different lexemes
- Inflectional morphology can be mastered through exposure to **partially overlapping subsets of paradigms**
- This **cognitively plausible** model explains why native speakers have the intuition that full paradigms exist

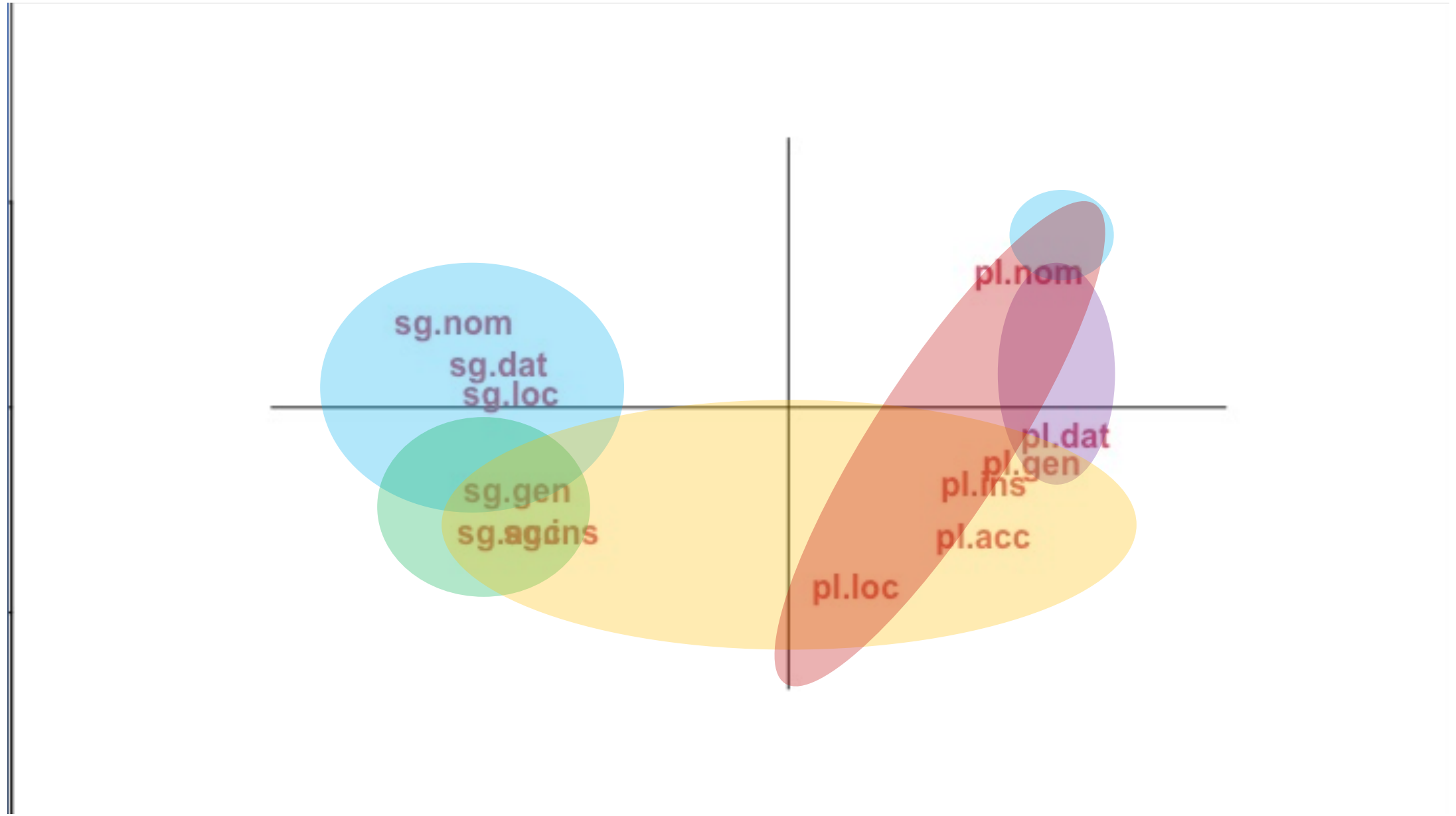
The “space” of the paradigm of Russian masculine animate nouns as determined by corpus data

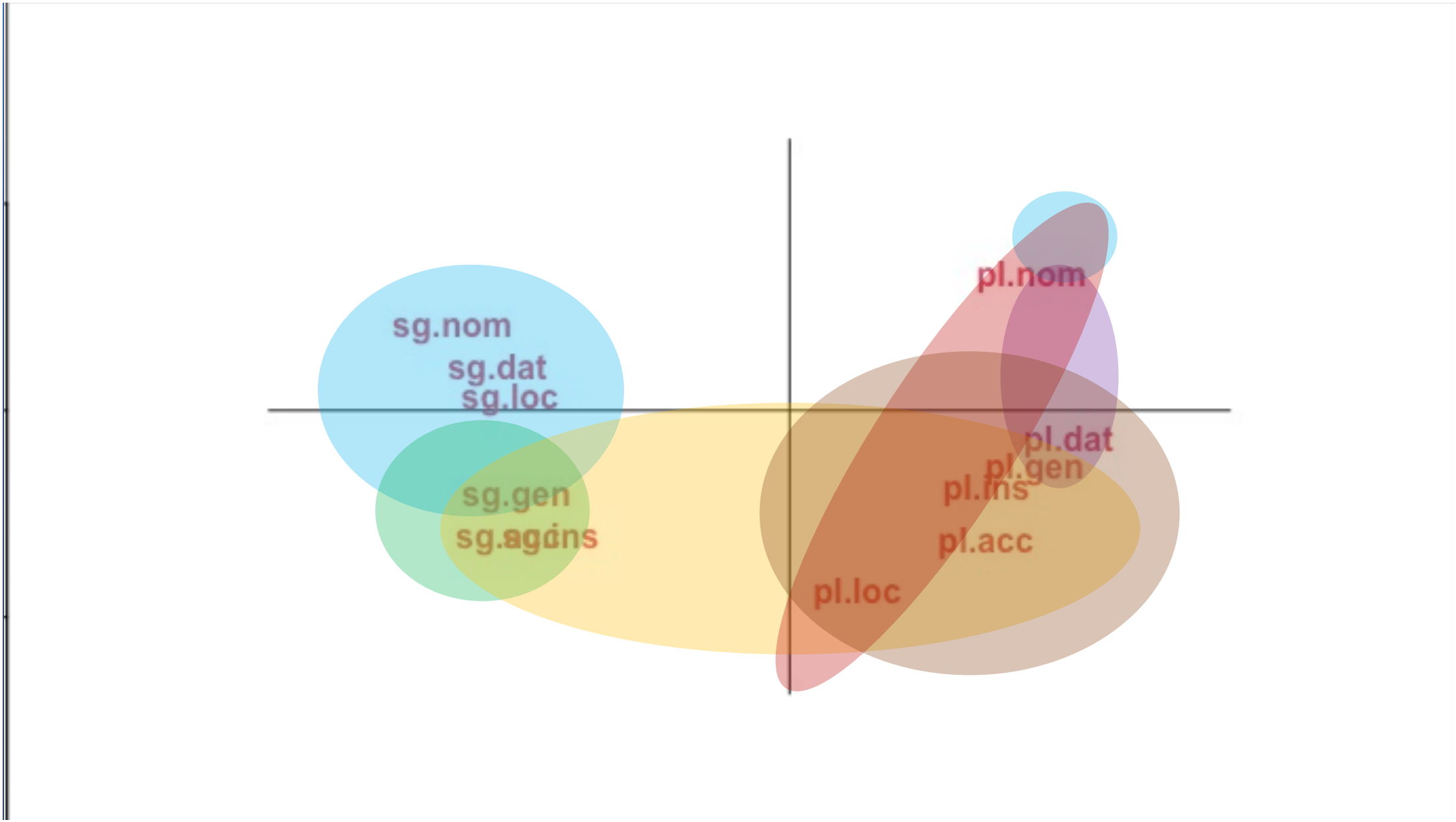

pl.nom
sg.nom
sg.dat
sg.loc
pl.dat
pl.gen
pl.ins
sg.gen
pl.acc
pl.loc

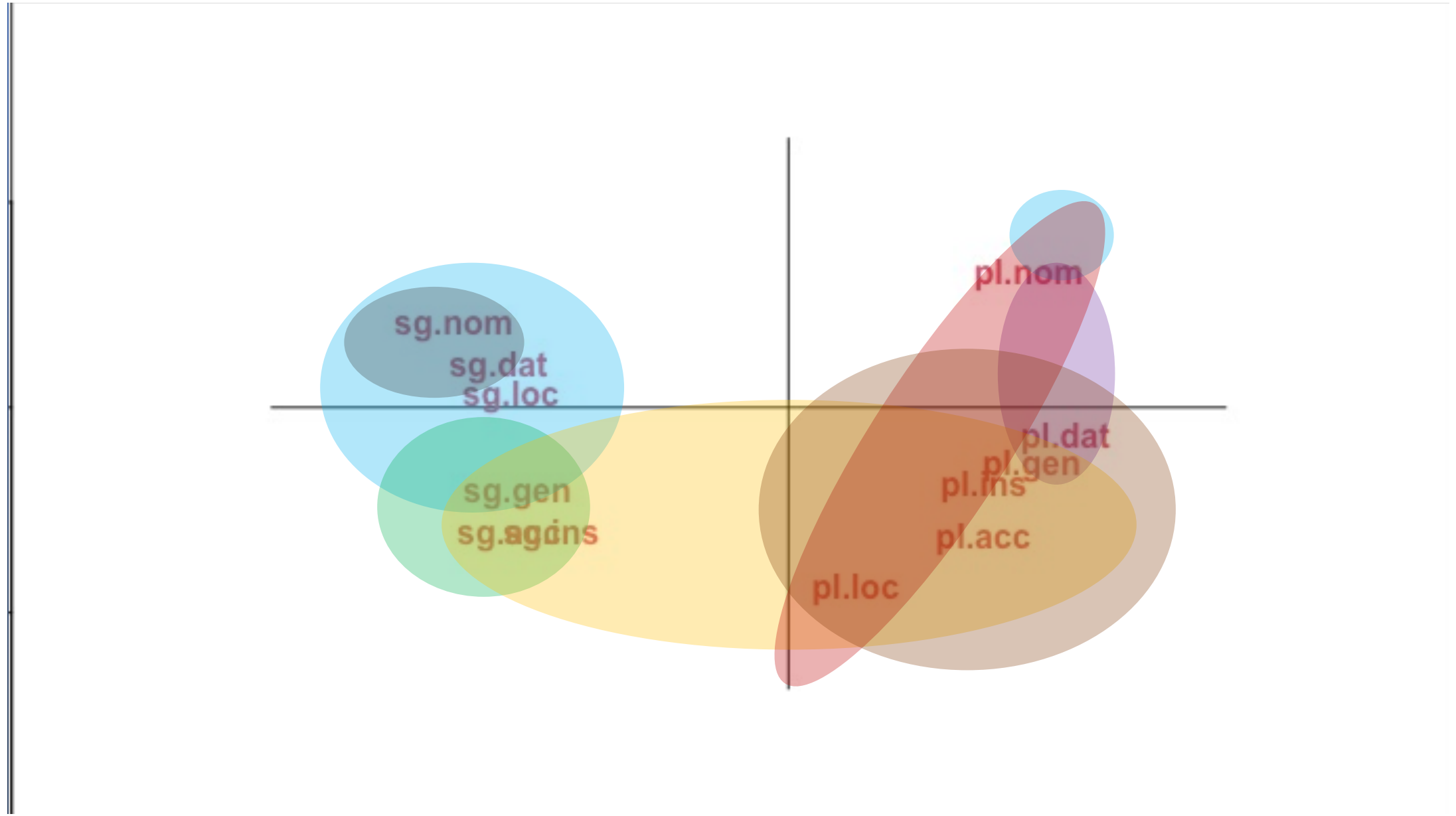

pl.nom
sg.nom
sg.dat
sg.loc
pl.dat
pl.gen
pl.ins
sg.gen
pl.acc
pl.loc

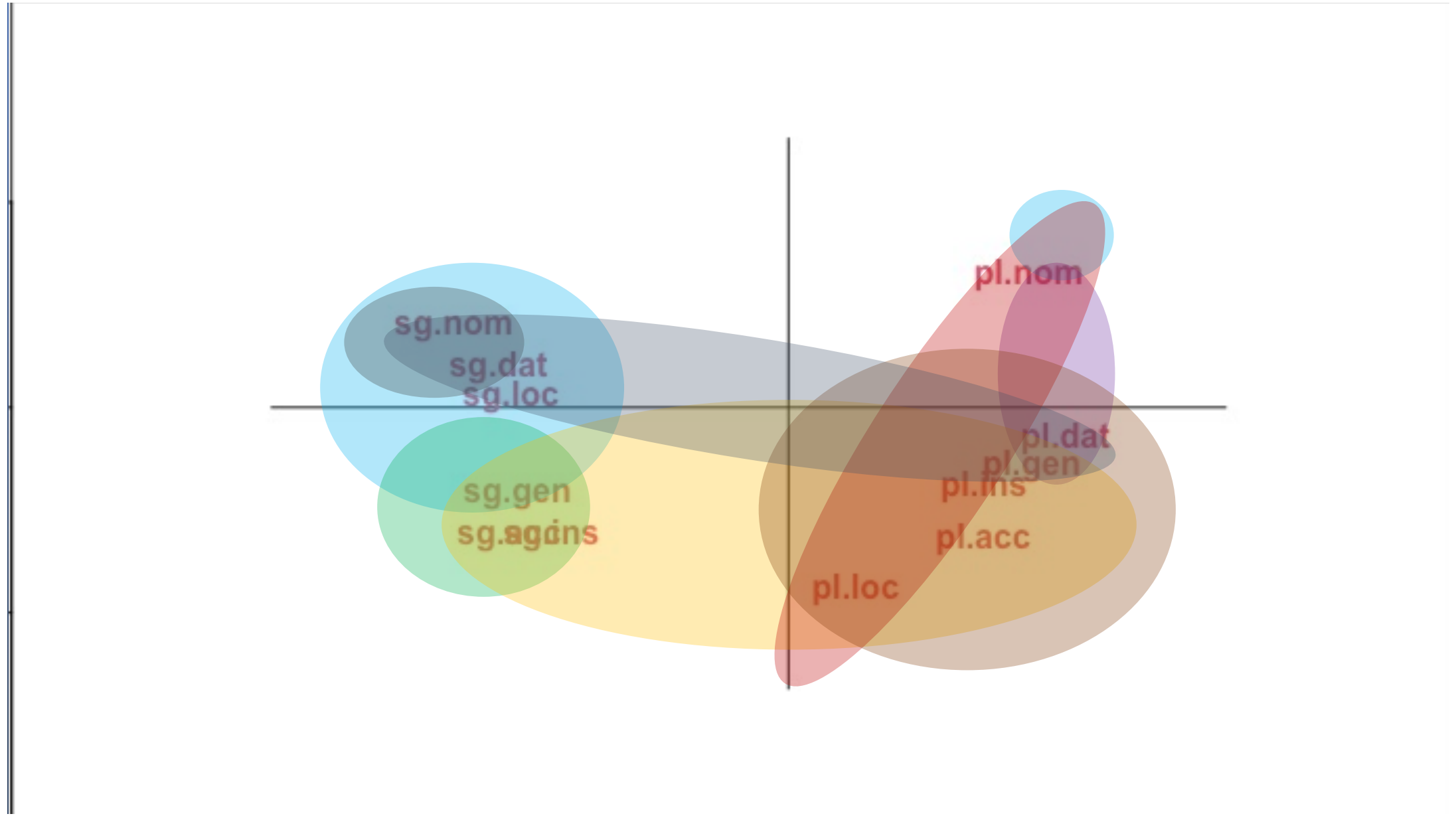

sg.nom
sg.dat
sg.loc
sg.gen
sg.acc
sg.ins
pl.nom
pl.dat
pl.gen
pl.ins
pl.acc
pl.loc

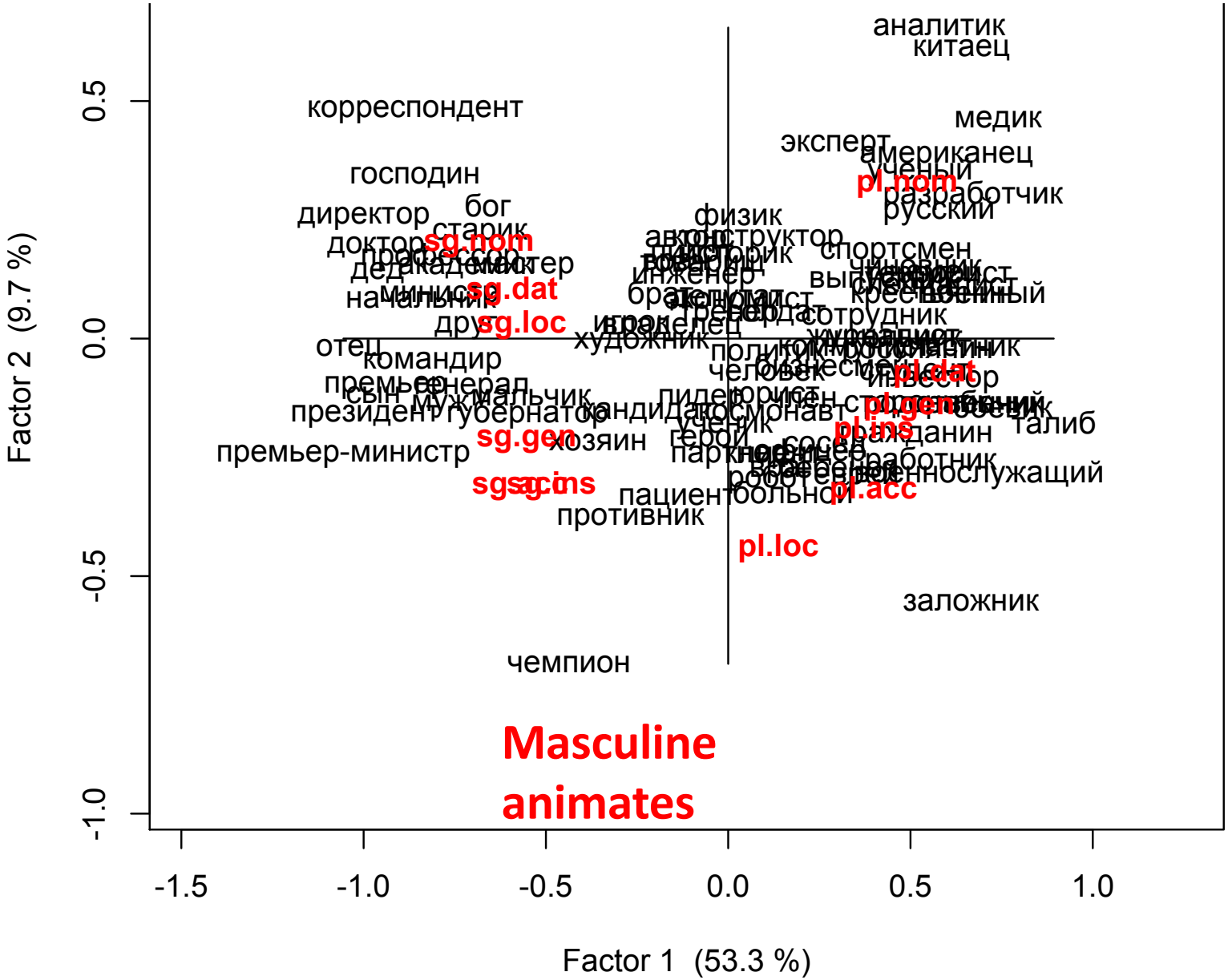

pl.nom
sg.nom
sg.dat
sg.loc
pl.dat
pl.gen
pl.ins
sg.gen
pl.acc
pl.loc


sg.nom
sg.dat
sg.loc
sg.gen
sg.acc
sg.ins
pl.nom
pl.dat
pl.gen
pl.ins
pl.acc
pl.loc

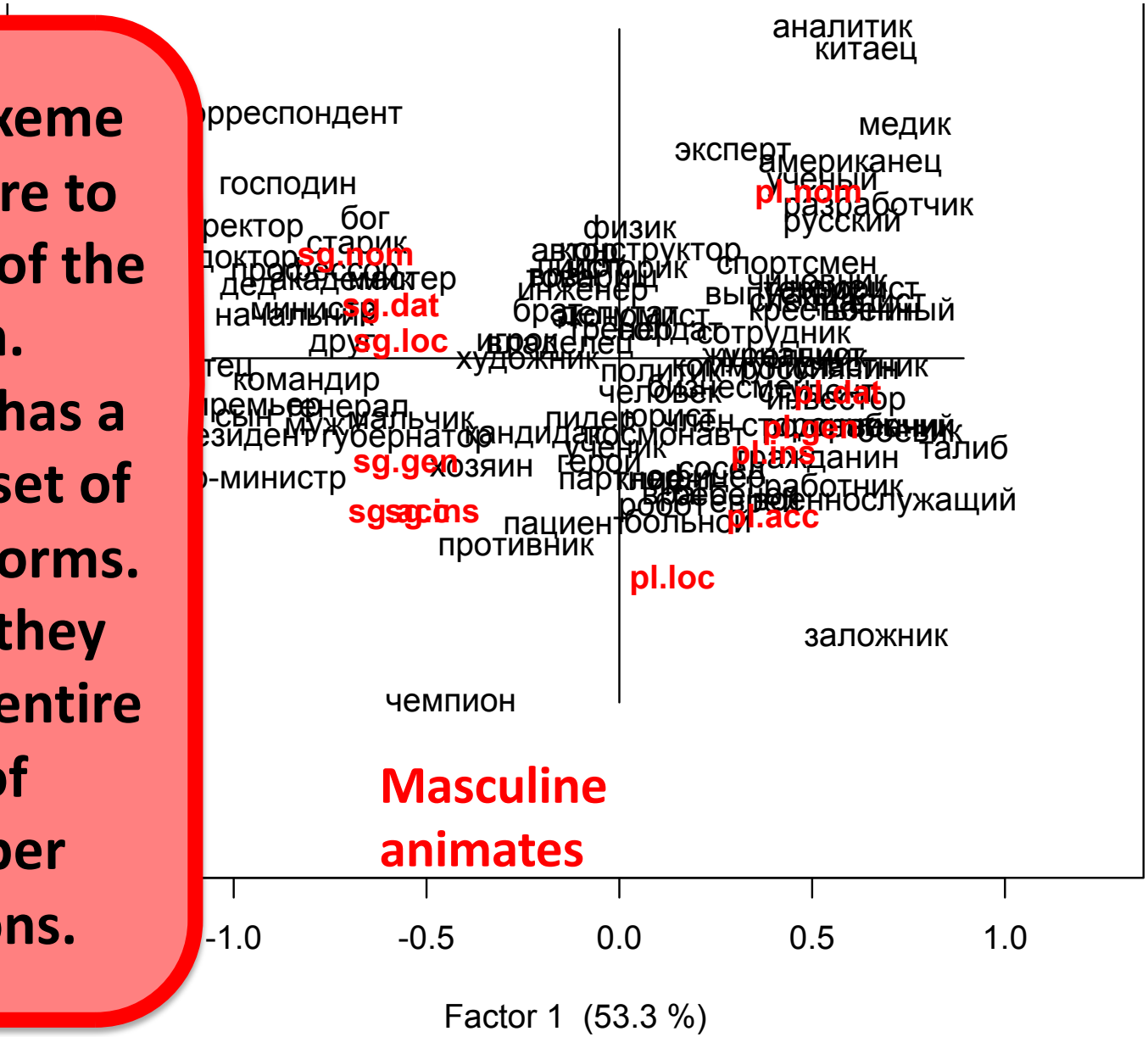

sg.nom
sg.dat
sg.loc
sg.gen
sg.acc
sg.ins
pl.nom
pl.dat
pl.gen
pl.ins
pl.acc
pl.loc

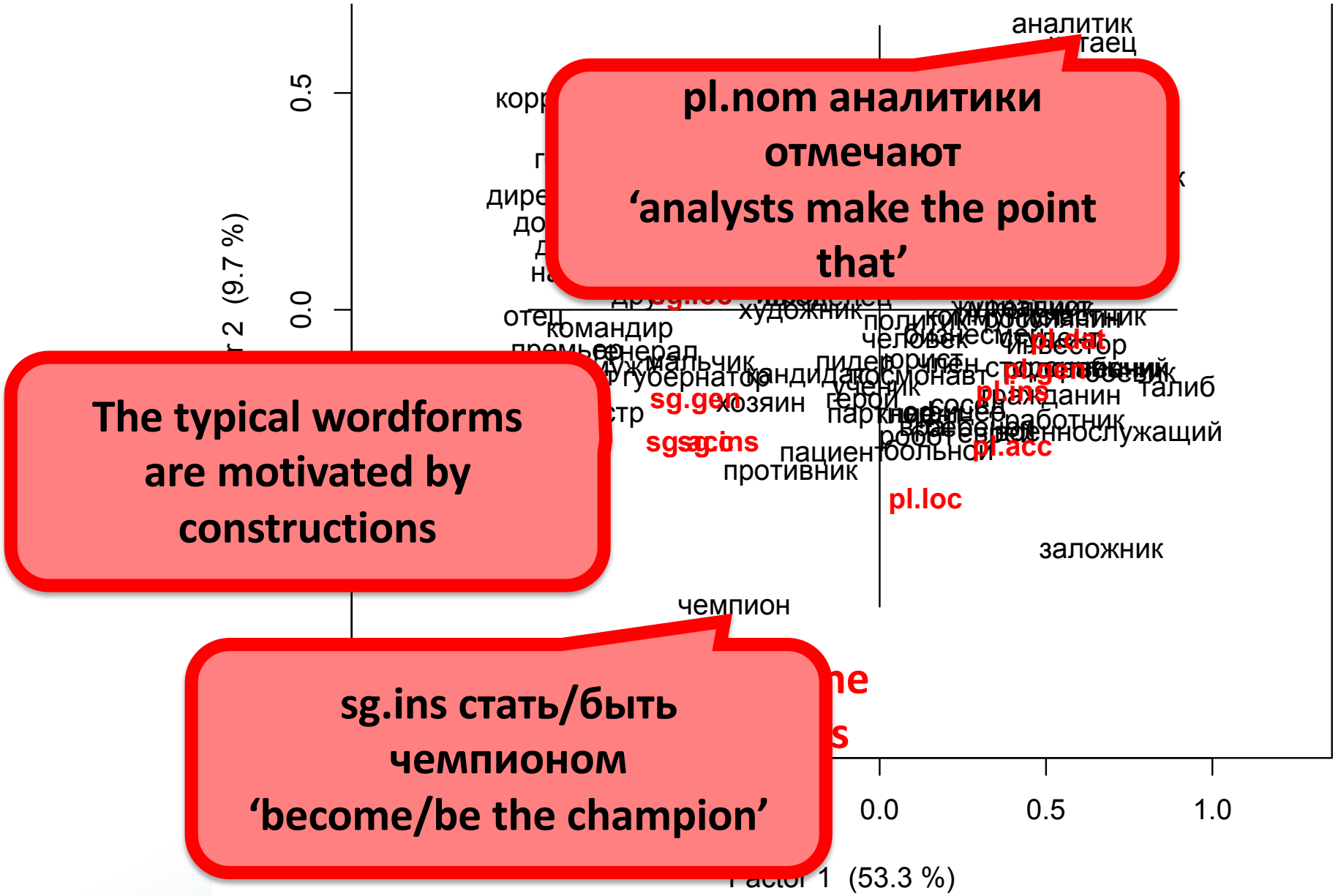

sg.nom
sg.dat
sg.loc
sg.gen
pl.nom
pl.dat
pl.gen
pl.ins
pl.acc
pl.loc


pl.nom
sg.nom
sg.dat
sg.loc
pl.dat
pl.gen
pl.ins
sg.gen
pl.acc
pl.loc


аналитик
китаец
медик
эксперт
американец
ученый
pl.nom
разработчик
русский
корреспондент
господин
директор
бог
старик
физик
спортсмен
sg.nom
sg.dat
sg.loc
инженер
сотрудник
художник
отец
командир
человек
инвестор
pl.dat
pl.gen
pl.ins
талиб
президент
губернатор
кандидат
космонавт
ученик
sg.gen
хозяин
гражданин
премьер-министр
работник
военнослужащий
sg.acc
pl.acc
пациент
больной
противник
pl.loc
заложник
чемпион
Masculine animates
Factor 2 (9.7 %)
Factor 1 (53.3 %)
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

Any single lexeme gives exposure to only a subset of the paradigm.
Each lexeme has a different subset of most typical forms.
Collectively they populate the entire "space" of case/number combinations.



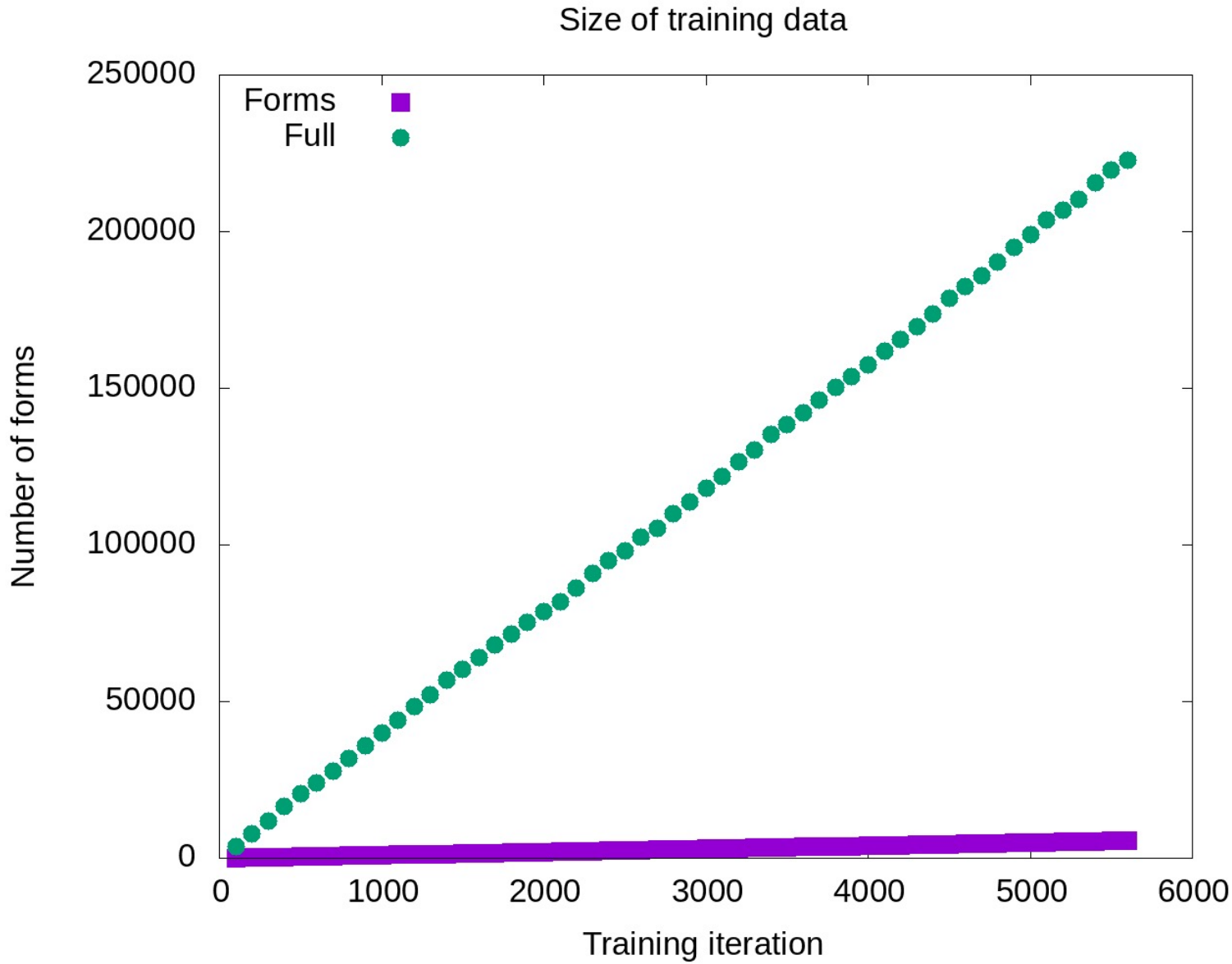

pl.nom аналитики отмечают ‘analysts make the point that’
The typical wordforms are motivated by constructions
sg.ins стать/быть чемпионом ‘become/be the champion’
аналитик
чемпион
заложник
sg.gen
pl.loc
pl.acc
pl.ins
0.0
0.5
1.0
Factor 1 (53.3 %)
Factor 2 (9.7 %)

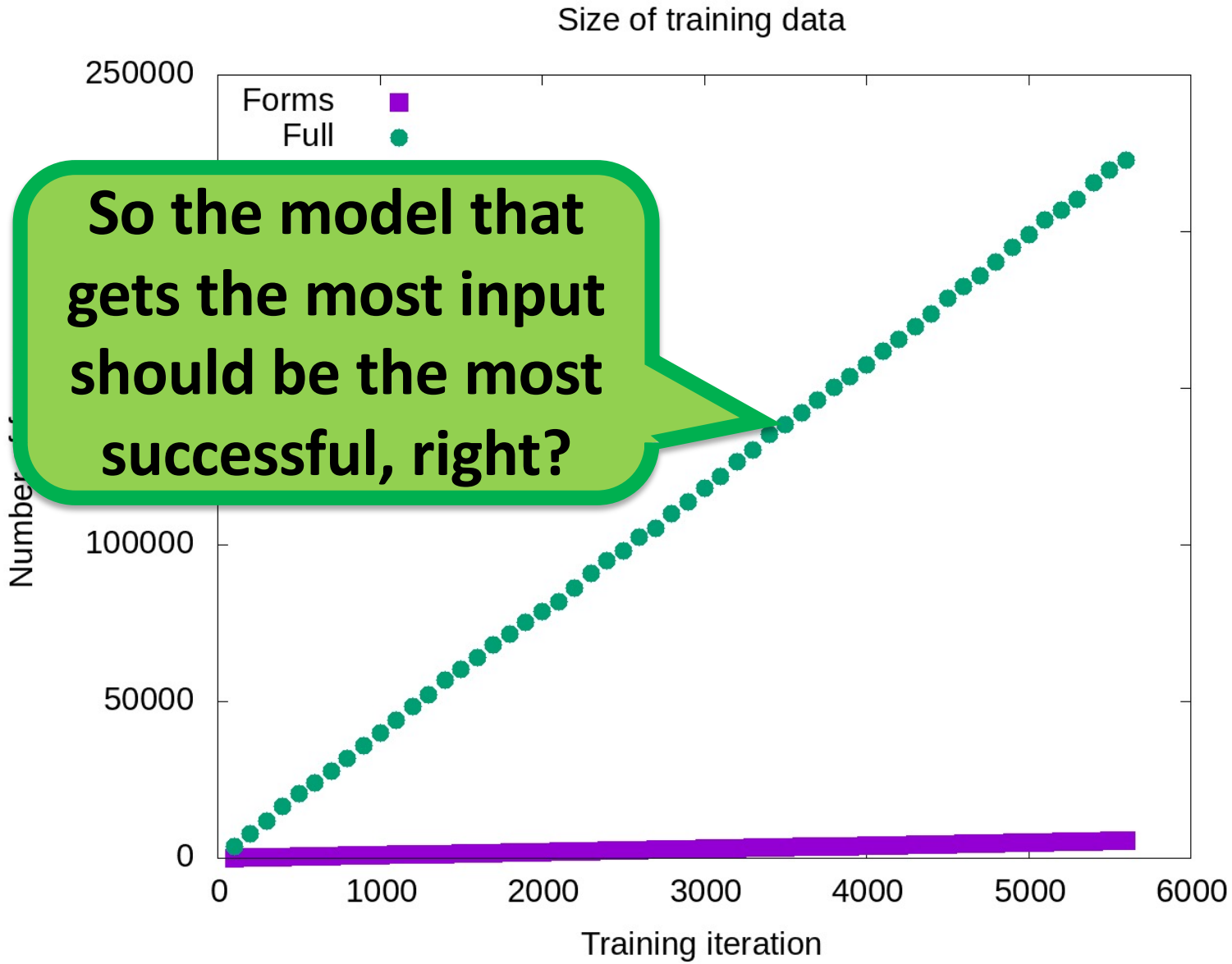
A machine-learning experiment

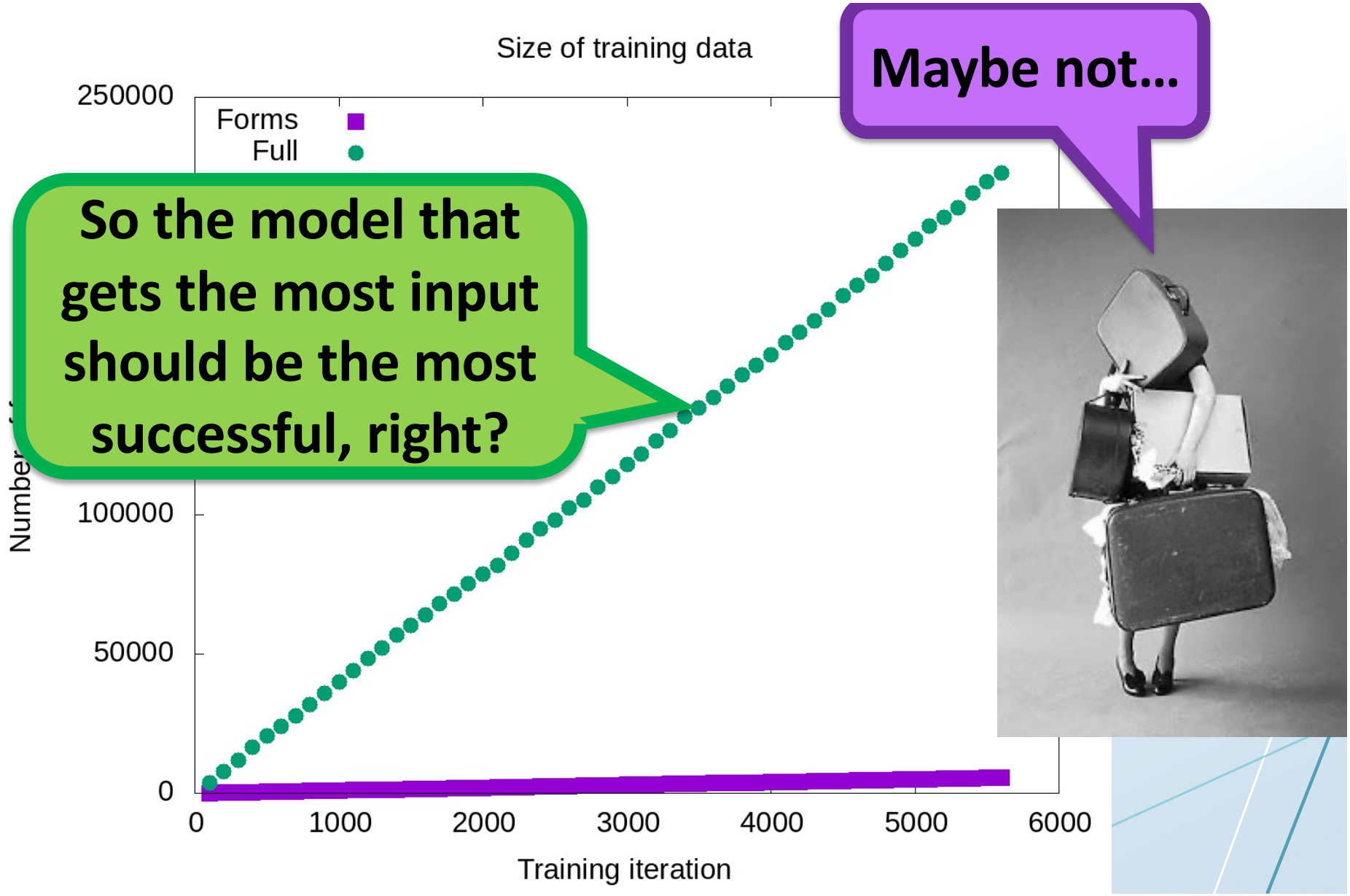
- Based on an ordered list of the most frequent forms for nouns, verbs, and adjectives in SynTagRus (1M gold standard)
- Machine learning: training, testing
  - Given the 100 most frequent forms, predict the next 100 most frequent forms
  - Given the 200 most frequent forms, predict the next 100 most frequent forms
  - Given the 300 most frequent forms, predict the next 100 most frequent forms
  - … until 5400, when SynTagRus runs out of data
- Testing is always: Predict paradigm forms for 100 previously unseen words
- Two versions of experiment:
  - Training on entire paradigm, **all forms**
  - Training only on the **single most frequent form**

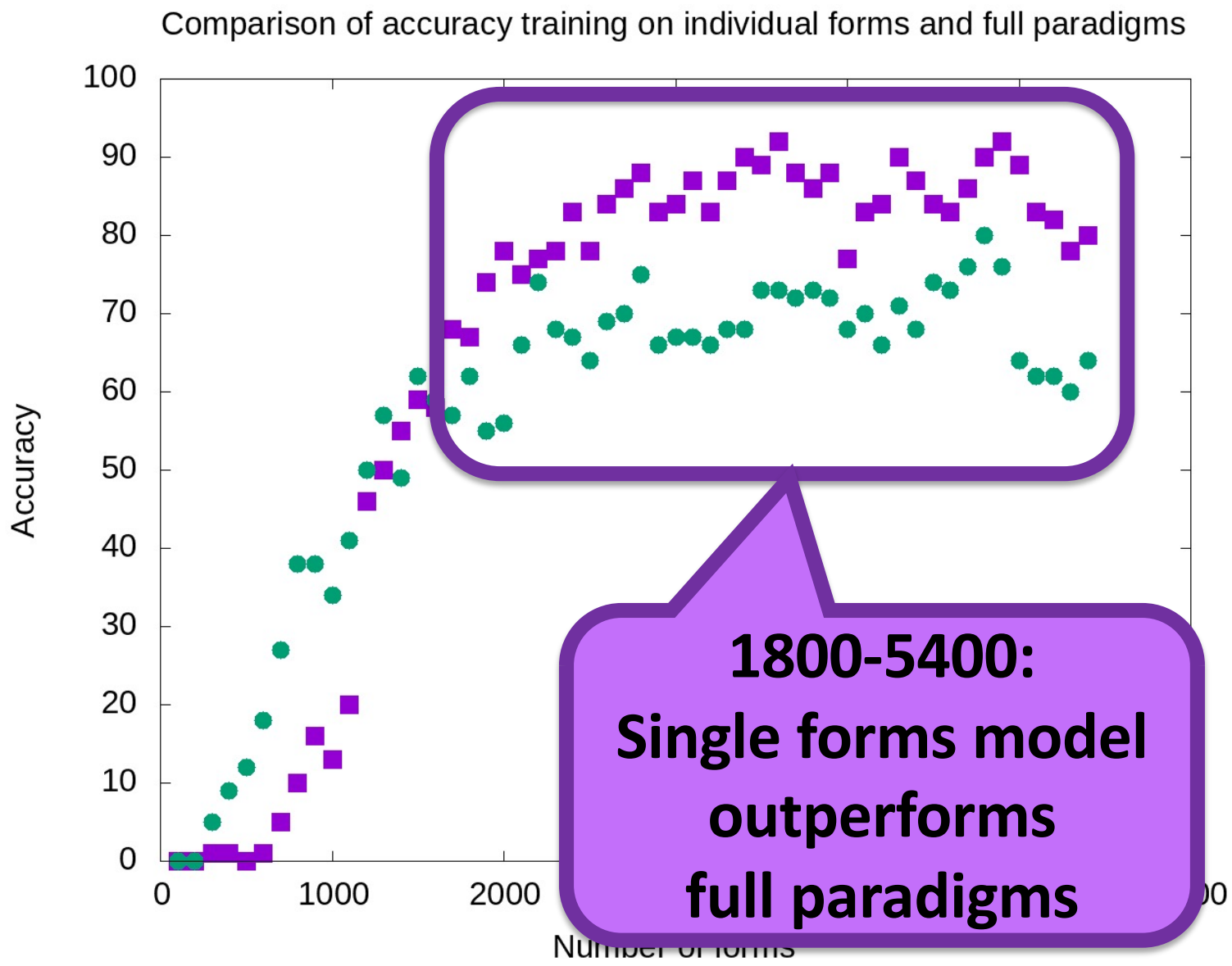
Data for training and testing from SynTagRus

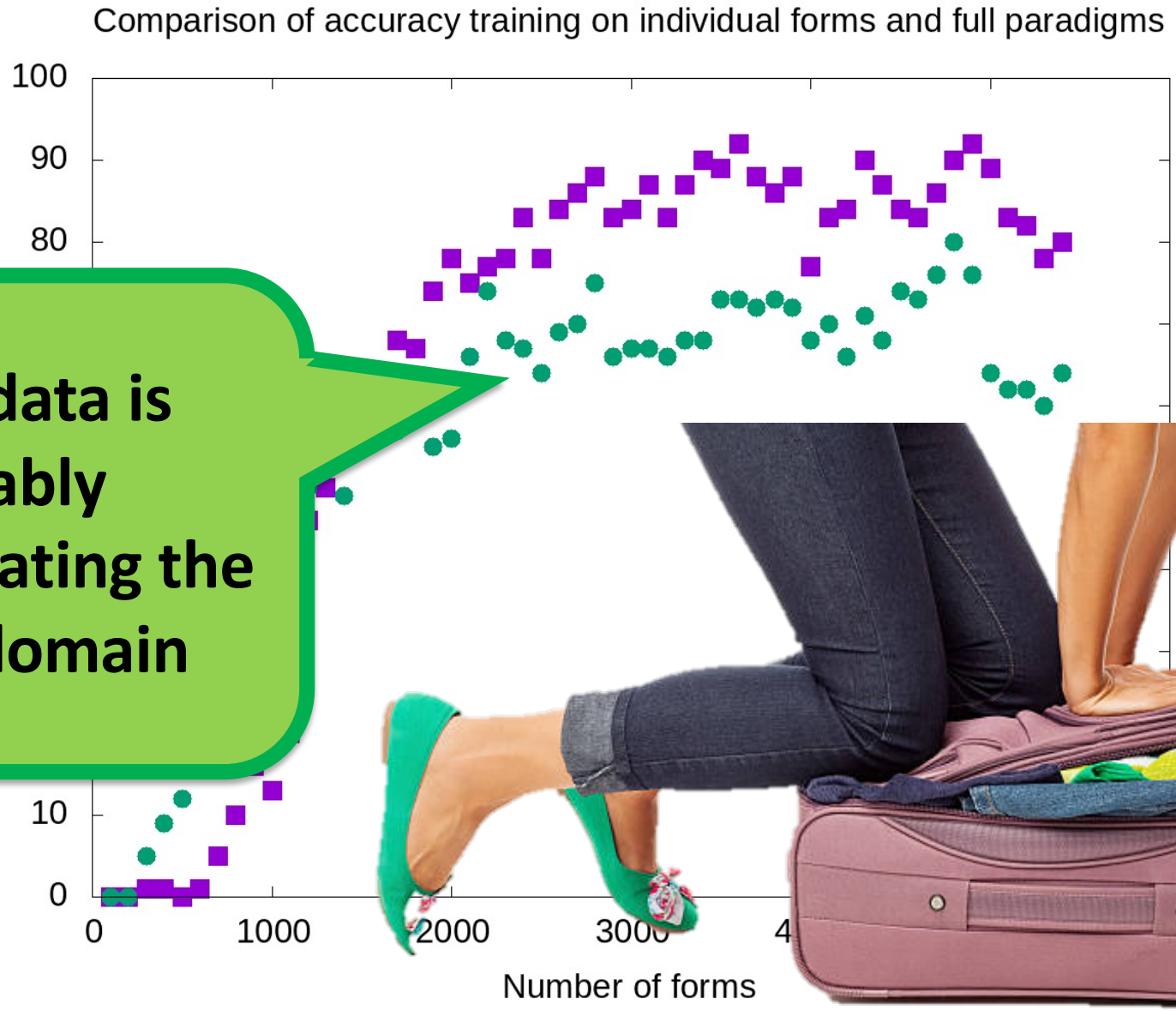
Frequency & Form	Lemma	POS	Parse of form
1447 может	мочь	VERB	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act
1286 года	год	NOUN	Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing
999 лет	год	NOUN	Animacy=Inan|Case=Gen|Gender=Masc|Number=Plur
832 году	год	NOUN	Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing
813 время	время	NOUN	Animacy=Inan|Case=Acc|Gender=Neut|Number=Sing
678 россии	россия	NOUN	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing
571 могут	мочь	VERB	Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act
571 люди	человек	NOUN	Animacy=Anim|Case=Nom|Gender=Masc|Number=Plur
543 россии	россия	NOUN	Animacy=Inan|Case=Loc|Gender=Fem|Number=Sing
436 является	являться	VERB	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act
416 случае	случай	NOUN	Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing
411 людей	человек	NOUN	Animacy=Anim|Case=Gen|Gender=Masc|Number=Plur
403 страны	страна	NOUN	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing
400 жизни	жизнь	NOUN	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing


Size of training data
Forms
Full
Number of forms
Training iteration
250000
200000
150000
100000
50000
0
0
1000
2000
3000
4000
5000
6000


Size of training data
250000
Forms
Full
100000
50000
0
0
1000
2000
3000
4000
5000
6000
Training iteration
So the model that gets the most input should be the most successful, right?

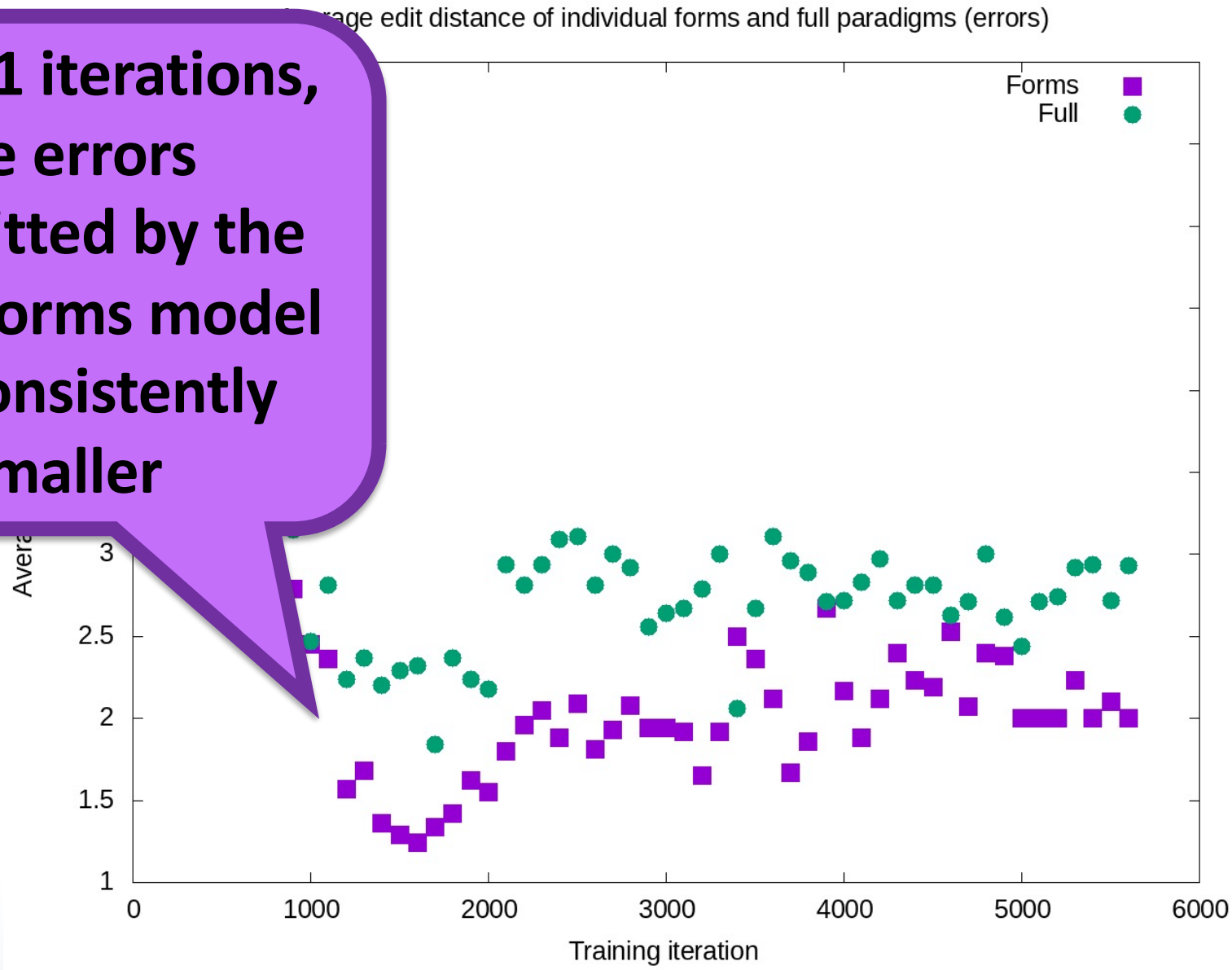

Size of training data
250000
Forms
Full
100000
50000
0
0
1000
2000
3000
4000
5000
6000
Training iteration
So the model that gets the most input should be the most successful, right?
Maybe not…


Comparison of accuracy training on individual forms and full paradigms
Accuracy
100
90
80
70
60
50
40
30
20
10
0
0
1000
2000
1800-5400:
Single forms model
outperforms
full paradigms



Excess data is probably overpopulating the search domain

After 11 iterations, the errors committed by the single forms model are consistently smaller



Machine-learning experiment: Summary

- Learning is potentially enhanced by focus **only on the most typical wordforms** attested for each lexeme: **accuracy increases** and **severity of errors decreases**
- This finding is **consistent with a usage-based cognitively plausible model**

How Can We Escape From Overstuffed Paradigms?

- Textbooks have always focused on certain forms and constructions
- Now we can do this in a scientific, consistent way

SMARTool = Strategic Mastery of Russian Tool

The SMARTool is available at: http://uit-no.github.io/smartool/

Free web resource for L2 learners of Russian that implements findings of corpus research and a learning simulation experiment to optimize the acquisition of Russian vocabulary and morphology

Partially funded by SIU [Center for Internationalization of Education], now known as DIKU [Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education], in collaboration with the National Research University Higher School of Economics in Moscow

The SMARTool:

- **Inspired by research** on the distribution and simulated learning of Russian wordforms **(cognitively plausible)**
- Strategic focus on the **highest frequency wordforms and contexts** that motivate their use **(usage-based)**
- **Reduces the task** of learning a basic vocabulary of about 3,000 lexemes **by over 90%**
- Can be **continuously updated** and custom-tailored
- Potentially **portable to other languages** with rich inflectional morphology

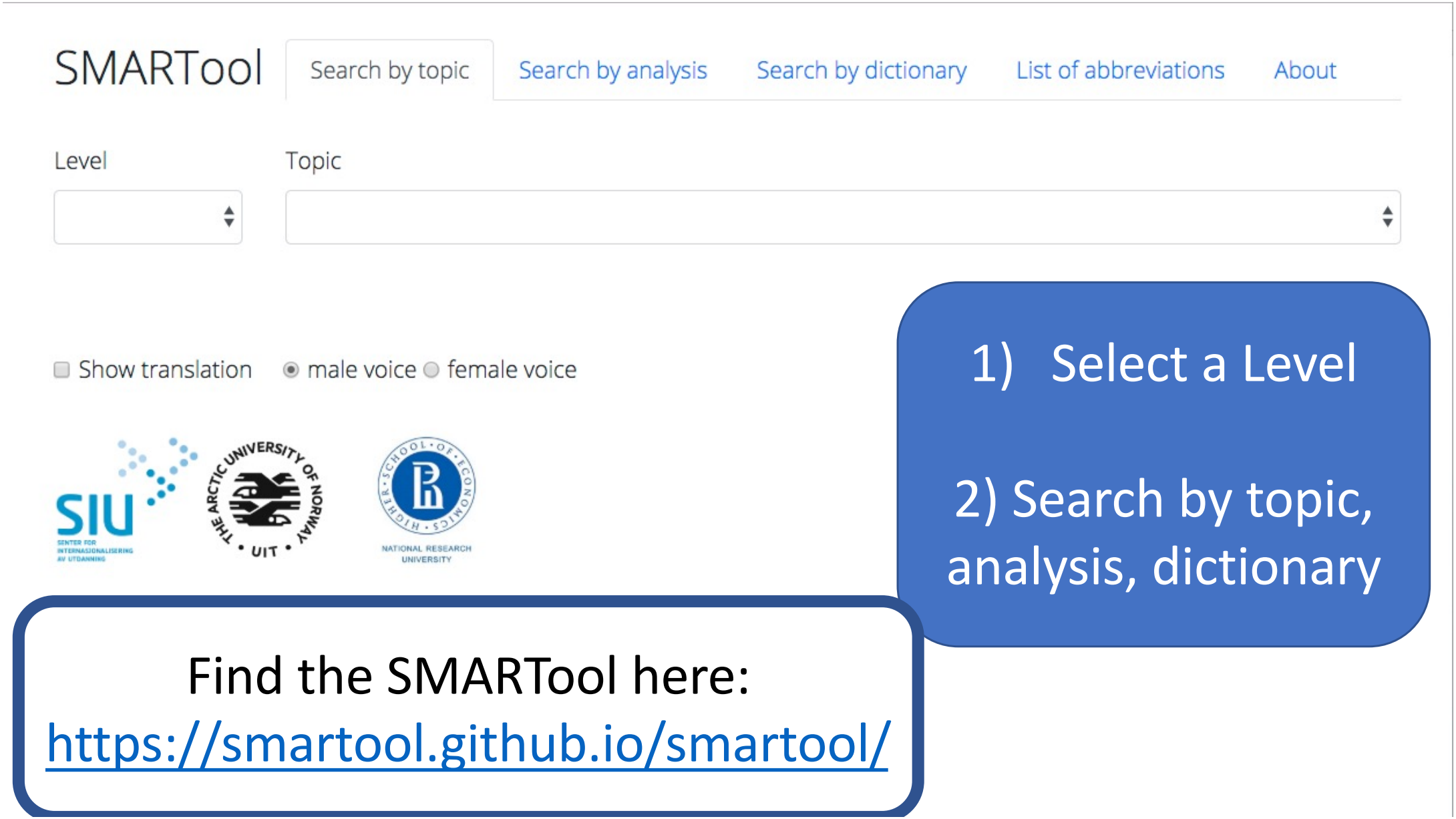
Vocabulary Selection from 5 Textbooks and Лексический минимум; Balanced for Nouns, Verbs, Adjs (RNC ratio)

CEFR Level	ACTFL Equivalent	Russian Equivalent	SMARTool number of lexemes
A1 “Beginner”	Novice Low-Mid	ТЭУ Элементарный уровень	500
A2 “Elementary”	Novice High	ТБУ Базовый уровень	500
B1 “Intermediate”	Intermediate Low-Mid	ТРКИ-1 I Сертификационный уровень	1,000
B2 “Upper Intermediate”	Intermediate High-Advanced Low	ТРКИ-2	1,000

Typical Contexts Illustrated by Examples

- For each lexeme we identify 3 most common wordforms and **most typical grammatical constructions** and **lexical collocations**, and provide corpus-inspired **example sentences**
- Based on queries:
  - SynTagRus Corpus
  - the Russian National Corpus (http://ruscorpora.ru)
  - the Collocations Colligations Corpora (http://cococo.cosyco.ru/)
  - the Russian Constructicon (https://constructicon.github.io/russian/)

RUSSIAN
CONSTRUCTICON


SMARTool
Search by topic
Search by analysis
Search by dictionary
List of abbreviations
About
Level
Topic
Show translation
male voice
female voice
SIU
THE ARCTIC UNIVERSITY OF NORWAY
UIT
HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY
1) Select a Level
2) Search by topic, analysis, dictionary
Find the SMARTool here:
https://smartool.github.io/smartool/

A PLACE FOR US:

https://dataverse.no/dataverse/trolling

TROLLing

- is an international archive of linguistic data and statistical code
- is built on the Dataverse platform from Harvard University and complies with DataCite, the international standard for storing and citing research data
- is compliant with CLARIN, the EU research infrastructure for language-based resources
- assigns a permanent URL to each post
- uses metadata that ensures visability and retrieval through international services
- is professionally managed by the University Library of Tromsø and an international steering committee

dataverse.no/dataverse/trolling

DataverseNO Open Research Data Certified by: CORE TRUST SEAL

Search ▾ Deposit Guide Support Log In

DataverseNO > **TROLLing**

Contact Share

Sign Up **Getting started with TROLLing**

CLARIN C CENTRE

Search this dataverse... Find Advanced Search

- [x] **Dataverses (0)**
- [x] **Datasets (111)**
- [ ] Files (2,998)

Publication Year

2014 (21)
2020 (19)
2016 (17)
2021 (14)
2017 (12)

More...

Author Name

1 to 10 of 111 Results Sort ▾

Replication Data for: Eye Behavior during Syntactic Movement Evidence for Processing Approach to Persian Syntax

Jun 28, 2021

Alaee, Majid; Rasekh-Mahand, Mohammad; Tehrani-Doost, Mahdi, 2021, "Replication Data for: Eye Behavior during Syntactic Movement Evidence for Processing Approach to Persian Syntax", https://doi.org/10.18710/TZBAOR, DataverseNO, V1

[Dataset abstract:] The datasets include experimental material on an eye-tracking analysis in Persian syntax and how syntactic movement in this language is affected by weight factor. It also includes statistical datasets and the results of statistical analysis. In addition, some...

Replication Data for: Davvisámi earutkeahtes oamasteapmi

Jun 24, 2021

Janda, Laura A; Antonsen, Lene, 2020, "Replication Data for: Davvisámi earutkeahtes oamasteapmi", https://doi.org/10.18710/QGXLQB, DataverseNO, V2, UNF:6:TEN1pDpmWcpB9xsXiJqcZg== [fileUNF]

Getting started with TROLLing

http://site.uit.no/trolling/getting-started/

- Promotional video
- Instructional videos
- User guide
- TROLLing banner

Basic steps in TROLLing

- Create an account
  - Needed only for archiving -- you do not need an account to search or download data
  - This step is self-explanatory, but there is an instructional video
  - It may take a day or two for your account to be approved
- Create a study
  - Enter metadata, upload files in persistent formats, get DOI
- Search for a study
  - TROLLing terms of use

After you have uploaded your study…

- You submit it for approval and receive an acknowledgement
- It will be approved and released by an administrator and you will receive an email
- You will be able to edit your study later if needed and resubmit
- Previous versions of your study are archived, but only the latest version shows up in initial searches

Search for a study

- All of the cataloging information (metadata) is searchable, including:
  - author
  - affiliation
  - country / nation
  - date of production and distribution
  - keywords, e.g. language
  - topic classification
- Advanced search
  - possible to include and exclude combinations of items

Terms of use that users must agree to when downloading files

Affim that:

- You will not use the materials that you downloaded to identify individuals or organizations
- You will not download or use materials that would be prohibited by law
- You will cite the data as stipulated in the Data Citation Information in any publications or reports that you make using the data