The corpus of contact-influenced Russian of Northern Siberia and the Russian Far East: Assessing a “post-pidgin continuum” in the circumpolar region

ICASS X, Arkhangelsk, 15.06.2021

Irina Khomchenkova[123] (irina.khomchenkova@yandex.ru)
Polina Pleshak[(1)5] (polinapleshak@yandex.ru)
Natalya Stoynova[124] (stoynova@yandex.ru)

1 - Institute of Linguistics, RAS;
2 - Vinogradov Russian Language Institute, RAS;
3 - Lomonosov Moscow State University;
4 - NRU HSE; Moscow, Russia
5 - University of Maryland

This study was supported by RSF grant №17-18-01649.

1. Introduction

- **Our corpus**:
  - “The corpus of contact-influenced Russian of Northern Siberia and the Russian Far East” (http://web-corpora.net/ruscontact/corpus.html)
  - a small spoken corpus
  - Russian speech of bilingual speakers of indigenous languages of Northern Siberia and the Russian Far East (mostly Samoyedic and Tungusic)
  - manual annotation of “contact-induced features”
- Total size: **262,159** words.
- Search on contact features available: **180,105** words.

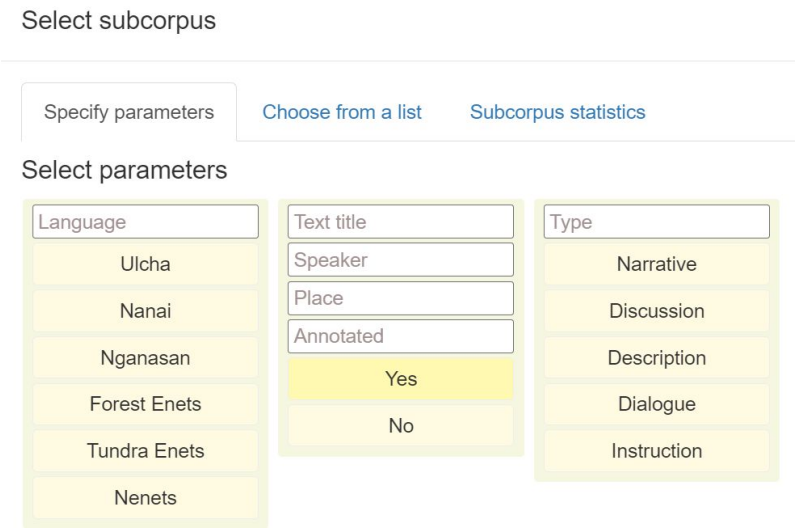
Within the project: “Dynamics of language contact in the circumpolar region” (http://iling-ran.ru/main/departments/typol_compar/circumpolar/).

1. Introduction

- The texts differ in the degree of deviation from monolinguals' Russian. The corpus represents in outline a post-pidgin continuum attested in the area.
- **Our aims:**
  → to assess this continuum, basing on our annotation of contact-induced features
  - to describe inter-speaker variation;
  - to reveal contact-induced features that determine this variation.

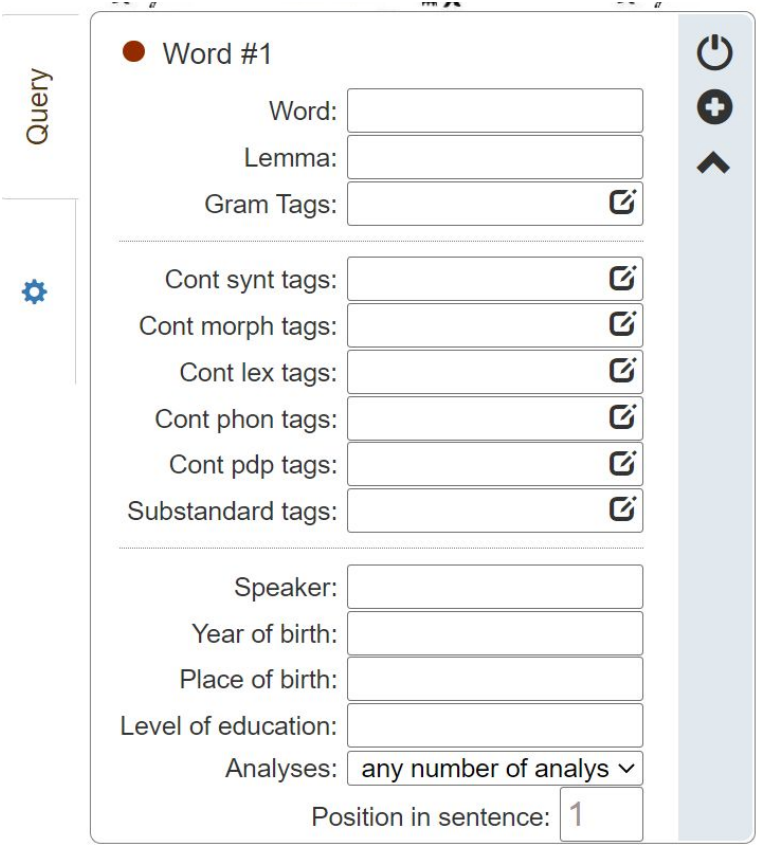
2. Corpus: texts

- **Spontaneous oral texts**:
  - short narratives (folklore, biographies);
  - descriptions (ethnographic texts, recipes, ...);
  - everyday dialogues with linguists.
- Collected by us and by our colleagues as a **“by-product”** of current language documentation projects.



2. Corpus: search

- Search on:
  - grammatical features
  - **contact features**
  - sociolinguistic information
- Output:
  - audio & transcription



2. Corpus: Contact tags

- **morphology** (including word-formation, inflection, use of grammatical categories): 10 tags
- **syntax** (clause-level): 24 tags, the most elaborated and frequently used
- **pdp** [polypredication, discourse, prosody]: 4 tags (complex sentences) + 5 tags (discourse) + 3 tags (prosody)
- **phonetics**: 18 tags, almost all of them are very specific
- **lexicon**: 3 tags

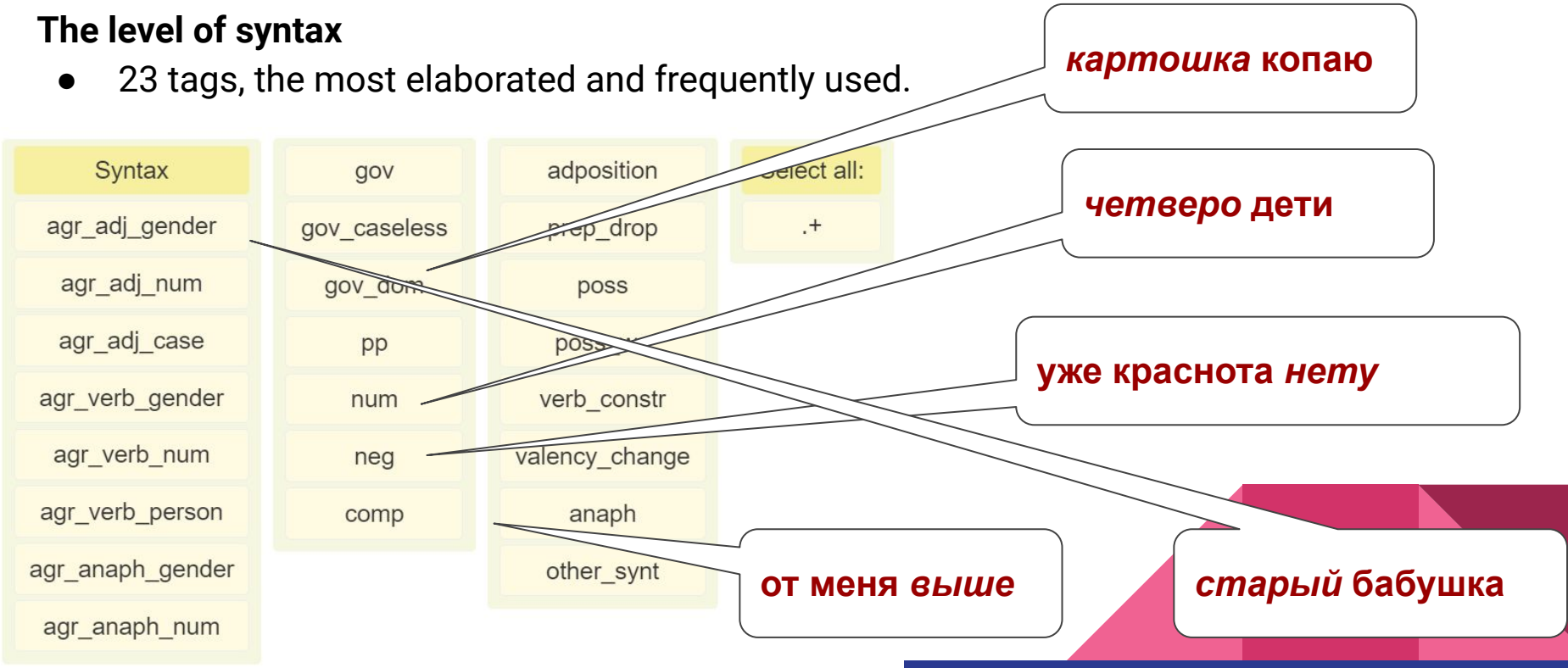
⇒ **67 tags in total**

- **“substandard”** (non-contact): 6 tags, dialectal, regional, register features (phonetics, stress, morphology, syntax, lexicon, other)

2. Corpus: Contact tags

The level of syntax

- 23 tags, the most elaborated and frequently used.



2. Corpus: speakers

- **Samoyedic speakers** (Taimyr peninsula: speakers of Nganasan, Enets, Nenets)
- **Tungusic speakers (**Amur region: speakers of Nanai and Ulcha)

- **Comparable sociolinguistic situations**:
  - endangered languages.
- **Typical speakers**:
  - acquired Russian at school age;
  - now, use both languages or (almost) only Russian;
  - the “best” narrators: 1-4 classes of school education.
- A continuum from **“near-pidgin”** to **“near-standard”** texts.
- Local pidgins (extinct):
  - for Samoyedic speakers - Taimyr Pidgin Russian (Govorka)
  - for Tungusic speakers - a variety of Siberian Pidgin Russian

3. Assessing a “post-pidgin continuum”

- We assessed this continuum, basing on our annotation of contact-induced features.
- Clustering speakers:
  - according to the distribution of different types of contact-induced features in their speech.
- Clustering contact-induced features:
  - according to their contribution in the inter-speaker variation.

3. Assessing a “post-pidgin continuum”

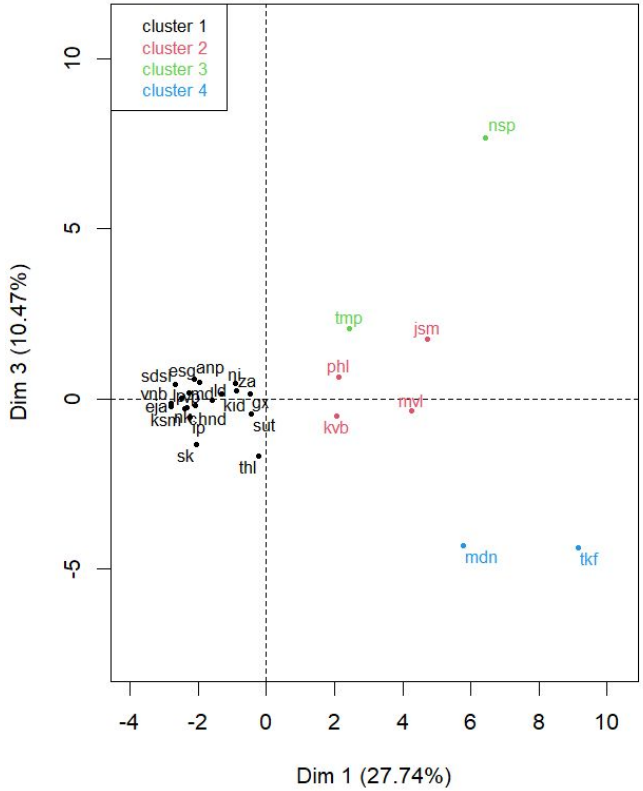
- **Method**: Principal Component Analysis (PCA) & Hierarchical clustering on principal components (HCPC)

  R-package Factoshiny (http://factominer.free.fr/graphs/factoshiny.html)
- **Data**:
  - 27 Tungusic speakers, 28 Samoyedic speakers
  - 38 features (morphology & syntax & polypredication) + total N of phonetic, prosodic, discourse, lexical, substandard features as supplementary parameters
  - for each feature & each speaker: N of uses per 1,000 tokens

A	D	E	F	G	H	P	Q	R	S
speaker	edu	lang	yob	tokens	total_contact	agr_adj_case	agr_adj_gend	agr_adj_num	agr_anaph_g
anp	secondary	enf	1952	469	19.1897655	0	2.132196	0	0
chnd	higher	nio	1946	6231	16.0487883	0	0.160488	0.320976	0.320976
eja	secondary	yrk	1970	717	0	0	0	0	0
esg	higher	enf	1962	1284	4.6728972	0	0	0	0
gx	primary	yrk	1950	2771	26.34428	0.360881	2.526164	0.360881	0.360881

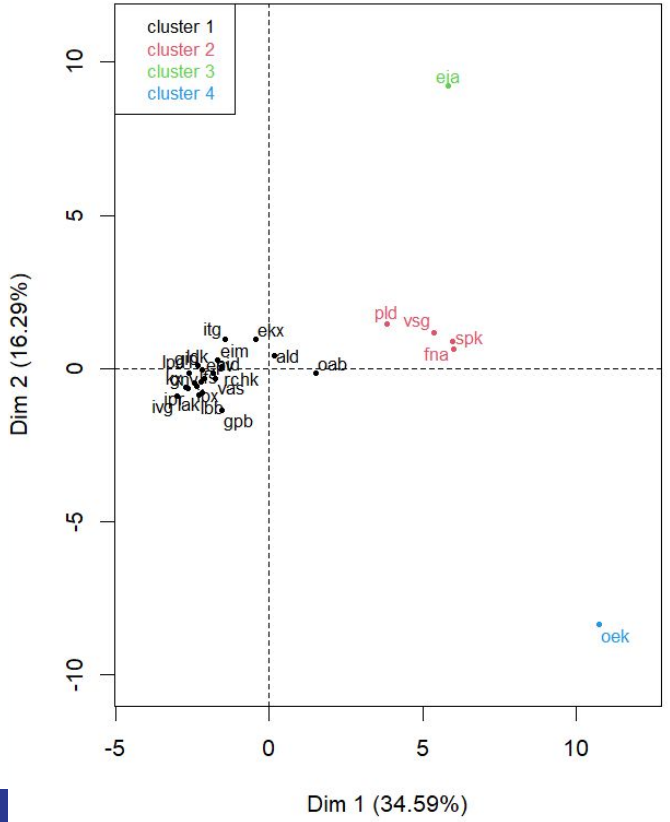
3.1. Clusterization of speakers: Samoyedic data

- **Cluster 1: “Near-standard”**
  - younger
  - negative correlation with: agr_adj_gender, gov_dom, agr_verb_gender, num, agr_verb_num, …
- **Cluster 2: “Intermediate”**
  - Nganasan
  - positive correlation with: categ_change, num, pp, agr_adj_**gender**, ...
- **Cluster 3: “Substandard”**
  - older, high % of dialectal and regional features
  - positive correlation with: subord_rel, infl_v, agr_anaph_gender, wo, ...
- **Cluster 4: “Near-pidgin: postpositions”**
  - older, known as Govorka mesolect speakers
  - positive correlation with: **adposition**, verb_constr, refl, gender, asp, poss



3.1. Clusterization of speakers: Tungusic data

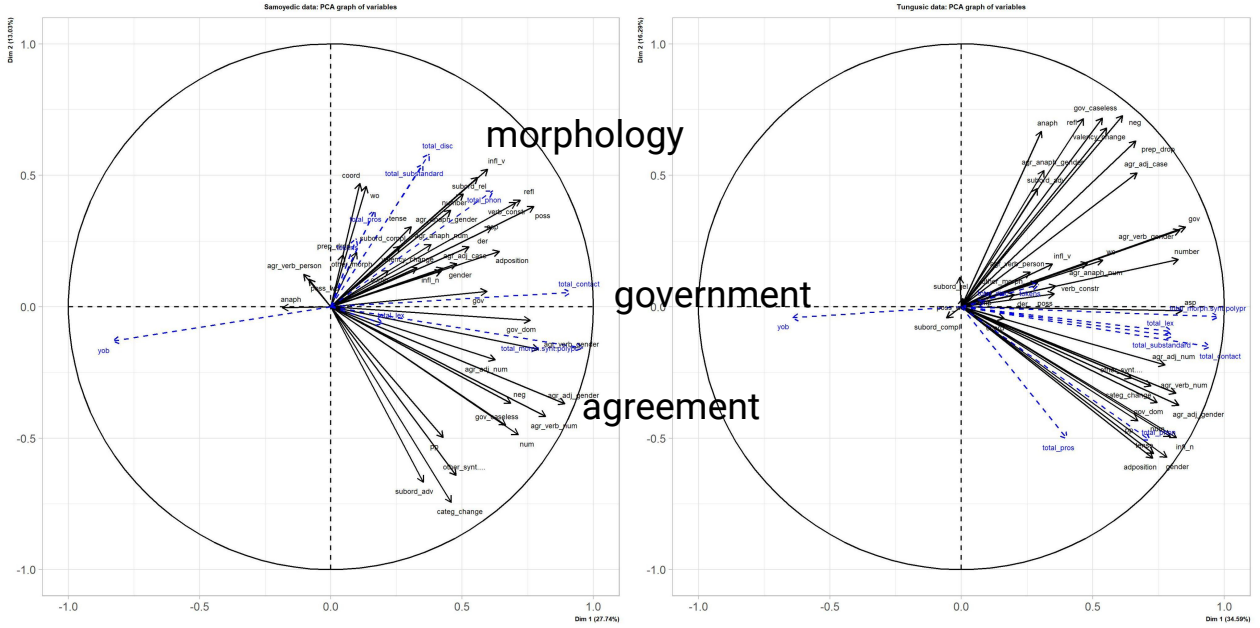
- **Cluster 1: “Near-standard”**
  - younger, higher or secondary education
  - negative correlation with: gov, agr_verb_gender, number, asp, agr_verb_num, neg, prep_drop, ...
- **Cluster 2: “Intermediate: number disagreement”**
  - positive correlation with: wo, agr_anaph_**num**, number, agr_adj_**num**, verb_constr, agr_verb_**num**, poss, ...
- **Cluster 3: “Near-pidgin: no inflection”**
  - positive correlation with: **agr_adj_case**, **gov_caseless**, neg, prep_drop, anaph, refl, valency_change
- **Cluster 4: “Nonstandard outlier”**
  - positive correlation with: gov_dom, gender, infl_n, adposition, tense, agr_adj_gender, num, ...



3.2. Clusterization of contact-induced features

- Different features behave more or less similarly.
- There are differences between the Samoyedic and Tungusic samples.

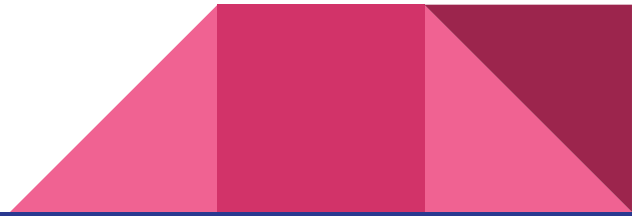
➔ Each vector represents a contact-induced feature.
➔ Blue vectors represent supplementary features (not included in the analysis).
➔ The length of a vector represents its contribution in inter-speaker variation.
➔ The adjacent vectors represent features that are distributed across speakers in a similar way.



4. Summary

- A corpus with the annotation of contact-influenced features.
- It is a helpful instrument for assessing a deviation from monolinguals' benchmarks for different speakers, using precise quantitative measures.

- Clusters of speakers more or less follow our intuition.
- A large cluster of near-standard speakers vs. different very small clusters of non-standard ones.
- Sets of features that determine each cluster of speakers do not form clear natural classes.



4. Further research

- The annotation needs further elaboration:
  - some inconsistencies and errors in the current version of the annotation;
  - some contact-induced features were annotated as “substandard” and vice versa
- We need more data: there are narrators with very few texts.
- However, the data of this type seem to be insightful for further research:
  - a more detailed analysis of contact-induced features should be conducted;
  - is there any natural logic in their clusterization?
  - e.g. is there any difference between the features associated with pattern-copying vs. those associated with incomplete acquisition?
  - most features behave similarly; which ones are outliers (e.g. do not contribute in inter-speaker variation)?