Theoretical and experimental approaches to dialectal variation and contact-induced change: a case study of Tundra Nenets

Nikolett Mus

Research Institute for Linguistics, Hungarian Academy of Sciences

Linguistic diversity, minority languages and digital research infrastructures
Hamburg, 20–21 September 2018

Background

Typology of Negation in Ob-Ugric and Samoyedic Languages

- 2008–2011, University of Vienna
- typological framework of negation
- description of a certain – previously undescribed – linguistic phenomenon in the examined languages
- online database of the constructions
- annotated texts

Background (cont.)

Languages Under Influence. Uralic syntax changing in an asymmetrical contact situation

- 2016–2017, Research Institute of Linguistics HAS
- descriptions and analyses of potential contact induced syntactic changes, e.g. a change from SOV to SVO
- database

Background (cont.)

Nominal Structures in Uralic Languages

- 2017–2021, Research Institute of Linguistics HAS
- a description of noun phrases and nominal predication in Uralic languages
- an online, open access database of the construction types

Background (cont.)

Workshop series on Uralic prosody

- University of Tartu, Estonia; Research Institute of Linguistics HAS
- empirical research on prominence, rhythm, and intonation
- typological questions of Uralic prosody
- tools for prosodic stylization or annotation
- corpora for Uralic languages

Background (cont.)

Theoretical and experimental approaches to dialectal variation and contact-induced change: a case study of Tundra Nenets

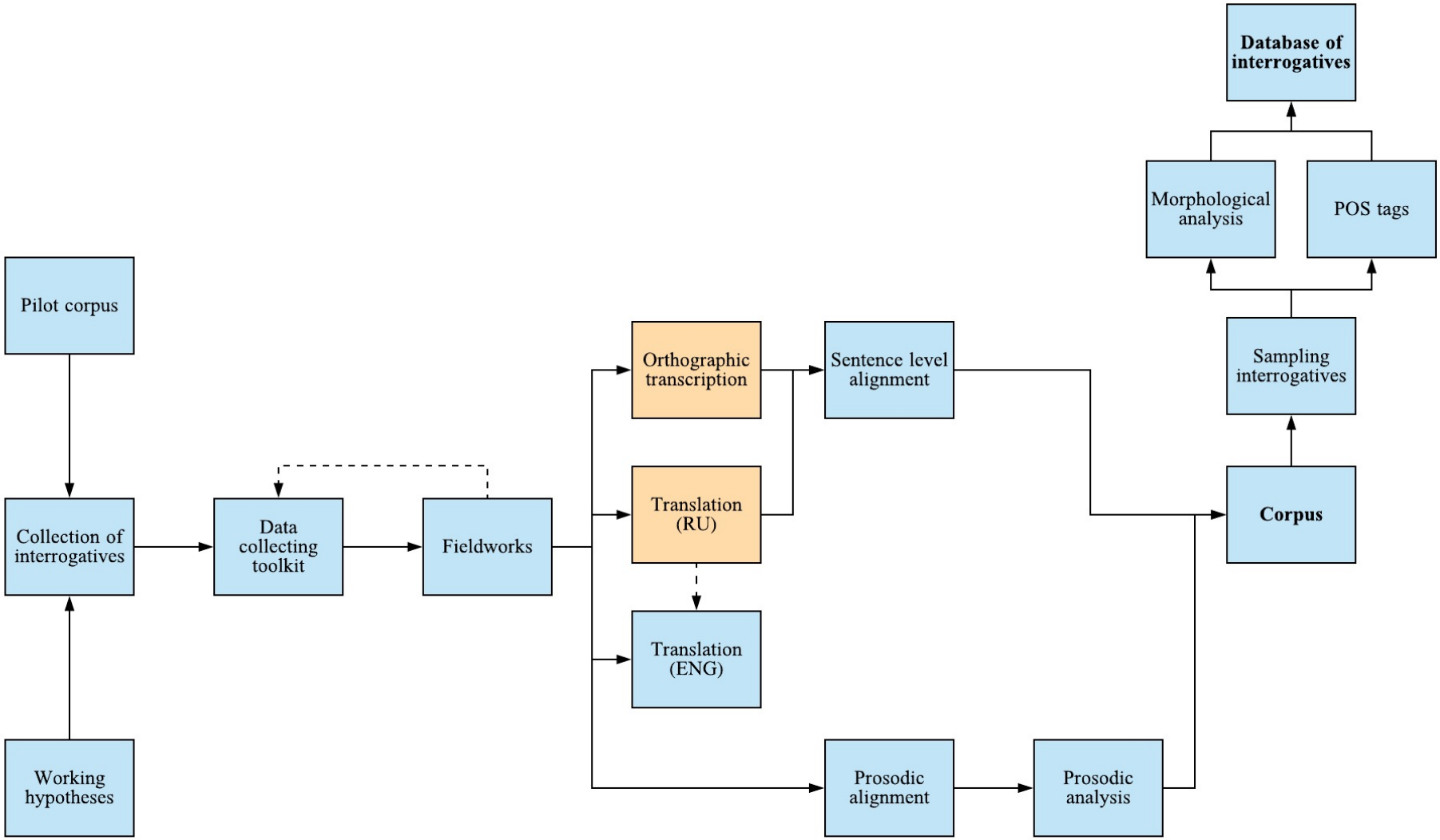
- Research network and research infrastructure
  - 2018–2022
  - Research Institute of Linguistics HAS
  - National Research, Development and Innovation Office
  - participants
    - Katalin, Mády (RIL HAS)
    - Réka, Metzger
    - Nikolett, Mus (PI; RIL HAS)
    - Uwe, Reichel (LMU Munich)
    - Péter, Rebrus (RIL HAS)

Background (cont.)

Aims and expected results

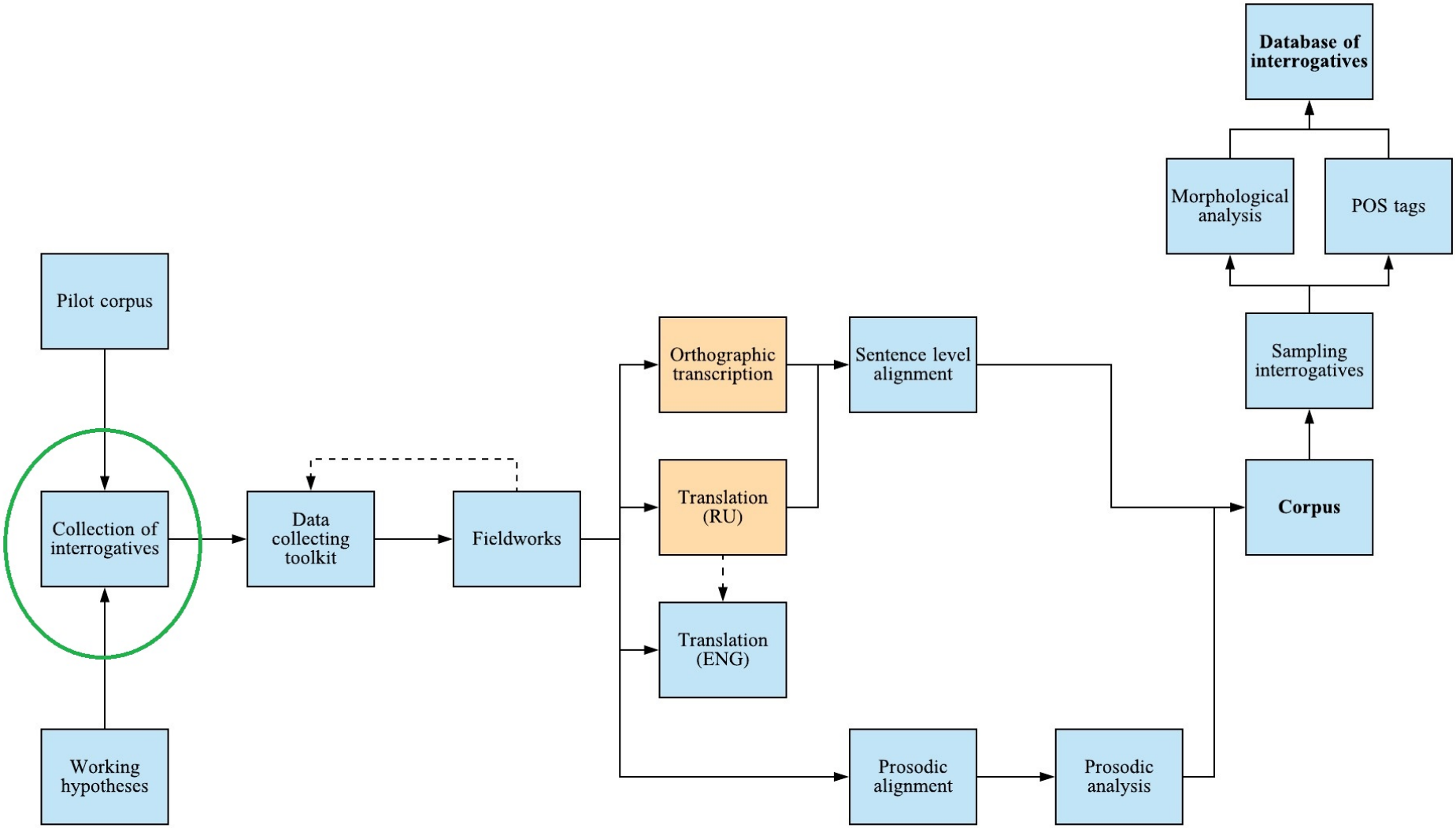
- a comprehensive description of interrogative clauses in Tundra Nenets
- a comparative analysis of interrogatives in Tundra Nenets dialects
- a typological questionnaire
- a language-independent data collection toolkit
- an online database of Tundra Nenets interrogative structures
- a parallel (and comparable) corpus of Tundra Nenets spoken data (collected during our fieldworks)

Workflow


Database of interrogatives
Morphological analysis
POS tags
Pilot corpus
Sampling interrogatives
Orthographic transcription
Sentence level alignment
Collection of interrogatives
Data collecting toolkit
Fieldworks
Translation (RU)
Corpus
Translation (ENG)
Working hypotheses
Prosodic alignment
Prosodic analysis

Outline

Collection of interrogatives



Theoretical starting point

- The Uralic languages spoken in the Russian Federation are under a strong Russian influence.
- This contact often leads to the restructuring of the traditional SOV-type Uralic languages to the Russian SVO-type, i.e. from head-final to head-initial types.
- There are social factors/constraints that may slow down the contact induced changes.

Theoretical starting point (cont.)

- We will focus on two speech communities
  - the one spoken on the Yamal Peninsula by traditional reindeer herders
  - the other spoken in Dudinka, by people who settled down in the city and live a (more or less) urban life

H The two dialectal variations of Tundra Nenets are influenced by the Russian language to a different extent:

  - the dialect spoken on the Yamal Peninsula is supposed to represent a more conservative head-final structure
  - signs of breaking up the SOV structure in the variation spoken in Dudinka is assumed.

Subject of the research

- Our general hypothesis will be tested on a specific type of clauses, i.e. on interrogatives.

$\Rightarrow$ interrogatives are usually presumed to be universal (Sadock & Zwicky 1985; Huddleston 1994; König & Siemund 2007; Velupillai 2012)

$\Rightarrow$ interrogative marking strategies show correlations with the basic word order of languages (Greenberg 1966)

Empirical starting point

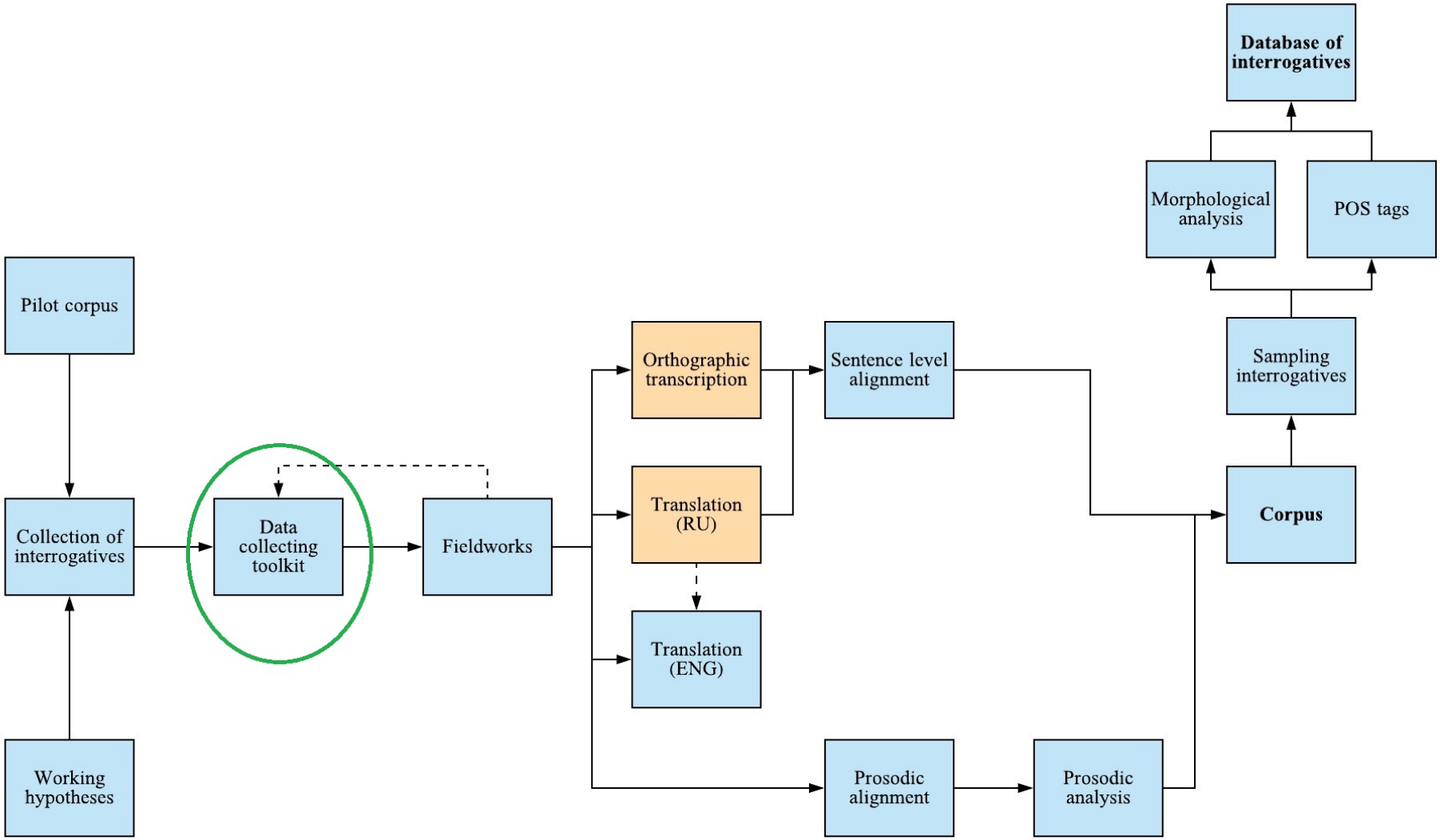
- building a pilot corpus (raw data)
  - selection of written and spoken texts
  - published and/or electronically accessible sources e.g. newspapers, TV podcasts, corpora, database
  - Yamal and Taymyr dialects

  $\rightarrow$ OCR, web scraping

- literature review

$\Rightarrow$ Collection of interrogatives: raw data, texts in `.txt`

Data collecting toolkit



Methodological considerations

- aims
  - cross-linguistic comparability of the data
  - language independent, but! not culture independent toolkit
- methods
  - combining different data gathering techniques
  - examining the same linguistic phenomena with various procedures

Methods of documentational linguistics

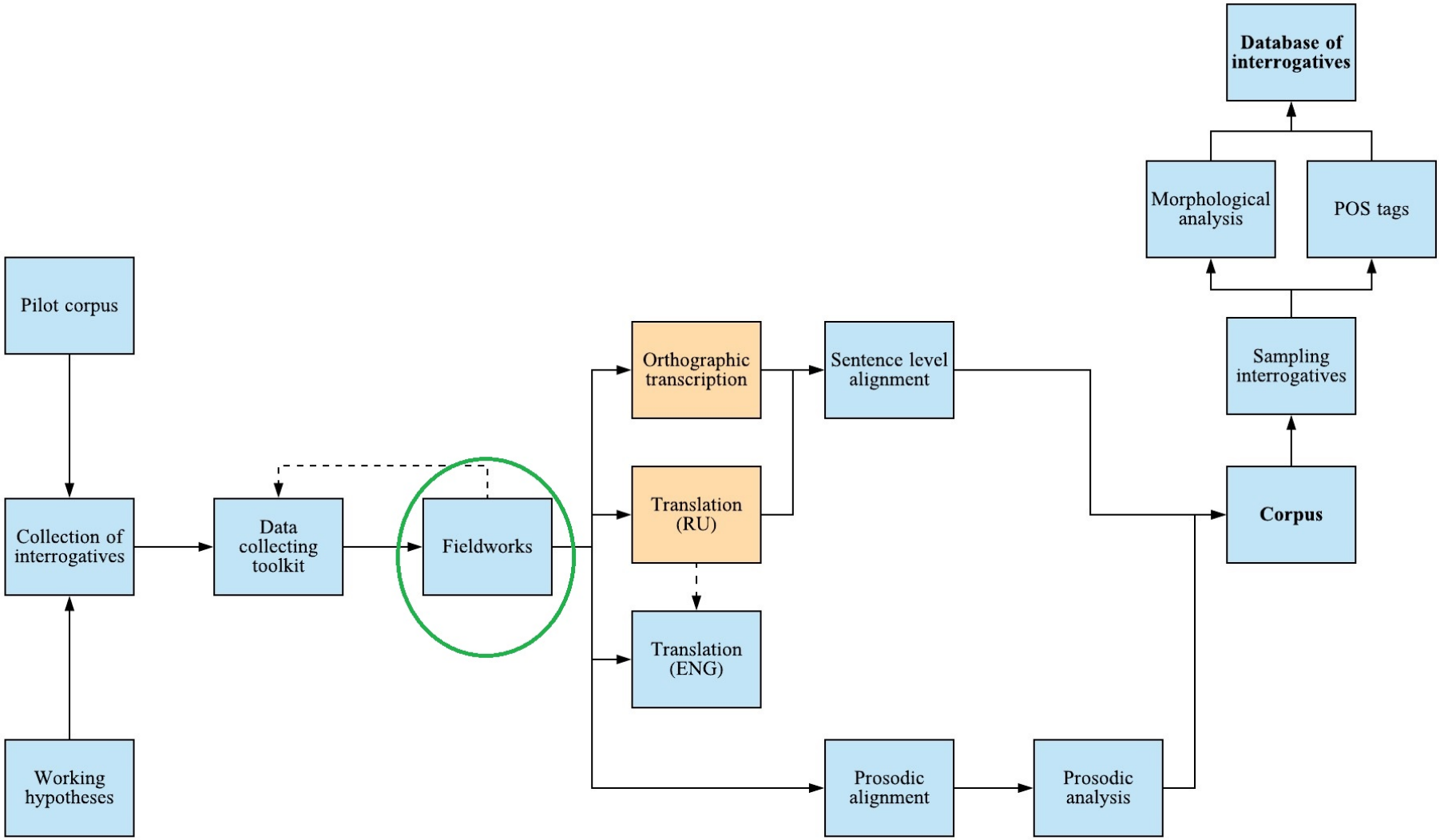
- focus on the speaker's performance
- collecting reliable, representative and natural data
- documenting metadata
- techniques
  - observed events
  - staged events

Methods of experimental linguistics

- focus (also) on the speaker's competence
- measuring grammaticality, preference and reaction times
- systematic manipulation of variables
- techniques
  - sentence-picture matching
  - questionnaires

$\Rightarrow$ Data collecting toolkit

Fieldworks

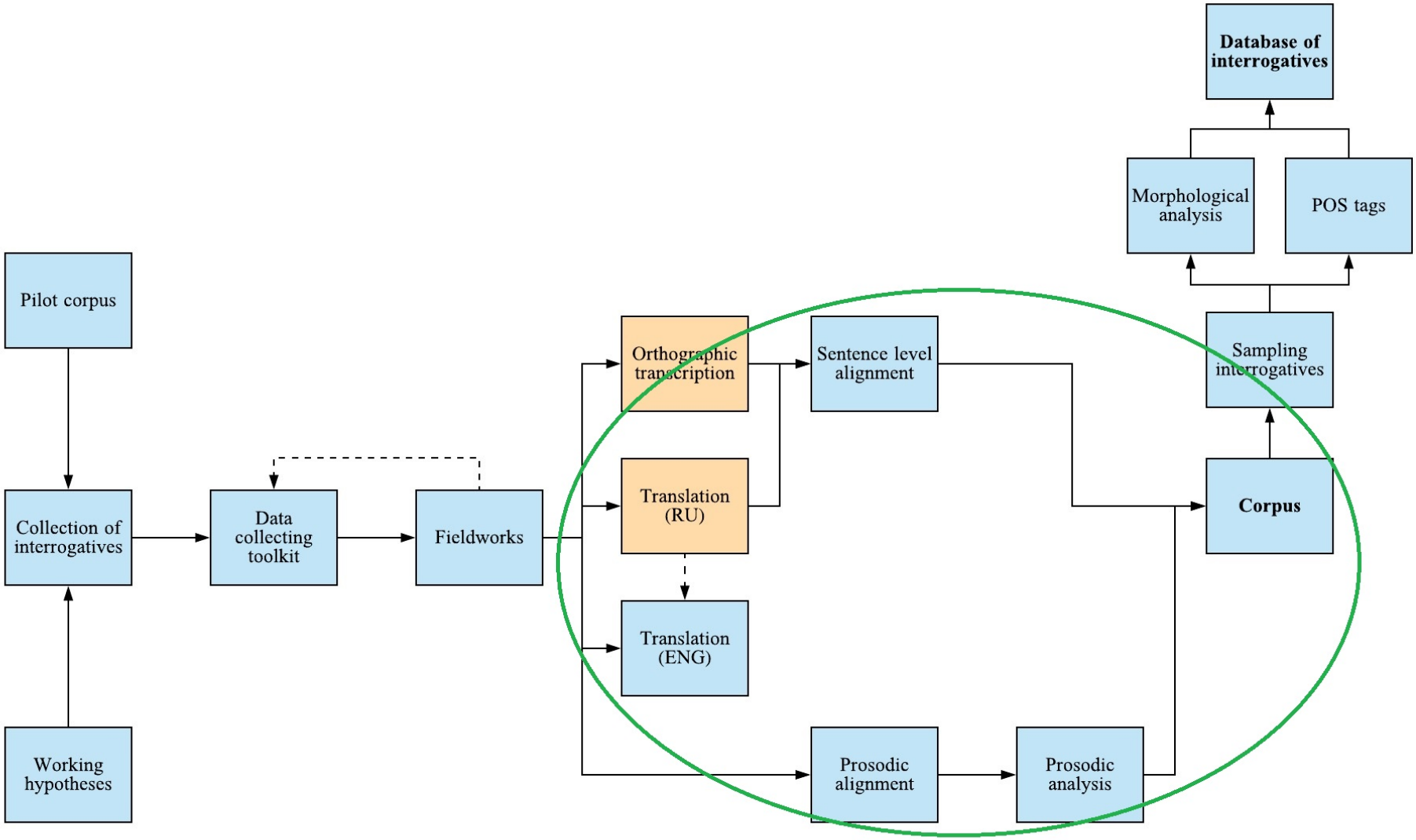

Database of interrogatives
Morphological analysis
POS tags
Pilot corpus
Orthographic transcription
Sentence level alignment
Sampling interrogatives
Collection of interrogatives
Data collecting toolkit
Fieldworks
Translation (RU)
Corpus
Translation (ENG)
Working hypotheses
Prosodic alignment
Prosodic analysis

Fieldworks

- Fieldwork 1
  - in Moscow with a native speaker
  - the main goal is to test the toolkit
- 2020 expedition to the Yamal Peninsula
- 2021 expedition to Dudinka

Fieldworks (cont.)

Corpus


Database of interrogatives
Morphological analysis
POS tags
Pilot corpus
Orthographic transcription
Sentence level alignment
Sampling interrogatives
Collection of interrogatives
Data collecting toolkit
Fieldworks
Translation (RU)
Corpus
Translation (ENG)
Working hypotheses
Prosodic alignment
Prosodic analysis

Processing levels

- Spoken data
  - Recordings → Automatic prosodic alignment → Automatic prosodic analysis (Uwe Reichel)
- Written data
  - Recordings → Manual Cyrillic transcription → Automatic Latin transcription
  - Recordings → Manual Russian and English translations
  - Automatic sentence-level alignment of YRK – RUS – ENG

Problems with the Cyrillic transcription

- Same phoneme – different characters
  - In Tundra Nenets, two glottal stops are usually differentiated: a nasalizable /h/ and a non-nasalizable /q/.
  - They are marked with two different characters in the Cyrillic transcriptions too: nasalizable /’/ and non-nasalizable /”/, e.g. я’ ‘soot’ and я” ‘piece of hair’.
  - Although the two ‘phonemes’ are indicated differently in writing, their marking strategy is not systematic.
  - They are pronounced in the same way and do not differ in any acoustic properties (Staroverov 2006).

Problems with the Cyrillic transcription (cont.)

- Different phonemes – same character
  - The length of low vowels:
    - short /a/ – over-short /ə/ – reduced /°/ (Nikolaeva 2014)
    - short /a/ – over-short /ă/ (Staroverov 2006)
  - this difference shows up only in the first syllable: xada ‘grandmother’ xăda ‘nail’
  - in (most) Cyrillic transcriptions the length of vowels is not differentiated хада ‘grandmother’ and ‘nail’

Problems with Latin transcription

RUS:	ты”	яля	август’	ŋарка	яля.
FUT:	tī?	jāĺe	avgust?	ŋārka	jāĺe.
Nikolaeva:	tíq	yalya	avgust°h	ŋarka	yalya.
IPA:	ti:?	jalʲa	avgust?	ŋarka	jalʲa. → transliteration

Proposed strategy

- rough Cyrillic transcription by the speakers
- acoustic-phonetic examination and analysis of the phonemes
- clarification of the phoneme system
- drawing up transcription rules
- automatic Latin transcription

\+ correcting the rough transcription by using the conventions of Conversation Analysis

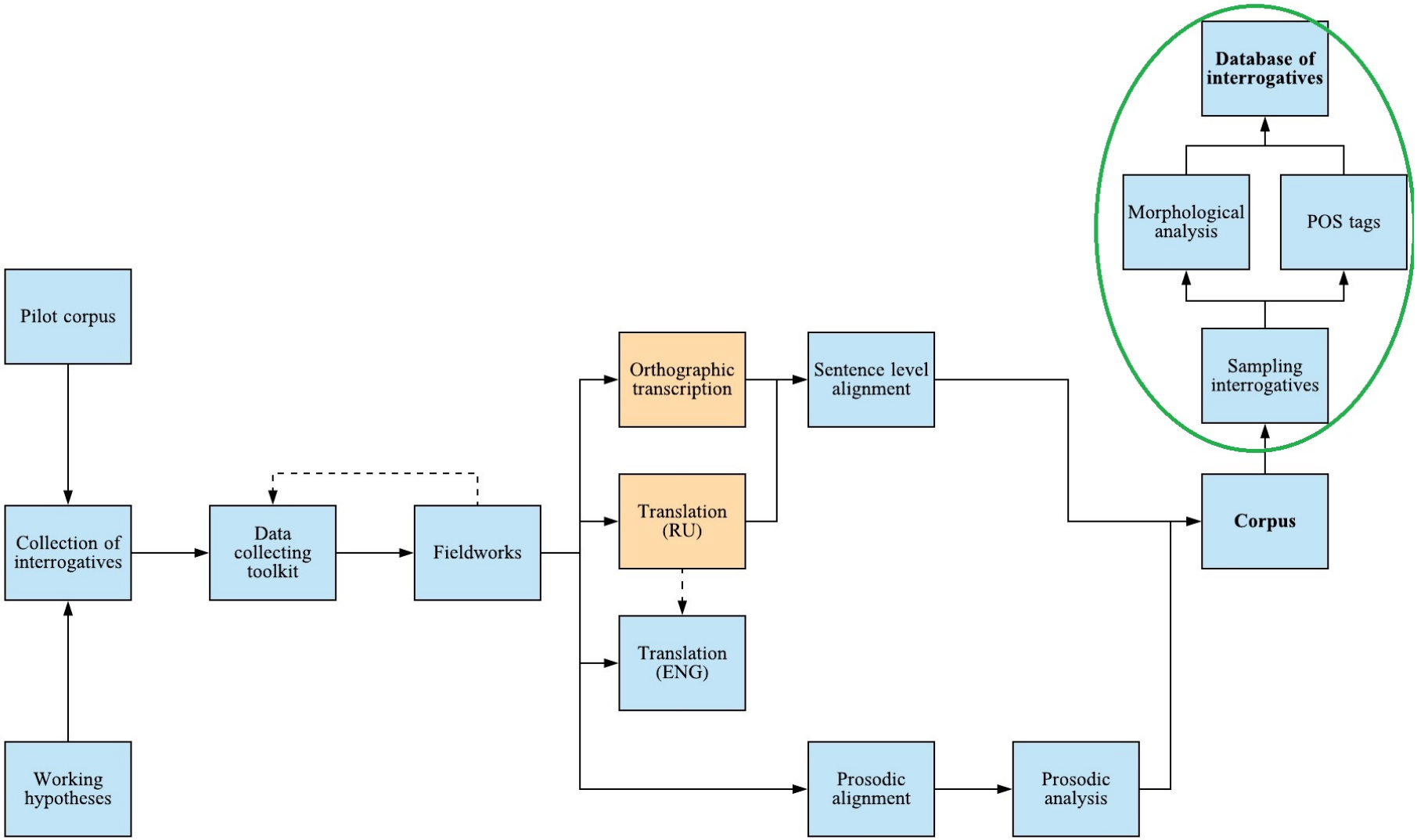
Sentence level alignment

- following the process of UraLUID
  - conversion of the `.txt` files (Cyrillic transcritpion and the ENG/Ru translations) into `.tsv`
  - uploading the texts to (No)SketchEngine $\rightarrow$ EN/RU/YRK$_{\text{Y}}$ and EN/RU/YRK$_{\text{T}}$ parallel corpus
  - creating ELAN and Praat annotation files automatically by Pympi module of python3

Metadata

- usual information, e.g. age, dialect, etc.
- photo-documentation
- managing metadata
  - (C)IMDI maker

Database of interrogatives

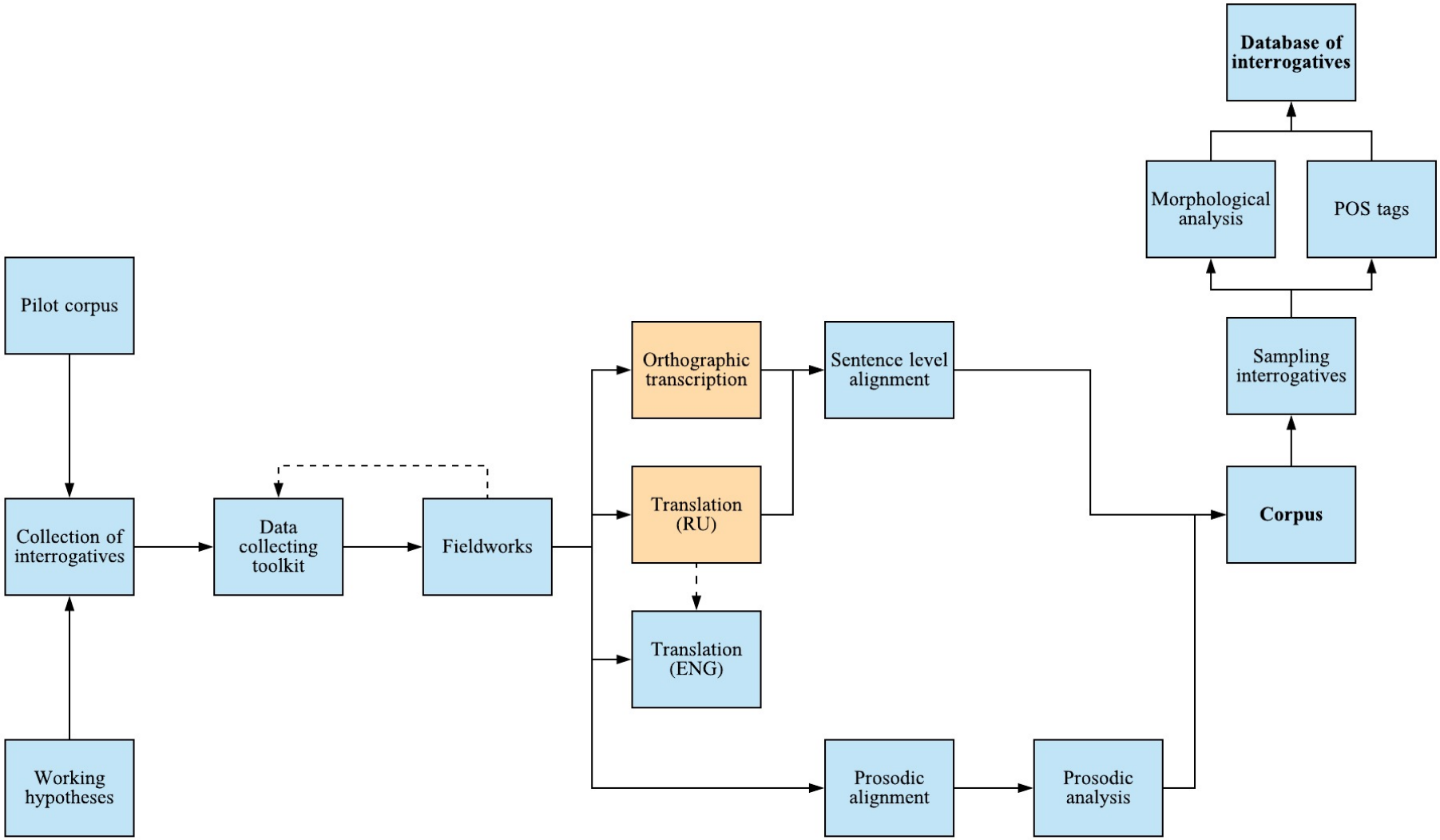


A note on morphological analysis

- morphological analysis
- POS tags $\Rightarrow$ manually (?)

- The Giellatekno Morphological Analyser
  - pilot-study: 1522 / 3462 tokens (44%)

Summary


Database of interrogatives
Morphological analysis
POS tags
Pilot corpus
Sampling interrogatives
Orthographic transcription
Sentence level alignment
Collection of interrogatives
Data collecting toolkit
Fieldworks
Translation (RU)
Corpus
Translation (ENG)
Working hypotheses
Prosodic alignment
Prosodic analysis

Summary (cont.)

- comparable corpus of both dialectal variations
  - Q1 Can we find systematic similarities/differences in the marking strategies of the interrogatives in the two dialectal variations of Tundra Nenets?
  - Q2 Are the systematic differences possible consequences of a contact-induced change?
  - Q3 Do the different circumstances of the dialectal variations result in different patterns of contact-induced changes?
  - Q4 Can we predict a path of a more general typological contact-induced change on the basis of the Tundra Nenets data and results?

Acknowledgements

Thank you for your attention!

The author wish to thank Veronika Hegedűs for her contribution.
The support of the research projects "Theoretical and experimental approaches to dialectal variation and contact-induced change: a case study of Tundra Nenets" (NKFI 129235) and "Languages under the Influence. Uralic Syntax Changing in an Asymmetrical Contact Situation" (OTKA 118079) is gratefully acknowledged.