Grammar & Corpora 2022,
University of Ghent
30.06.2022

# Multilingual SpeechReporting database: tools, methods and techniques

Katya Aplonova[1], Tatiana Nikitina[1], Timofey Arkhangelskiy[2], Abbie Hantgan-Sonko[3], Izabela Jordanoska[1], Elena Sokur[4], Lacina Silué[5], Rebecca Paterson[6]
[1]LACITO, CNRS; [2]University of Hamburg; [3]Language Conservancy Group; [4]NRU Higher School of Economics [5]INALCO , [6] SIL International & Western Oregon University



# Introduction

# Why a database of speech reporting?

- Speech reporting:
  *Mary said that she would come*
- Universal and claimed to distinguish human language from other kinds of communication (Coulmas 1985)
- Reported speech is a very special syntactic domain with a characteristic set of properties (Spronck & Nikitina 2019)
- Corpus research of reported speech based on natural data are extremely rare
- There is no way to extract the data about reported speech automatically, special annotations are needed

# Why not simply search for “say”?

Udihe, Tungusic

(1) Belie **diga-si-mi** **meteu-e-ni**
beauty eat-IM-GER finish-PST-3SG

Mam'a, ededе, jeu-xi ŋeneh-e-ti?
grandmother oh! what-DIR go-PST-3PL
‘The girl **finished eating**: "Grandmother, ooh, where did they go?"’

Macedonian, Slavic

(2) Kade e Role? Na ora-nje jet.
where be.PRS.3SG R. on plow-NLMZ be.PRS.3SG
‘“Where is Role?” “He is out plowing.”’

# The SpeechReporting database

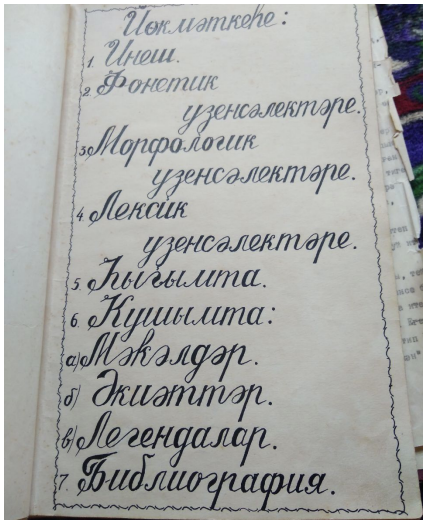
- The SpeechReporting database (Nikitina et al. 2021): collection of traditional folk stories in 10 languages in West Africa and Eurasia
- All of the texts are:
  - transcribed
  - semi-automatically glossed
  - translated
  - annotated for instances of reported speech

# SpeechReporting database: data

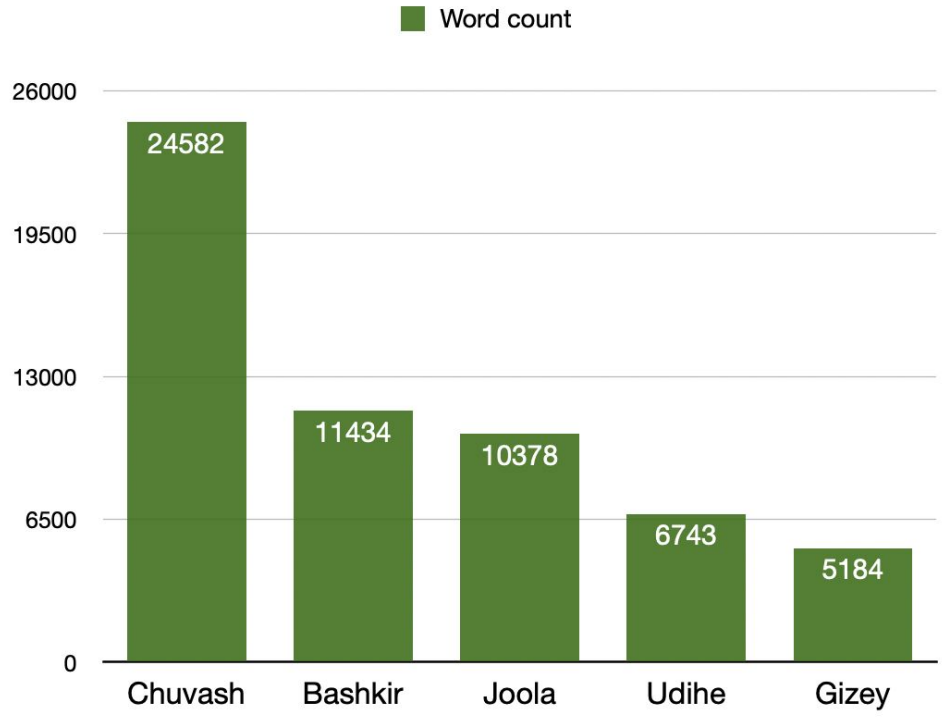
- Data is very diverse: multimedia files, archived transcriptions, data from the field, written/unwritten languages, recordings of professional storytellers as well as of regular people, one or multiple participants…

# SpeechReporting database: data

- Data is very diverse: multimedia files, archived transcriptions, data from the field, written/unwritten languages, recordings of professional storytellers as well as of regular people, one or multiple participants…

# SpeechReporting database: corpora available online



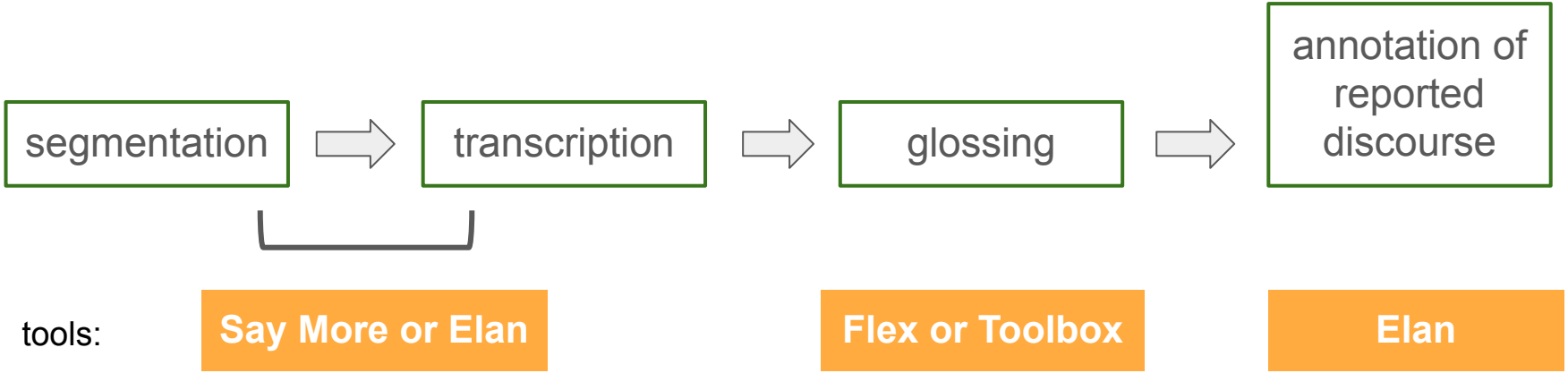

- database is regularly updated
- languages in progress: Kafire, Macedonian, Ut-Ma'in, Guro, Wan, Mwan

# Today’s talk

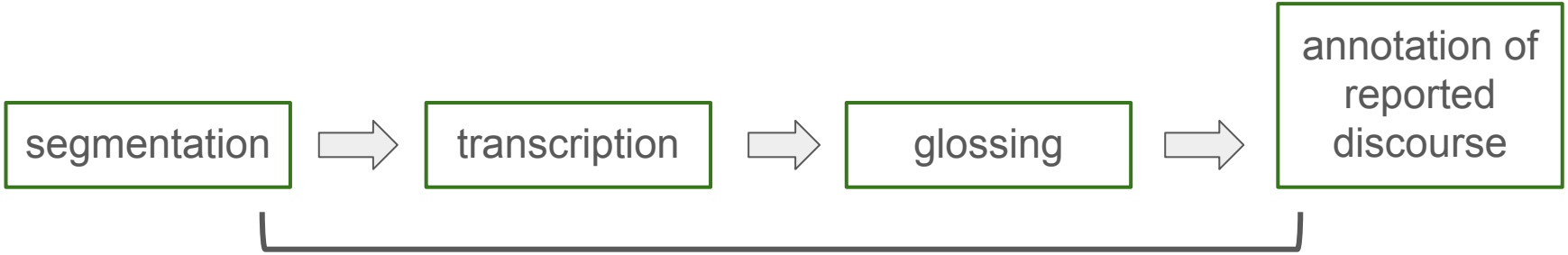
- **Our workflow** and the **tools** that we used to construct the corpora of the different languages comparable and suitable for cross-linguistic research.
- **Case study** that illustrates how our data can be used for **quantitative cross-linguistic comparison**: Speech-introducing function of interjections

# Workflow and Tools

# Workflow 1


segmentation
transcription
glossing
annotation of reported discourse
tools:
Say More or Elan
Flex or Toolbox
Elan

# Workflow 2




tools:

**Elan CorpA (Chanard 2015)**

# SpeechReporting template (Nikitina et al. 2019): annotation of reported discourse

- Elements of reported speech construction (RP): Discourse Report, Quotative…
- Syntactic type of the Reported Speech construction (QT)
- Semantic type of the quote (TYP): Statement, Question, Command…
- Reference to participants (PAR): Reported speaker/listener, Current speaker/listener

Bashkir, Turkic

| | | | | | |
|---|---|---|---|---|---|
| (3) | *min* | *šulaj* | *ti-p* | *ujla-j-əm* | *ti-gän* |
| | 1SG | so | say-CVB | think-IPFV-1SG | say-PC.PST |

‘“I think so” – he said’.

# SpeechReporting template (Nikitina et al. 2019): annotation of reported discourse

- Elements of reported speech construction (RP): **Discourse Report**, **Quotative**…
- Syntactic type of the Reported Speech construction (QT)
- Semantic type of the quote (TYP): Statement, Question, Command…
- Reference to participants (PAR): Reported speaker/listener, Current speaker/listener

(3)

| ***min*** | ***šulaj*** | ***ti-p*** | ***ujla-j-əm*** | ***ti-gän*** |
|---|---|---|---|---|
| 1SG | so | say-CVB | think-IPFV-1SG | say-PC.PST |

‘“I think so” – he said’.

# SpeechReporting template (Nikitina et al. 2019): annotation of reported discourse

- Elements of reported speech construction (RP): **Discourse Report**, **Quotative**…
- Syntactic type of the Reported Speech construction (QT): Quotative + Discourse Report
- Semantic type of the quote (TYP): Statement, Question, Command…
- Reference to participants (PAR): Reported speaker/listener, Current speaker/listener

(3)

| ***min*** | ***šulaj*** | ***ti-p*** | ***ujla-j-əm*** | ***ti-gän*** |
|---|---|---|---|---|
| 1SG | so | say-CVB | think-IPFV-1SG | say-PC.PST |

‘“I think so” – he said’.

# SpeechReporting template (Nikitina et al. 2019): annotation of reported discourse

- Elements of reported speech construction (RP): Discourse Report, Quotative…
- Syntactic type of the Reported Speech construction (QT)
- Semantic type of the quote (TYP): **Statement**, Question, Command…
- Reference to participants (PAR): Reported speaker/listener, Current speaker/listener

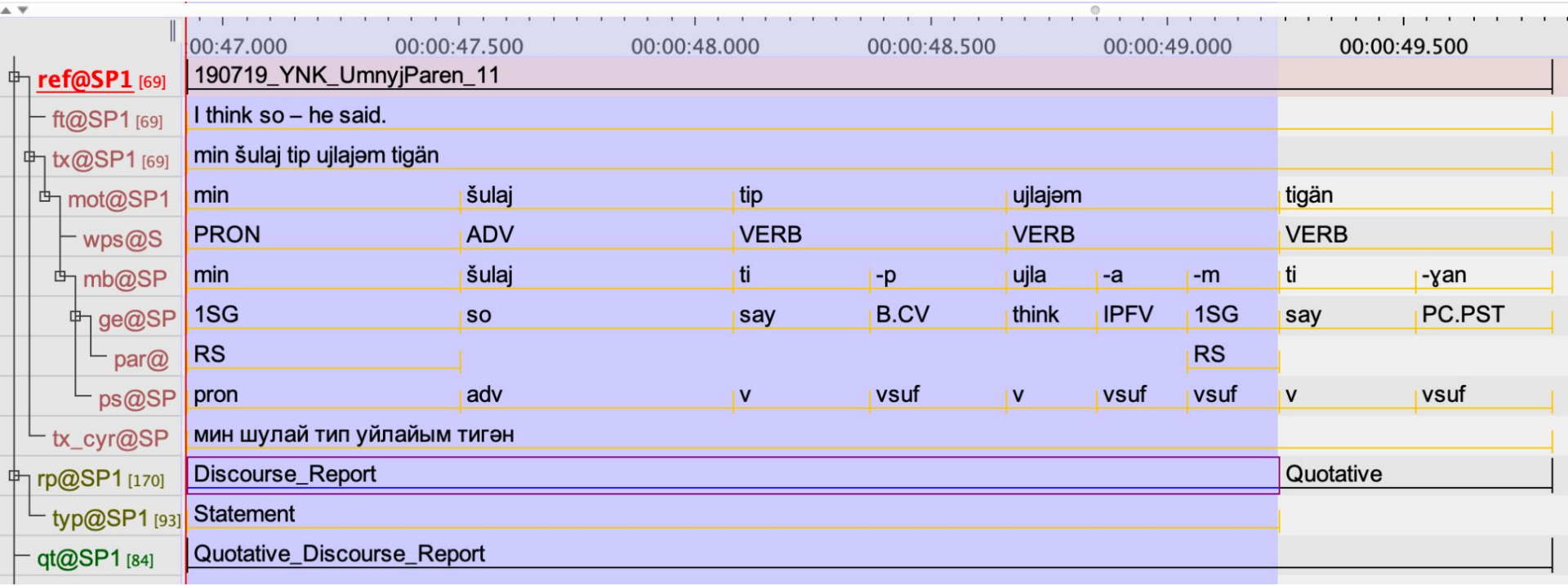
(3) ***min*** ***šulaj*** ***ti-p*** ***ujla-j-əm*** *ti-gän*
1SG so say-CVB think-IPFV-1SG say-PC.PST
‘“I think so” – he said’.

# SpeechReporting template (Nikitina et al. 2019): annotation of reported discourse

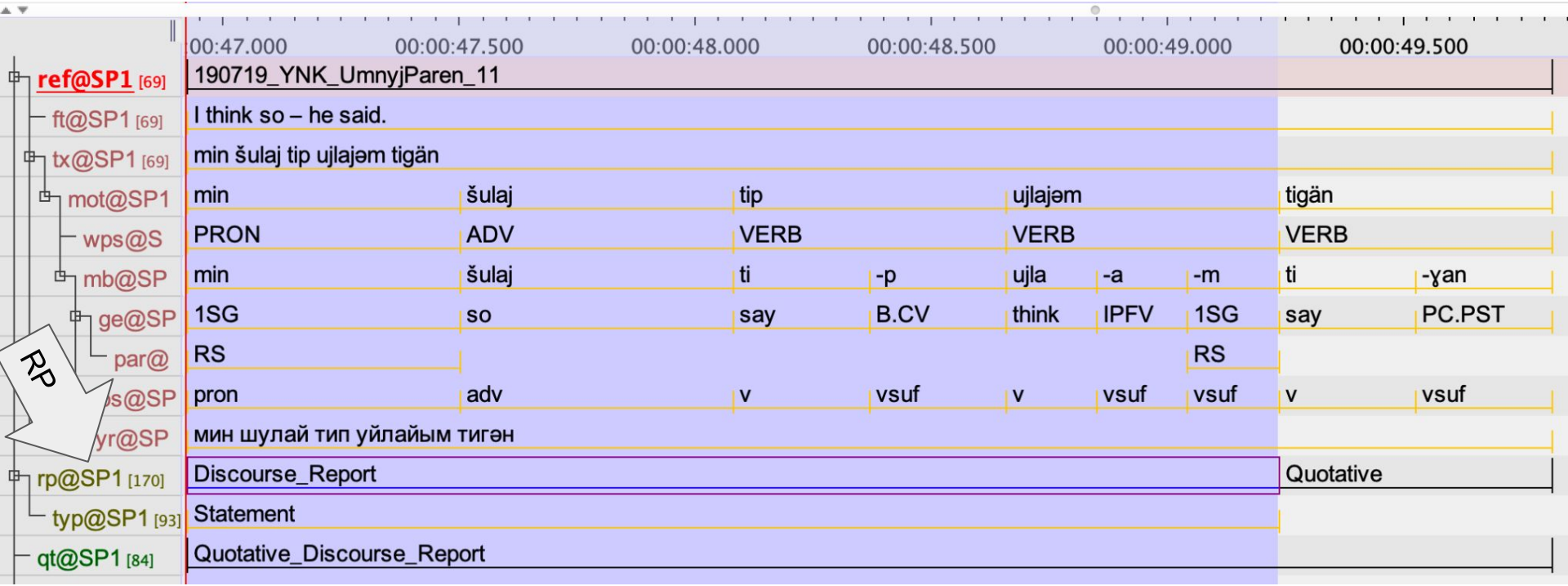
- Elements of reported speech construction (RP): Discourse Report, Quotative…
- Syntactic type of the Reported Speech construction (QT)
- Semantic type of the quote (TYP): Statement, Question, Command…
- Reference to participants (PAR): **Reported speaker**/listener, Current speaker/listener

(3) ***min*** *šulaj* *ti-p* *ujla-j-**əm*** *ti-gän*
1SG so say-CVB think-IPFV-1SG say-PC.PST
‘“I think so” – he said’.

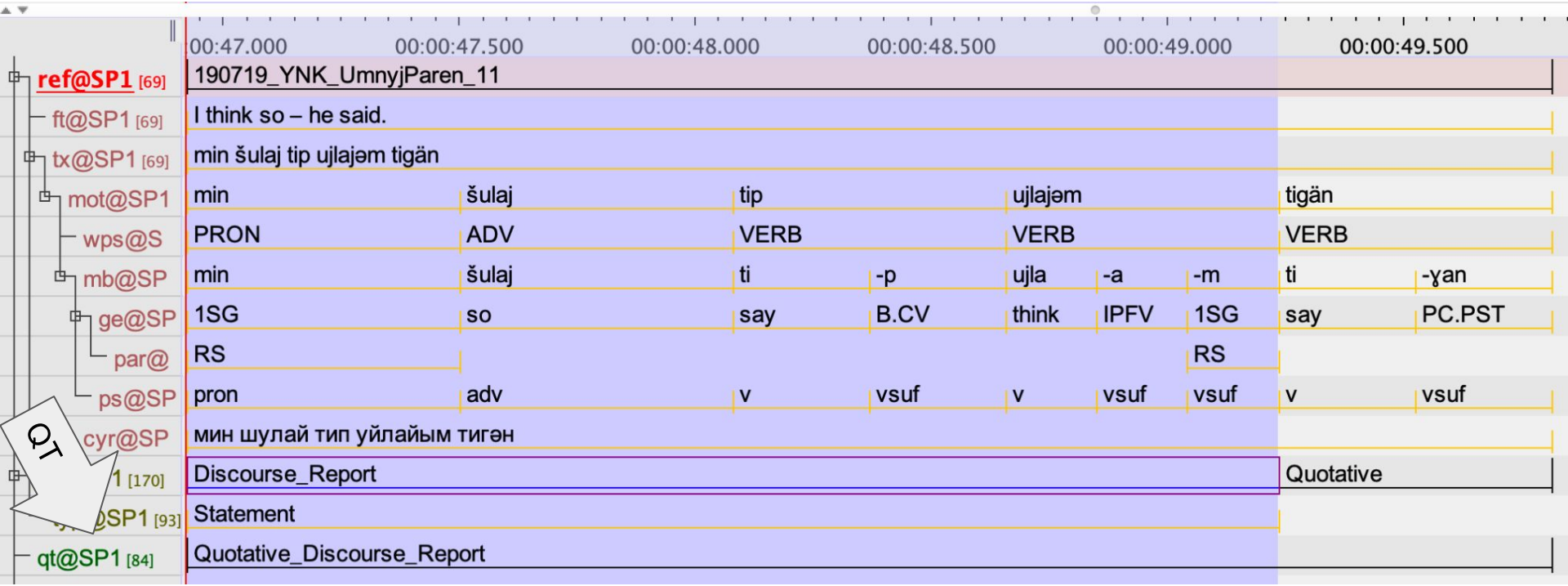
# SpeechReporting template: annotation of reported discourse



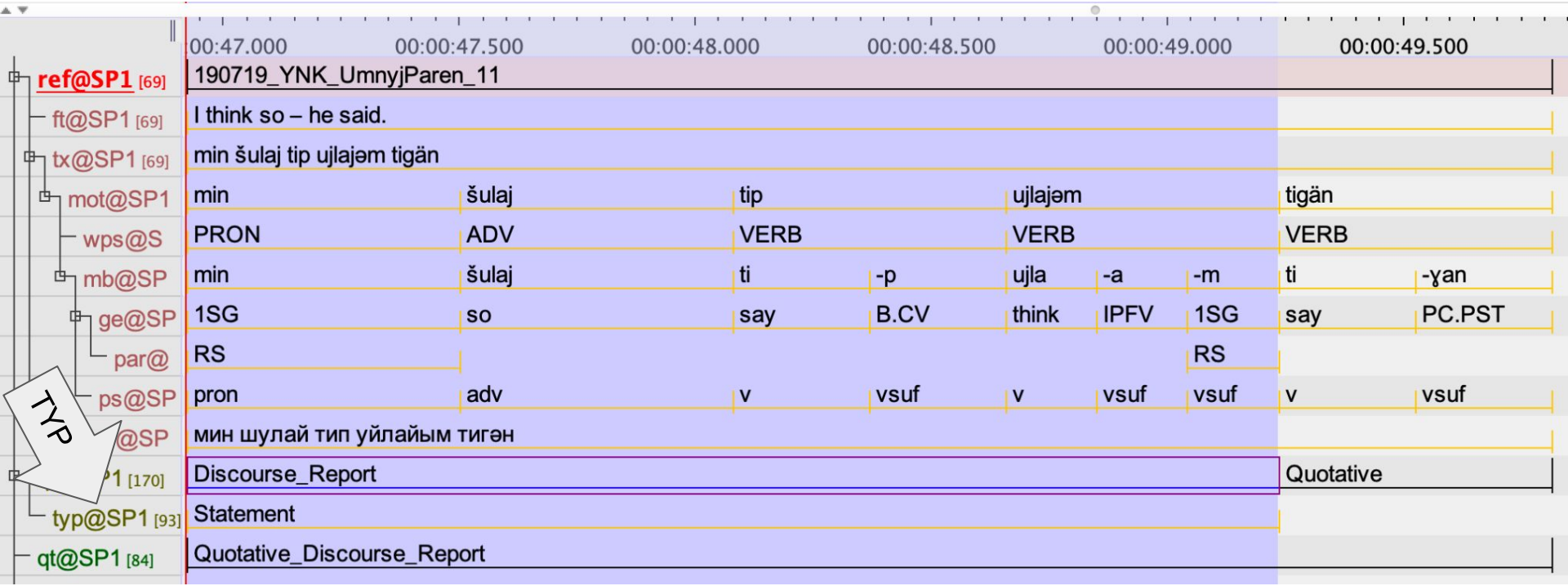
# SpeechReporting template: annotation of reported discourse



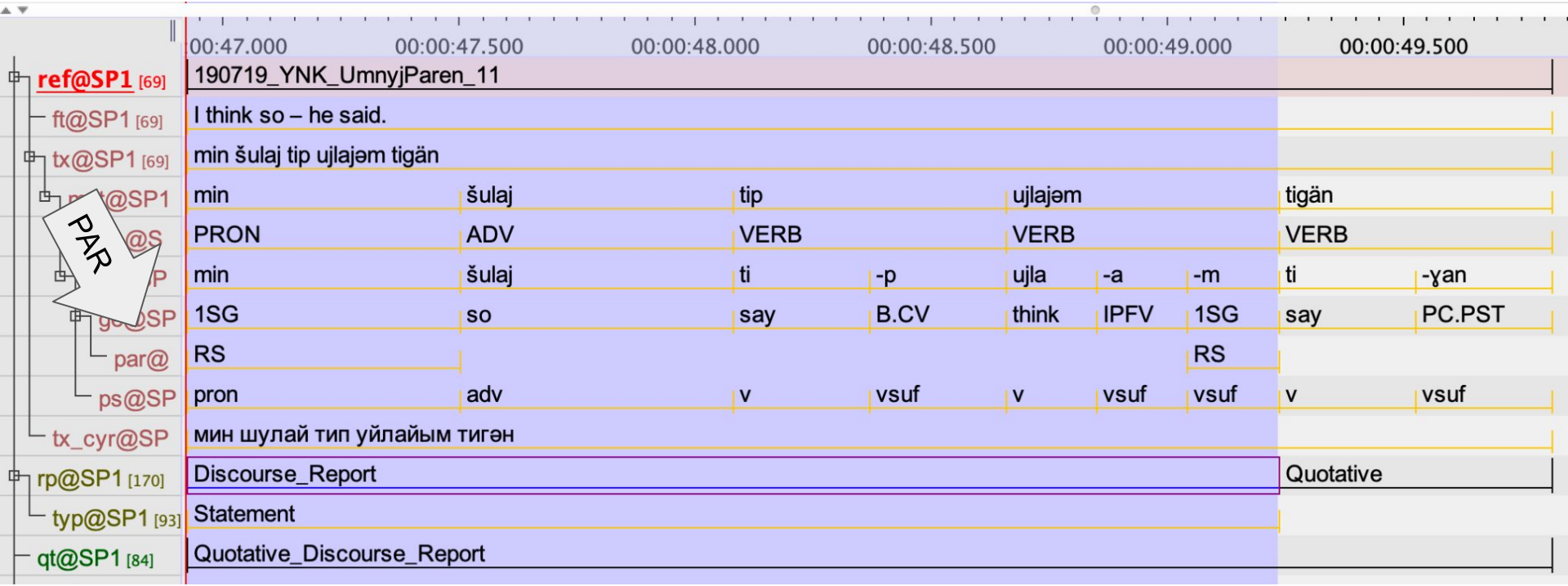
# SpeechReporting template: annotation of reported discourse



# SpeechReporting template: annotation of reported discourse



# SpeechReporting template: annotation of reported discourse



# Tsakorpus: web search interface for multimedia corpora

Multilingual corpus of the Discourse Reporting project EN | RU | ?

Back to search


Search result: 115 occurrences, 110 sentences

| min | šulaj | tip | ujlajəm | tigän |
|---|---|---|---|---|
| min | šulaj | ti | ujla | ti |
| min | šulaj | ti-p | ujla-a-m | ti-ɣan |
| 1SG | so | sayB.CV | thinkIPFV1SG | sayPC.PST |
| pron | adv | v, vsuf | v, vsuf | v, vsuf |
| RS | so | say | ∅-∅-RS | say |
| 1SG | | | think | |

I think so – he said.

190719_YNK_UmnyjParen_11

qt: ['Quotative_Discourse_Report']
rp: ['Discourse_Report', 'Quotative']
speaker: SP1
typ: ['Statement']

Only 20 documents

# Tsakorpus: Demo

**Visit the corpus yourselves:**
*http://discoursereporting.huma-num.fr/index.html*

# Speech introducing function of interjections

# Interjections and reported speech (Rosier 2000, Coulmas 1985)

- Interjections signal discourse report: *Mary said: Oh, I will come!*
  - When interjections are present: ‘direct’ speech
  - Lack of interjections: ‘indirect’ speech

→ Interjections are the ultimate criterion for distinguishing between direct and indirect speech

# Interjections and reported speech (our findings)

Quantitative analysis of our two biggest corpora (Chuvash and Wan):

- They interact with grammatical means of signaling the speech reporting;
- **Negative correlation** between the use of **interjections** and overt marking of reported speech by **specialized quotative** elements;
- This correlation is statistically **significant**.

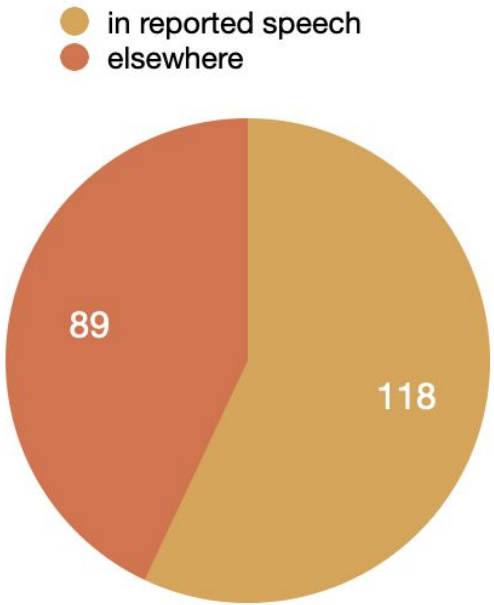
⟶ The role of interjections is even more essential than previously thought, it cannot be reduced to an direct/indirect opposition

# Wan (Mande) and Chuvash (Turkic)

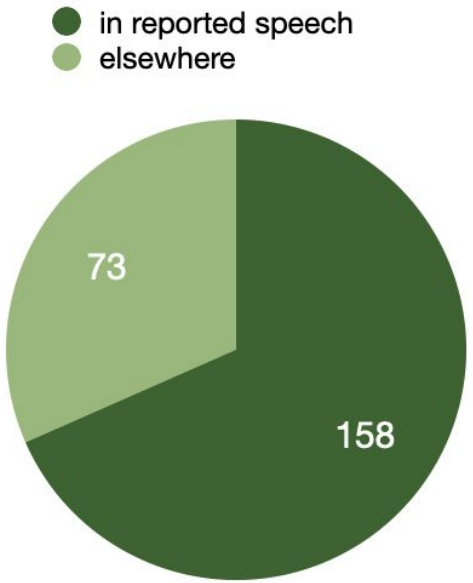
- Wan (Mande)
  - spoken in Côte d'Ivoire
  - approx. 30.000 speakers
- Chuvash (Turkic)
  - spoken in the Volga region of Russia
  - approx. 1 million speakers
- both languages have quotatives
- but the languages have a **different structure**: the quotative **precedes** the report in Wan and **follows** the report in Chuvash
- the observed effect is the same despite the ordering difference

→ this effect reflects a potentially **universal** function of interjections

# Distribution of interjections in the reported speech and elsewhere




Chuvash
(24.582 words in total)




Wan
(22.933 words in total)

# Distribution of **interjections** in the reported speech and elsewhere

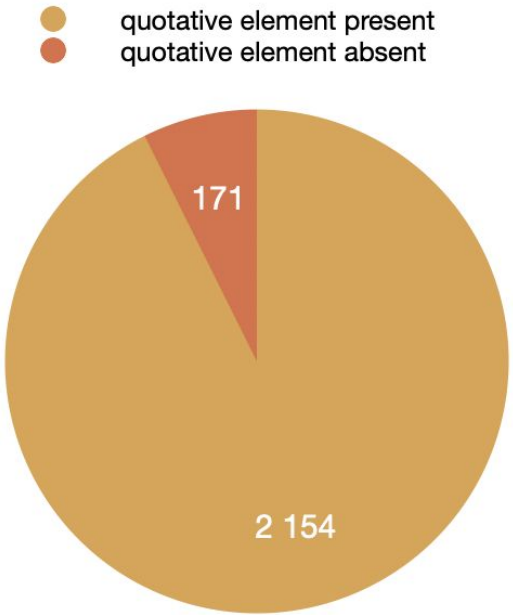
**Wan, Mande**

| *è* | *gé* | *íì* | *ɓā* | *dè* | *ɓā* *zòŋ̀* | *pà-ŋ̀* |
|---|---|---|---|---|---|---|
| 3SG.SUBJ | say | INTJ | LOG | father | LOGPROSP | be.able-PROSP |

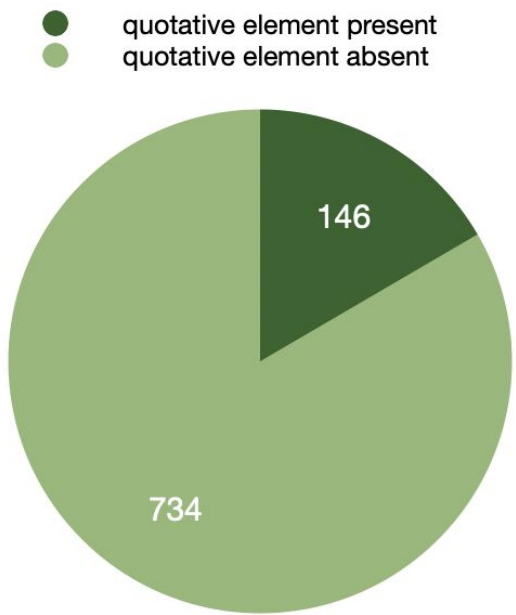
*à* *lé* *wà*
3SG on NEG

‘He (hare) said: “Ii, father, I will not be able to do it”.’

# Quotative elements in the reported speech




Chuvash
(24.582 words in total)




Wan
(22.933 words in total)

# Quotative elements in the reported speech: Chuvash (Turkic)

(4) *patʂa* ***kal-atʲ*** *atʲa* ***t-et***
tsar speak-PRS.3SG come.on QV-PRS.3SG

*ëɕkë* *töv-as* ***t-et***
feast do-PC.FUT QV-PRS.3SG
‘The tsar says: Let us make a feast.’

(5) *snatɕtɕët* *Ivan* ***kal-at***
it.means Ivan speak-PRS.3SG

*kil-e* *il-se* *kaj-ər*
home-ACC/DAT take-CV.COORD go-IMP.2PL
‘That's it, Ivan says: - Take [it] home…’

# Quotative elements in the reported speech: Wan (Mande)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (6a) | *è* | ***gé*** | *ɓlá* | *glɛ̰̄* | *pīlɔ̄ŋ̄* | *é* | *ɛ̀ŋ̀* | ***dóō*** |
| | 3SG. SUBJ | say | sheep | male | two | DEF | to | QUOT |

| | | | |
|---|---|---|---|
| *ɓāā* | *mī* | *zō* | *yɛ̀* |
| LOG+POSS | person | come:PST | here |

‘He says to the two rams: Our person came here.’

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (6b) | *ɓé* | *è* | ***gé*** | *lɛŋ* | *yɛ̰̀* | *ō* | *ɓā* | *dè* |
| | then | 3SG.SUBJ | say | to | how | PR | LOG | father |

‘He said to [him]: how [is it], my father?’

# Chuvash: interjection present → no quotative verb

| | Interjection | Total number of RS constructions |
|---|---|---|
| Quotative verb | 90 (4%) | 2154 (100%) |
| No quotative verb | **30 (18%)** | 171 (100%) |

*Table 1.* Correlation between interjections and quotatives in Chuvash

# Wan: interjection present → no quotative marker

| | Interjection | Total number of RS constructions |
|---|---|---|
| Quotative marker | 13 (9%) | 146 (100%) |
| No quotative marker | **149 (19%)** | 734 (100%) |

*Table 2.* Correlation between interjections and quotatives in Wan

# Concluding remarks

# SpeechReporting database provides

- meticulously annotated corpora of low-resourced languages
- open access tools for comparable annotation of reported discourse
- documented traditional narratives from two different regions
- potential of a new level of understanding of discursive phenomena

# Challenges

- **comparability**: hard to identify typologically applicable categories based on limited data
- **cross-checking data**: tedious task but necessary to improve quality of individual dataset
- **composition of the team**: need for more technical people who understand linguistics
- **thinking for the future**: annotation should be transparent and well documented
- **interdisciplinarity**: data for dissemination in the communities should be slightly different from linguistic documentation

# Thank you for your attention!

*http://discoursereporting.huma-num.fr/index.html*

# References


Chanard, Christian. 2019. *ELAN-CorpA*. CNRS-LLACAN. (JAVA.) (http://llacan.vjf.cnrs.fr/res_ELAN-CorpA.php)

Coulmas, Florian. 1985. Direct and Indirect speech. *Journal of Pragmatics 9 (1)*. 41-63.

Nikitina, T., Aplonova, E., Hantgan-Sonko, A., Jordanoska, I. and Perekhvalskaya, E.(eds.) *The SpeechReporting Corpus: Discourse Reporting in Storytelling.* Villejuif-Paris: CNRS-LACITO.

Nikitina, Tatiana & Hantgan, Abbie & Chanard, Christian. 2019. *Reported speech annotation template for ELAN* (SpeechReporting Corpus). Villejuif-Paris: LLACAN.

Spronck, Stef & Nikitina, Tatiana. 2019. Reported speech forms a dedicated syntactic domain. *Linguistic Typology* 23(1). 119–159.

Rosier, Laurence. 2000. Interjection, subjectivité, expressivité et discours rapporté à l'écrit : petits effets d'un petit discours. *Cahiers de praxématique* 34. 19-50.

# Additional slides

# Constructional typology of reported thought expressions

# Reported speech and reported thought

- Recent years have seen an upsurge of interest in the properties of speech reports, yet it remains an open question whether and to what extent the findings extend to attitude reports, and in particular to reported thought;
- Based on qualitative analysis of all project corpora, we suggest that in all languages, speakers have access to more than one morphosyntactic strategy for the encoding of reported thought;
- Strategies specialized for the reporting of thought co-exist with strategies associated with speech reports, and in some languages, lexical means can be used to *coerce* the reported thought reading of what is otherwise a reported thought construction.

# Three strategies of representing reported speech

- All languages seem to allow speakers to represent thought as if it were speech;
- At the same time they provide their speakers with means to distinguish reported thought from reported speech, at the morphosyntactic or lexical level:
  - Reported speech constructions are recruited for the expression of thought;
  - Some reported thought constructions have no equivalent among expressions of reported speech;
  - Reported speech constructions are coerced into a reported thought interpretation

# Conclusions

- Reported speech and reported thoughts are related but some languages have specialized means to represent thoughts;
- All languages from our sample have more than one morphosyntactic strategy to represent thoughts;
- There is a considerable heterogeneity of expressing thoughts not only across languages but within one language;
- The dual status of reported thought suggests that reported thought should not be treated as a cross-linguistic syntactic category in a way similar to reported speech (Spronck and Nikitina 2019).