Redundancy in person marking: **Subject pronoun expression** in a cross-linguistic multifactorial design

UNIVERSITY OF TARTU

UNIVERSITY OF BIRMINGHAM

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

Radboud University

Maarja-Liisa Pilvik
Mari Aigro
Piia Taremaa
Rodolfo Basile
Liina Lindström
Virve-Anneli Vihman

Justyna Mackiewicz
Dagmar Divjak
Petar Milin

Anastasia Chuprina

MEDAL
Methodological Excellence in Data-Driven Approaches to Linguistics

Funded by the European Union

UKRI UK Research and Innovation

Redundancy

Language as a communicative system is optimized for efficiency: language users aim to minimize their effort while ensuring that meanings are effectively conveyed (Levshina 2022: 6).

Despite this, languages commonly avoid optimal encoding and incorporate repetitions and redundant coding of information on all linguistic levels (Trudgill 2011; Leufkens 2020; 2022).

Redundant use of grammar has the function of making communication more robust and predictable, and it protects us against noise (Shiffrin & Schneider 1977, Levshina 2022: 9; Wit & Gillette 1999: 4; Chiari 2007: 12–13).

Pronominal subject expression

In many languages, subject pronouns are not obligatorily expressed (*pro-drop*, *null subject, variable subject pronoun expression*).

(Partial) pro-drop languages include Finno-Ugric, Slavic, Romance languages, but also Greek, Arabic, Hindi, Japanese, and others.

EST:	***Ma***	*kuula-**n***	*muusika-t*.
	1SG	listen-1SG.PRS	music-PART
POL:	***Ja***	*słucha-**m***	*muzyki*
	1SG	listen-1SG.PRS	music.GEN

‘I am listening to music.’

Pronominal subject expression

In many languages, subject pronouns are not obligatorily expressed (*pro-drop*, *null subject, variable subject pronoun expression*).

(Partial) pro-drop languages include Finno-Ugric, Slavic, Romance languages, but also Greek, Arabic, Hindi, Japanese, and others.

EST:	Ø	*kuula-**n***	*muusika-t*.
		listen-1SG.PRS	music-PART
POL:	Ø	*słucha-**m***	*muzyki*
		listen-1SG.PRS	music.GEN

‘I am listening to music.’

Pronominal subject expression

Berdicevskis, Schmidtke-Bode and Seržant (2020) based on WALS data and Treebank corpus data of Slavic languages:

> There is a cross-linguistic **tendency to use only one referential device for encoding subject** and to avoid double marking or non-marking. This is motivated by **efficiency of communication**
>
> > “that equilibrates production effort and the robustness of information transfer”,
>
> while double encoding
>
> > “is **obviously redundant** in pragmatically unmarked topic-comment clauses and is therefore **costlier than necessary**”.

Pronominal subject expression

Torres Cacoullos & Travis (2019: 655)

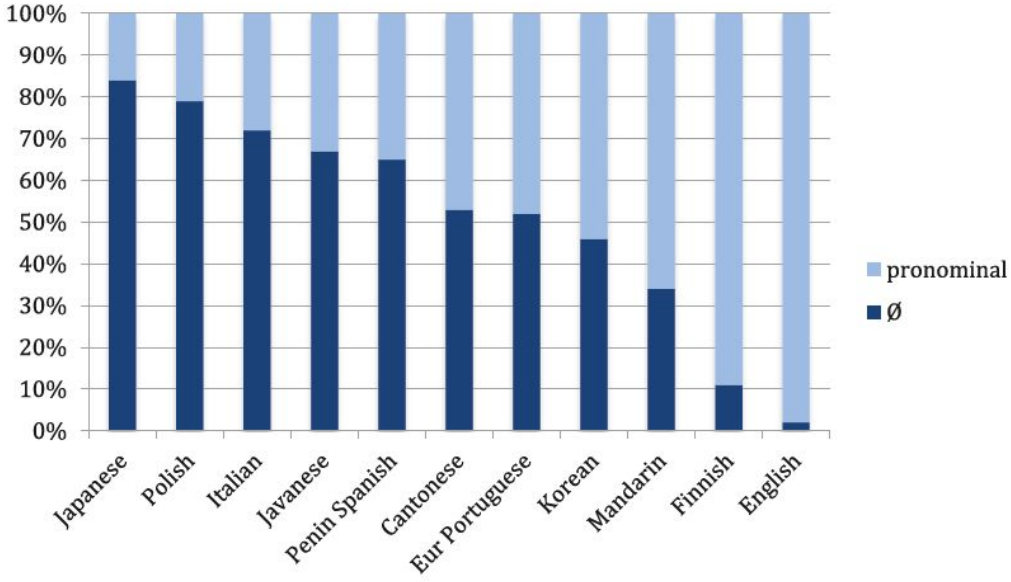


Figure 1: Rates of 1sg subject expression (Ø vs. pronominal) across different languages. Sources and rates of unexpressed 1sg subjects: Japanese 84% (Lee and Yonezawa 2008: 738, N = 1571), Polish 79% (Chociej 2011: 52, N = 536), Italian 72% (Nagy p.c. cf. Nagy et al. 2011, N = 224), Javanese 67% (Ewing 2014: 51, N = 289), Peninsular Spanish 65% (Posio 2013: 269, N = 787), Cantonese 53% (Nagy p.c. cf. Nagy et al. 2011, N = 362), European Portuguese 51% (Posio 2013: 269, N = 704), Korean 46% (Oh 2007: 466, N = 433), Mandarin 34% (Jia and Bayley 2002: 13, N = 393), Finnish 11% (Helasvuo 2014: 68, N = 1793), English ~ 2% (Torres Cacoullos and Travis 2014: 22, N = 6,600 (estimated)).

Pronominal subject expression

Torres Cacoullos & Travis (2019: 655)

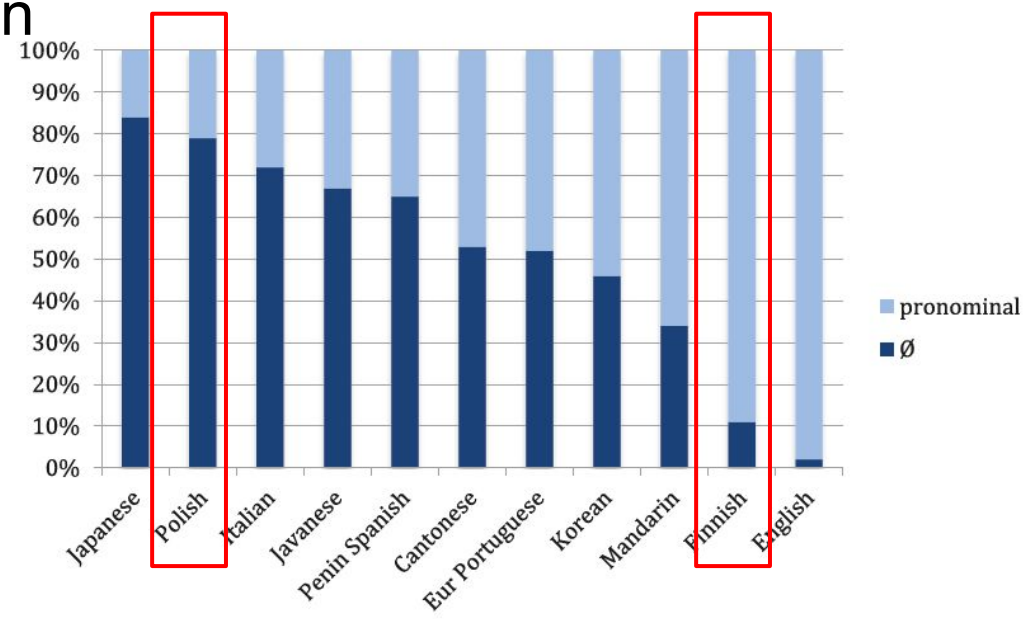


Figure 1: Rates of 1sg subject expression (Ø vs. pronominal) across different languages. Sources and rates of unexpressed 1sg subjects: Japanese 84% (Lee and Yonezawa 2008: 738, N = 1571), Polish 79% (Chociej 2011: 52, N = 536), Italian 72% (Nagy p.c. cf. Nagy et al. 2011, N = 224), Javanese 67% (Ewing 2014: 51, N = 289), Peninsular Spanish 65% (Posio 2013: 269, N = 787), Cantonese 53% (Nagy p.c. cf. Nagy et al. 2011, N = 362), European Portuguese 51% (Posio 2013: 269, N = 704), Korean 46% (Oh 2007: 466, N = 433), Mandarin 34% (Jia and Bayley 2002: 13, N = 393), Finnish 11% (Helasvuo 2014: 68, N = 1793), English ~ 2% (Torres Cacoullos and Travis 2014: 22, N = 6,600 (estimated)).

Pronominal subject expression

Torres Cacoullos & Travis (2019: 655)

"***Grammatical (dis)similarity*** *is detected not by the presence or absence of a feature,* ***nor by its overall rate of use****. The loci of cross-language comparisons are instead both* ***the probabilistic constraints and the variable context*** *within which they are operative.*" (Torres Cacoullos & Travis 2019: 682)

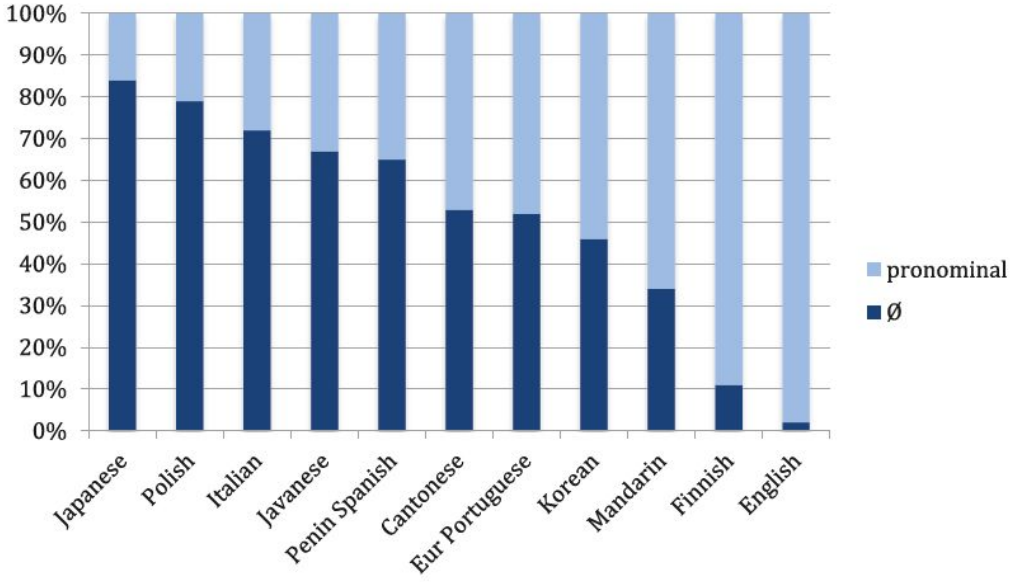


Figure 1: Rates of 1sg subject expression (Ø vs. pronominal) across different languages. Sources and rates of unexpressed 1sg subjects: Japanese 84% (Lee and Yonezawa 2008: 738, N = 1571), Polish 79% (Chociej 2011: 52, N = 536), Italian 72% (Nagy p.c. cf. Nagy et al. 2011, N = 224), Javanese 67% (Ewing 2014: 51, N = 289), Peninsular Spanish 65% (Posio 2013: 269, N = 787), Cantonese 53% (Nagy p.c. cf. Nagy et al. 2011, N = 362), European Portuguese 51% (Posio 2013: 269, N = 704), Korean 46% (Oh 2007: 466, N = 433), Mandarin 34% (Jia and Bayley 2002: 13, N = 393), Finnish 11% (Helasvuo 2014: 68, N = 1793), English ~ 2% (Torres Cacoullos and Travis 2014: 22, N = 6,600 (estimated)).

(Some) factors found to influence pronominal subject expression

STRUCTURAL-GRAMMATICAL

Syntactic complexity:

- coordination
  (Torres Cacoullos & Travis 2014)
- subordination
  (Lindström et al. 2009, Väänänen 2016)
- length of syntactic unit
  (Helasvuo & Kyröläinen 2016)
- transitivity
  (Orozco & Hurtado 2021)
- complexity of the verb phrase
  (Wagner 2018)

Grammatical categories:

- tense, aspect, polarity
  (Lindström et al. 2009, Väänänen 2016, Orozco & Hurtado 2021, Erker & Guy 2012, Nagy 2015)

(Some) factors found to influence pronominal subject expression

STRUCTURAL-GRAMMATICAL

Syntactic complexity:

- coordination
  (Torres Cacoullos & Travis 2014)
- subordination
  (Lindström et al. 2009, Väänänen 2016)
- length of syntactic unit
  (Helasvuo & Kyröläinen 2016)
- transitivity
  (Orozco & Hurtado 2021)
- complexity of the verb phrase
  (Wagner 2018)

Grammatical categories:

- tense, aspect, polarity
  (Lindström et al. 2009, Väänänen 2016, Orozco & Hurtado 2021, Erker & Guy 2012, Nagy 2015)

PROCESSING & MEMORY

Discourse continuity & unambiguity:

- recency of mention
  (Torres Cacoullos & Travis 2014, 2016)
- referential distance
  (Lindström et al. 2009, Väänänen 2016, Helasvuo & Kyröläinen 2016)
- subject/topic continuity
  (Givón 1983)
- referent switch
  (Erker & Guy 2012, Wagner 2018)

Structural persistence
(Helasvuo & Kyröläinen 2016, Torres Cacoullos & Travis 2014, 2016, 2019, Orozco & Hurtado 2021)

Constructional entrenchment / conventionalization
(Helasvuo 2014a, Laury, Helasvuo & Rauma 2020, Orozco & Hurtado 2021)

(Some) factors found to influence pronominal subject expression

STRUCTURAL-GRAMMATICAL

Syntactic complexity:

- coordination
  (Torres Cacoullos & Travis 2014)
- subordination
  (Lindström et al. 2009, Väänänen 2016)
- length of syntactic unit
  (Helasvuo & Kyröläinen 2016)
- transitivity
  (Orozco & Hurtado 2021)
- complexity of the verb phrase
  (Wagner 2018)

Grammatical categories:

- tense, aspect, polarity
  (Lindström et al. 2009, Väänänen 2016, Orozco & Hurtado 2021, Erker & Guy 2012, Nagy 2015)

PROCESSING & MEMORY

Discourse continuity & unambiguity:

- recency of mention
  (Torres Cacoullos & Travis 2014, 2016)
- referential distance
  (Lindström et al. 2009, Väänänen 2016, Helasvuo & Kyröläinen 2016)
- subject/topic continuity
  (Givón 1983)
- referent switch
  (Erker & Guy 2012, Wagner 2018)

Structural persistence
(Helasvuo & Kyröläinen 2016, Torres Cacoullos & Travis 2014, 2016, 2019, Orozco & Hurtado 2021)

Constructional entrenchment / conventionalization
(Helasvuo 2014a, Laury, Helasvuo & Rauma 2020, Orozco & Hurtado 2021)

SEMANTIC

Verb type, esp. cognition verbs
(Posio 2014, Helasvuo 2014b, Torres Cacoullos & Travis 2019)

(Some) factors found to influence pronominal subject expression

STRUCTURAL-GRAMMATICAL

Syntactic complexity:

- coordination
  (Torres Cacoullos & Travis 2014)
- subordination
  (Lindström et al. 2009, Väänänen 2016)
- length of syntactic unit
  (Helasvuo & Kyröläinen 2016)
- transitivity
  (Orozco & Hurtado 2021)
- complexity of the verb phrase
  (Wagner 2018)

Grammatical categories:

- tense, aspect, polarity
  (Lindström et al. 2009, Väänänen 2016, Orozco & Hurtado 2021, Erker & Guy 2012, Nagy 2015)

PROCESSING & MEMORY

Discourse continuity & unambiguity:

- recency of mention
  (Torres Cacoullos & Travis 2014, 2016)
- referential distance
  (Lindström et al. 2009, Väänänen 2016, Helasvuo & Kyröläinen 2016)
- subject/topic continuity
  (Givón 1983)
- referent switch
  (Erker & Guy 2012, Wagner 2018)

Structural persistence
(Helasvuo & Kyröläinen 2016, Torres Cacoullos & Travis 2014, 2016, 2019, Orozco & Hurtado 2021)

Constructional entrenchment / conventionalization
(Helasvuo 2014a, Laury, Helasvuo & Rauma 2020, Orozco & Hurtado 2021)

SEMANTIC

Verb type, esp. cognition verbs
(Posio 2014, Helasvuo 2014b, Torres Cacoullos & Travis 2019)

REGISTER DIFFERENCES

(Helavsuo & Kyröläinen 2016, Helasvuo 2014b)

Questions

1. How ‘redundant’ is 1st and 2nd person marking in 4 languages originating from 2 different language families: Estonian (Finno-Ugric), Finnish (Finno-Ugric), Russian (Slavic), and Polish (Slavic)?

    What is the overall distribution of expressed vs. unexpressed pronouns in the 4 languages?

Questions

1. How ‘redundant’ is 1st and 2nd person marking in 4 languages originating from 2 different language families: Estonian (Finno-Ugric), Finnish (Finno-Ugric), Russian (Slavic), and Polish (Slavic)?

   What is the overall distribution of expressed vs. unexpressed pronouns in the 4 languages?

2. When does redundant marking happen?

   Which factors are significantly associated with subject pronoun expression intra- and cross-linguistically?

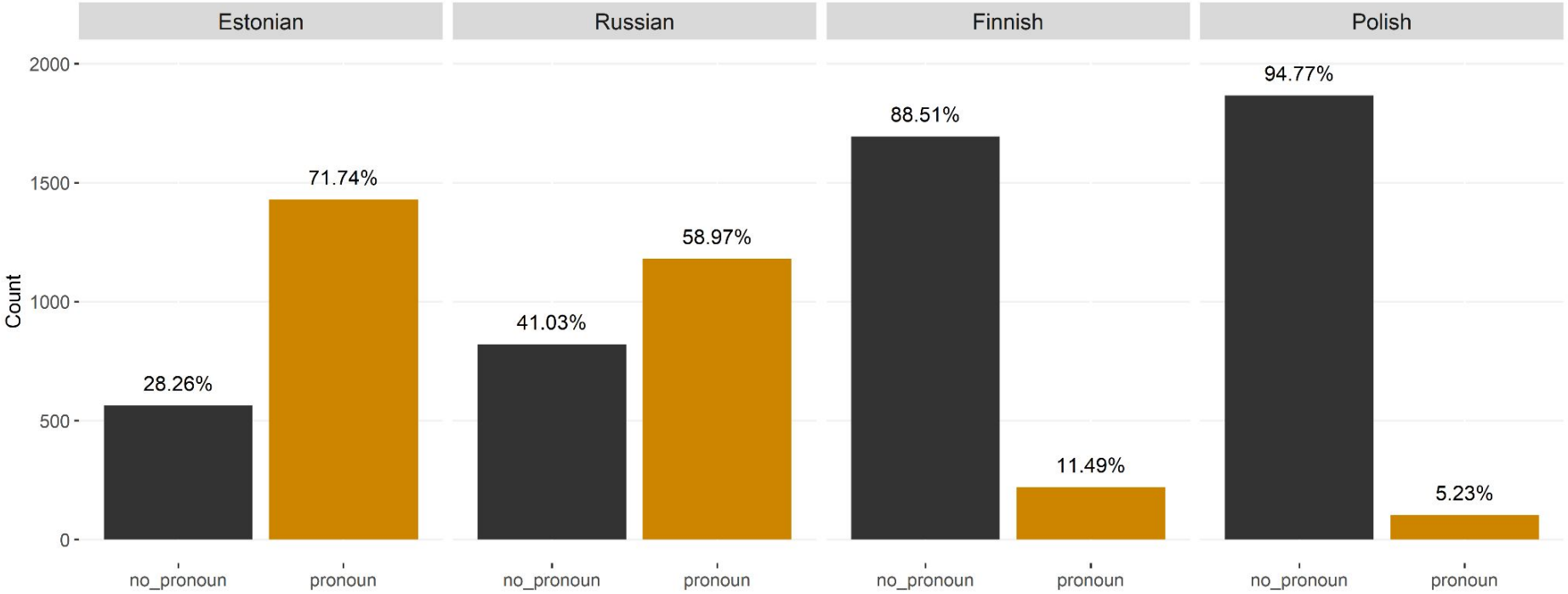
   In which direction do the effects affect subject pronoun expression?

Data

- Only 1SG, 2SG, 1PL,and 2PL (sampled proportionally to their frequencies in the corpora)
- Only indicative mood
- No negation
- No coordination
- No nonreferential uses (e.g., particles)

	Finno-Ugric	Slavic
Spontaneous speech	**ESTONIAN** The Phonetic Corpus of Estonian Spontaneous Speech (Lippus et al. 2023) ~ 600,000 tokens **1992** obs.	**RUSSIAN** Russian National Corpus spoken subcorpus (Savchuk et al. 2024) ~ 7,000,000 tokens **2001** obs. (1193)
Subtitles	**FINNISH** Open Subtitles 2018 (Lison & Tiedemann 2016) ~ 175,000,000 tokens **1914** obs. (1910)	**POLISH** Open Subtitles 2018 (Lison & Tiedemann 2016) ~ 496,000,000 tokens **1969** obs.

Results 1: Distribution of expressed vs. unexpressed pronouns



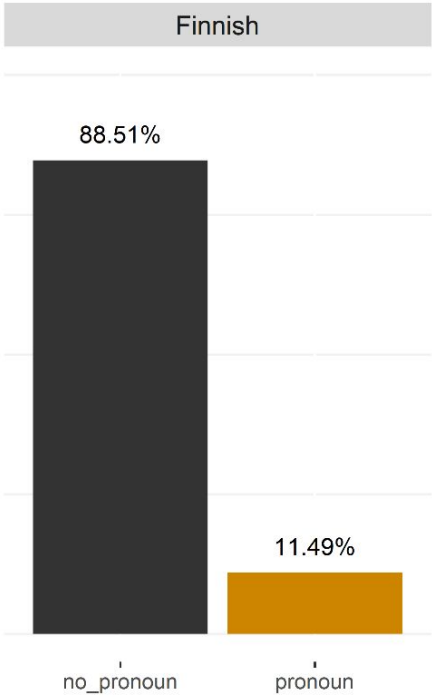
Results 1: Distribution of expressed vs. unexpressed pronouns

Expressed pronoun rates in Finnish

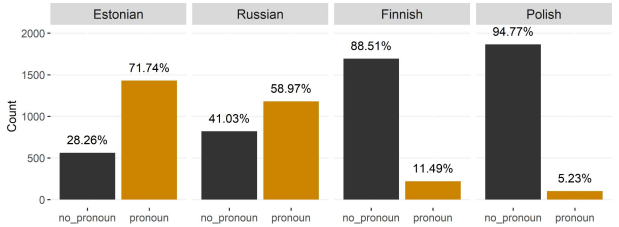
Helasvuo 2014a:
conversational data 1SG **89%,** 2SG **76%**

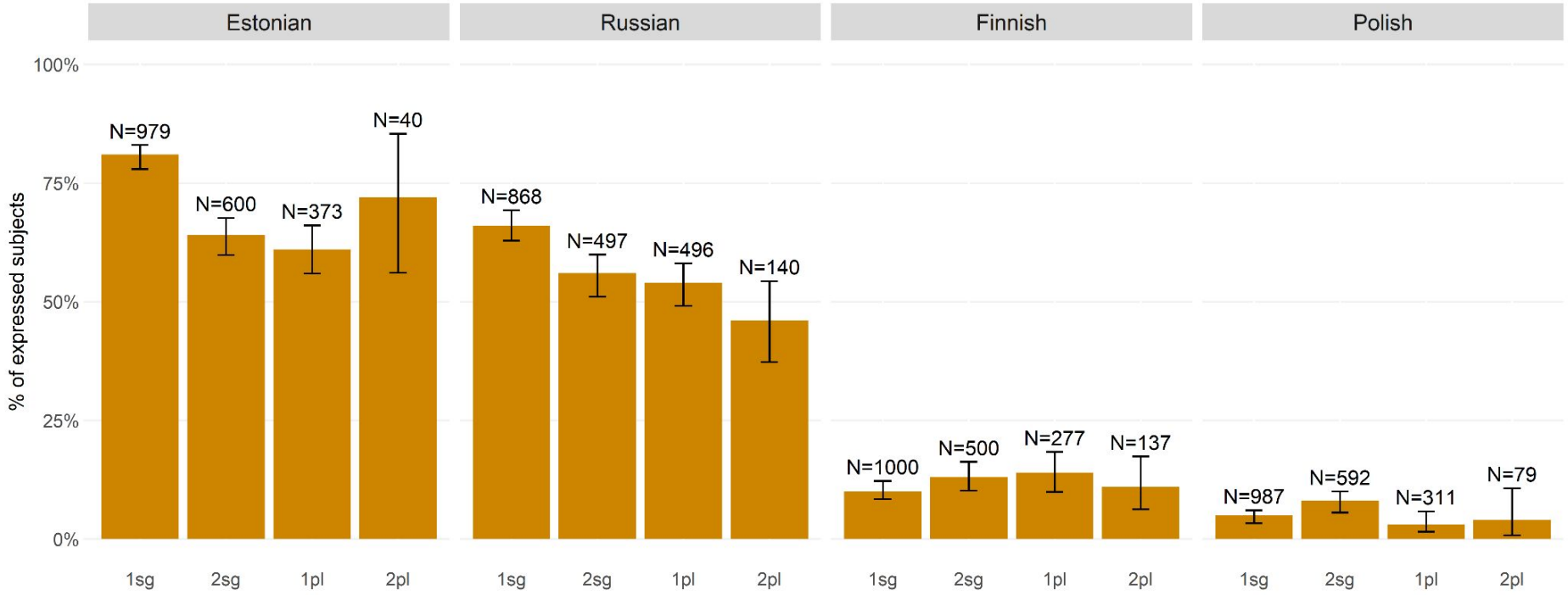
Duvallon & Chalvin 2004:
radio debates **60%**

Helasvuo 2014b:
written Finnish **20%**
text messages **19%**



Results 1


Estonian
Russian
Finnish
Polish
Count
2000
1500
1000
500
0
28.26%
71.74%
41.03%
58.97%
88.51%
11.49%
94.77%
5.23%
no_pronoun
pronoun


Estonian
Russian
Finnish
Polish
% of expressed subjects
100%
75%
50%
25%
0%
N=979
N=600
N=373
N=40
N=868
N=497
N=496
N=140
N=1000
N=500
N=277
N=137
N=987
N=592
N=311
N=79
1sg
2sg
1pl
2pl

Explanatory factors (structural)

- **Tense of the finite verb form** (TENSE: past, present, future)
- **Complexity of the verb** (VERB_COMPLEXITY: main, non-main)
- **Complexity of the verb phrase** (VP_COMPLEXITY: compl, no_compl)
- **Subordination** (CLAUSE_TYPE: main, subord)
- **Sentence type** (SENTENCE_TYPE: declarative, interrogative)
- **Person** (PERSON: 1, 2)
- **Number** (NUMBER: SG, PL)

Explanatory factors (structural, semantic)

- **Tense of the finite verb form** (TENSE: past, present, future)
- **Complexity of the verb** (VERB_COMPLEXITY: main, non-main)
- **Complexity of the verb phrase** (VP_COMPLEXITY: compl, no_compl)
- **Subordination** (CLAUSE_TYPE: main, subord)
- **Sentence type** (SENTENCE_TYPE: declarative, interrogative)
- **Person** (PERSON: 1, 2)
- **Number** (NUMBER: SG, PL)

- **Semantic type of the main verb** (VERB_TYPE: communicative, mental, perception, other)

Explanatory factors (structural, semantic & discourse-related)

- **Tense of the finite verb form** (TENSE: past, present, future)
- **Complexity of the verb** (VERB_COMPLEXITY: main, non-main)
- **Complexity of the verb phrase** (VP_COMPLEXITY: compl, no_compl)
- **Subordination** (CLAUSE_TYPE: main, subord)
- **Sentence type** (SENTENCE_TYPE: declarative, interrogative)
- **Person** (PERSON: 1, 2)
- **Number** (NUMBER: SG, PL)

- **Semantic type of the main verb** (VERB_TYPE: communicative, mental, perception, other)

- **Referent switch compared to the previous clause** (REFERENT_SWITCH: same, switch)
- **Mention of the referent in the immediately preceding context** (REF_SAMESPEAKER: yes, no, REF_ANYSPEAKER: yes, no)
- **Form of the mention in the immediately preceding context** (REFFORM_SAMESPEAKER: verb, pronoun, no, REFFORM_ANYSPEAKER: verb, pronoun, no)

Results 2

Which factors are significantly associated with subject pronoun expression?

1. Univariate analyses:

- Chi-squared test of independence;
- Simple log-linear models

Results 2

Which factors are significantly associated with subject pronoun expression?

1. Univariate analyses:

- Chi-squared test of independence;
- Simple log-linear models

	EST	RUS	FIN	POL
TENSE (past, present, future$_{SL}$)		***		
VERB_COMPLEXITY (main, non-main)			**	***
VP_COMPLEXITY (compl, no_compl)	***	***		(*)
CLAUSE_TYPE (main, subord)	***			***
SENTENCE_TYPE (declarative, interrogative)				
PERSON (1, 2)	***	**		**
NUMBER (SG, PL)	***	**		
VERB_TYPE (comm_ment_perc, other)	***	***	*	
REF_SAMESPEAKER (yes, no)	**	*	**	**
REF_ANYSPEAKER (yes, no)	**	*	**	*
REFFORM_SAMESPEAKER (verb, pronoun, no)	***	***	**	***
REFFORM_ANYSPEAKER (verb, pronoun, no)	***	***	**	***
REFERENT_SWITCH (same, switch)	***	***		NA

Results 2

Which factors are significantly associated with subject pronoun expression?

	EST	RUS	FIN	POL
TENSE (past, present, future$_{SL}$)		***		
VERB_COMPLEXITY (main, non-main)			**	***
VP_COMPLEXITY (compl, no_compl)	***	***		(*)
CLAUSE_TYPE (main, subord)	***			***
SENTENCE_TYPE (declarative, interrogative)				
PERSON (1, 2)	***	**		**
NUMBER (SG, PL)	***	**		
VERB_TYPE (comm_ment_perc, other)	***	***	*	
REF_SAMESPEAKER (yes, no)	**	*	**	**
REF_ANYSPEAKER (yes, no)	**	*	**	*
REFFORM_SAMESPEAKER (verb, pronoun, no)	***	***	**	***
REFFORM_ANYSPEAKER (verb, pronoun, no)	***	***	**	***
REFERENT_SWITCH (same, switch)	***	***		NA

Results 2

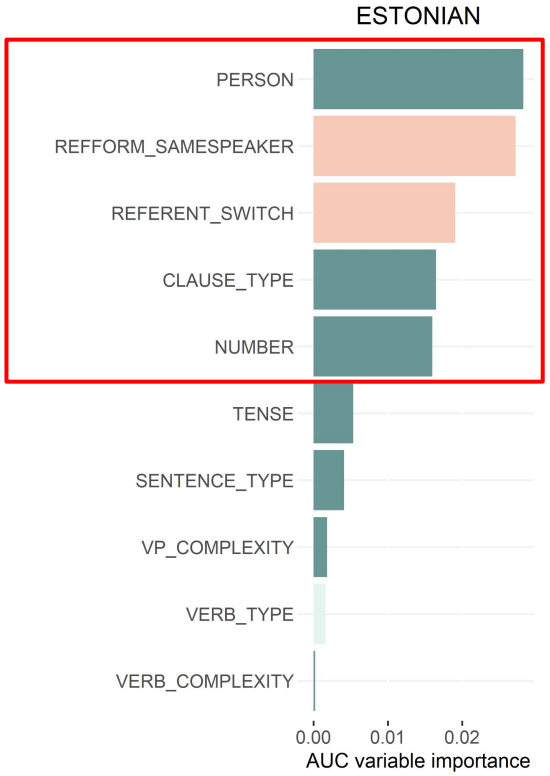
Which factors are significantly associated with subject pronoun expression?

2. Multivariate analyses:

- Conditional Random Forests;
- Complex log-linear models

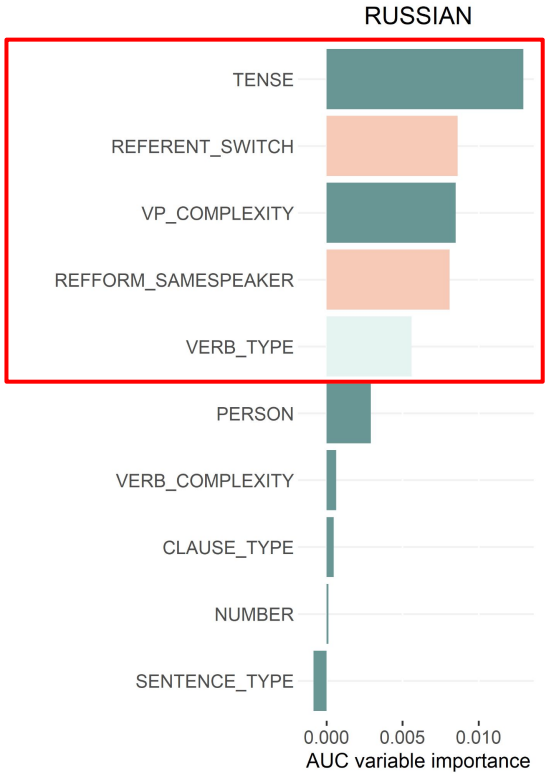
	EST	RUS	FIN	POL
TENSE (past, present, future$_{SL}$)		***		
VERB_COMPLEXITY (main, non-main)			**	***
VP_COMPLEXITY (compl, no_compl)	***	***		(*)
CLAUSE_TYPE (main, subord)	***			***
SENTENCE_TYPE (declarative, interrogative)				
PERSON (1, 2)	***	**		**
NUMBER (SG, PL)	***	**		
VERB_TYPE (comm_ment_perc, other)	***	***	*	
REF_SAMESPEAKER (yes, no)	**	*	**	**
REF_ANYSPEAKER (yes, no)	**	*	**	*
REFFORM_SAMESPEAKER (verb, pronoun, no)	***	***	**	***
REFFORM_ANYSPEAKER (verb, pronoun, no)	***	***	**	***
REFERENT_SWITCH (same, switch)	***	***		NA

Results 2



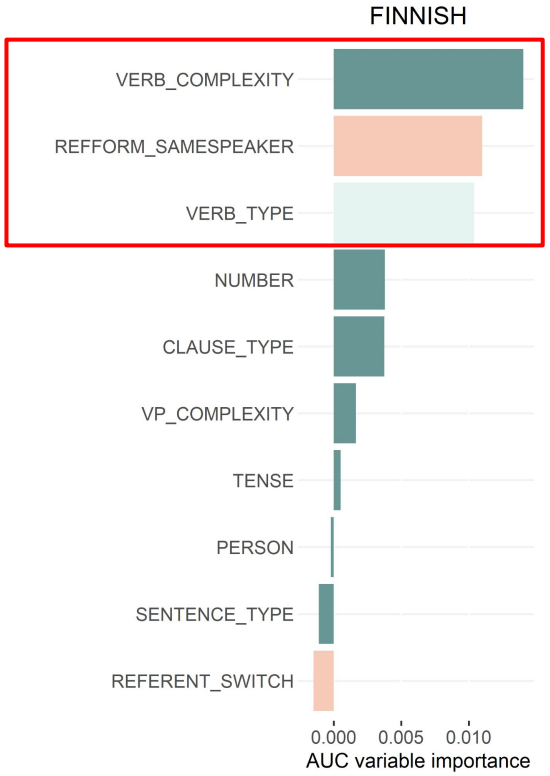
	EST	RUS	FIN	POL
TENSE (past, present, future$_{SL}$)		***		
VERB_COMPLEXITY (main, non-main)			**	***
VP_COMPLEXITY (compl, no_compl)	***	***		(*)
CLAUSE_TYPE (main, subord)	***			***
SENTENCE_TYPE (declarative, interrogative)				
PERSON (1, 2)	***	**		**
NUMBER (SG, PL)	***	**		
VERB_TYPE (comm_ment_perc, other)	***	***	*	
REF_SAMESPEAKER (yes, no)	**	*	**	**
REF_ANYSPEAKER (yes, no)	**	*	**	*
REFFORM_SAMESPEAKER (verb, pronoun, no)	***	***	**	***
REFFORM_ANYSPEAKER (verb, pronoun, no)	***	***	**	***
REFERENT_SWITCH (same, switch)	***	***		NA

Results 2



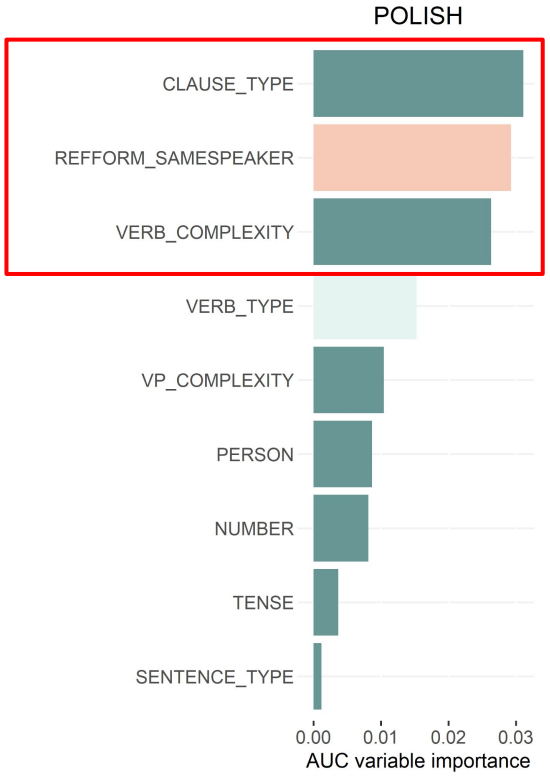
	EST	RUS	FIN	POL
TENSE (past, present, future$_{SL}$)		***		
VERB_COMPLEXITY (main, non-main)			**	***
VP_COMPLEXITY (compl, no_compl)	***	***		(*)
CLAUSE_TYPE (main, subord)	***			***
SENTENCE_TYPE (declarative, interrogative)				
PERSON (1, 2)	***	**		**
NUMBER (SG, PL)	***	**		
VERB_TYPE (comm_ment_perc, other)	***	***	*	
REF_SAMESPEAKER (yes, no)	**	*	**	**
REF_ANYSPEAKER (yes, no)	**	*	**	*
REFFORM_SAMESPEAKER (verb, pronoun, no)	***	***	**	***
REFFORM_ANYSPEAKER (verb, pronoun, no)	***	***	**	***
REFERENT_SWITCH (same, switch)	***	***		NA

Results 2



	EST	RUS	FIN	POL
TENSE (past, present, future$_{SL}$)		***		
VERB_COMPLEXITY (main, non-main)			**	***
VP_COMPLEXITY (compl, no_compl)	***	***		(*)
CLAUSE_TYPE (main, subord)	***			***
SENTENCE_TYPE (declarative, interrogative)				
PERSON (1, 2)	***	**		**
NUMBER (SG, PL)	***	**		
VERB_TYPE (comm_ment_perc, other)	***	***	*	
REF_SAMESPEAKER (yes, no)	**	*	**	**
REF_ANYSPEAKER (yes, no)	**	*	**	*
REFFORM_SAMESPEAKER (verb, pronoun, no)	***	***	**	***
REFFORM_ANYSPEAKER (verb, pronoun, no)	***	***	**	***
REFERENT_SWITCH (same, switch)	***	***		NA

Results 2



	EST	RUS	FIN	POL
TENSE (past, present, future$_{SL}$)		***		
VERB_COMPLEXITY (main, non-main)			**	***
VP_COMPLEXITY (compl, no_compl)	***	***		(*)
CLAUSE_TYPE (main, subord)	***			***
SENTENCE_TYPE (declarative, interrogative)				
PERSON (1, 2)	***	**		**
NUMBER (SG, PL)	***	**		
VERB_TYPE (comm_ment_perc, other)	***	***	*	
REF_SAMESPEAKER (yes, no)	**	*	**	**
REF_ANYSPEAKER (yes, no)	**	*	**	*
REFFORM_SAMESPEAKER (verb, pronoun, no)	***	***	**	***
REFFORM_ANYSPEAKER (verb, pronoun, no)	***	***	**	***
REFERENT_SWITCH (same, switch)	***	***		NA

Results 2

	EST	RUS	FIN	POL
REFFORM_SAMESPEAKER (verb, pronoun, no)	***	***	**	***
VERB_TYPE (comm_ment_perc, other)	(***)	***	*	
PERSON (1, 2)	***	(**)		(**)
VP_COMPLEXITY (compl, no_compl)	(***)	***		(*)
REFERENT_SWITCH (same, switch)	***	***		NA
CLAUSE_TYPE (main, subord)	***			***
VERB_COMPLEXITY (main, non-main)			**	***
NUMBER (SG, PL)	***	(**)		
TENSE (past, present, future$_{SL}$)		***		
SENTENCE_TYPE (declarative, interrogative)				

Results 2

	EST	RUS	FIN	POL
REFFORM_SAMESPEAKER (verb, pronoun, no)	pronoun ↑ verb ↓	pronoun ↑ verb ↓	pronoun/no ↑ verb ↓	pronoun/no ↑ verb ↓
VERB_TYPE (comm_ment_perc, other)	(***)	comm_ment_perc ↑ other ↓	other ↑ comm_ment_perc ↓	
PERSON (1, 2)	1 ↑ 2 ↓	(**)		(**)
VP_COMPLEXITY (compl, no_compl)	(***)	compl ↑ no_compl ↓		(*)
REFERENT_SWITCH (same, switch)	switch ↑ same ↓	switch ↑ same ↓		NA
CLAUSE_TYPE (main, subord)	subord ↑ main ↓			main ↑ subord ↓
VERB_COMPLEXITY (main, non-main)			main ↑ non-main↓	main ↑ non-main↓
NUMBER (SG, PL)	SG ↑ PL ↓	(**)		
TENSE (past, present, future$_{SL}$)		future ↑ past ↓		
SENTENCE_TYPE (declarative, interrogative)				

Results 2

	EST	RUS	FIN	POL
REFFORM_SAMESPEAKER (verb, pronoun, no)	pronoun ↑ verb ↓	pronoun ↑ verb ↓	pronoun/no ↑ verb ↓	pronoun/no ↑ verb ↓
VERB_TYPE (comm_ment_perc, other)	(comm_ment_perc ↑ other ↓)	comm_ment_perc ↑ other ↓	other ↑ comm_ment_perc ↓	
PERSON (1, 2)	1 ↑ 2 ↓	(1 ↑ 2 ↓)		(2 ↑ 1 ↓)
VP_COMPLEXITY (compl, no_compl)	(compl ↑ no_compl ↓)	compl ↑ no_compl ↓		(compl ↑ no_compl ↓)
REFERENT_SWITCH (same, switch)	switch ↑ same ↓	switch ↑ same ↓		NA
CLAUSE_TYPE (main, subord)	subord ↑ main ↓			main ↑ subord ↓
VERB_COMPLEXITY (main, non-main)			main ↑ non-main↓	main ↑ non-main↓
NUMBER (SG, PL)	SG ↑ PL ↓	(SG ↑ PL ↓)		
TENSE (past, present, future$_{SL}$)		future ↑ past ↓		
SENTENCE_TYPE (declarative, interrogative)				

Answering the research questions

1. How ‘redundant’ is 1st and 2nd person marking in 4 languages originating from 2 different language families: Estonian (Finno-Ugric), Finnish (Finno-Ugric), Russian (Slavic), and Polish (Slavic)?

Answering the research questions

1. How ‘redundant’ is 1st and 2nd person marking in 4 languages originating from 2 different language families: Estonian (Finno-Ugric), Finnish (Finno-Ugric), Russian (Slavic), and Polish (Slavic)?

There are **significant differences between the 4 languages**: highest subject pronoun expression rate in Estonian (74.71%), lowest in Polish (5.23%).

- Mode of communication is important! **Reduction of grammatical redundancy in subtitles.** → Edited, space restrictions, processing of two signals (auditive and visual language) at a time. More contextual (incl. visual) cues help the speaker disambiguate the referent.

Answering the research questions

2. When does redundant marking happen?

Answering the research questions

2. When does redundant marking happen?

Factors related to (online) **processing, memory, and discourse continuity** show unidirectional effects in all languages.

Recent activation of the same referent in discourse has a decreasing effect on pronoun expression, but less so when the reference is made using a verb form. →

- Potential effects of **narrative vs. dialogic** discourse.

Answering the research questions

2. When does redundant marking happen?

Factors related to (online) **processing, memory, and discourse continuity** show unidirectional effects in all languages.

Recent activation of the same referent in discourse has a decreasing effect on pronoun expression, but more so when the reference is made using a verb form. →

- Potential effects of **narrative vs. dialogic** discourse.

Explicit signaling of referent switches is important only in spoken spontaneous language. →

- **Processing constraints** in (on-line) spoken language;
- **visual cues** help with referent tracking in subtitles;
- spoken data dialogic, subtitles might include more narrative/monologous sequences;
- different **pragmatic and stylistic considerations** in spoken data and subtitles.

Answering the research questions

2. When does redundant marking happen?

Answering the research questions

2. When does redundant marking happen?

Factors related to **grammar and semantics** show more diverse effects (both with regard to significance and direction of the effects). Mode of communication > language family.

Answering the research questions

2. When does redundant marking happen?

Factors related to **grammar and semantics** show more diverse effects (both with regard to significance and direction of the effects). Mode of communication > language family.

Similar effects between spoken and subtitle data are found with verb phrase complexity.

- Explicit pronoun preferred with complex verb phrases ← **need to disambiguate between participants**.

Answering the research questions

2. When does redundant marking happen?

Factors related to **grammar and semantics** show more diverse effects (both with regard to significance and direction of the effects). Mode of communication > language family.

Similar effects between spoken and subtitle data are found with verb phrase complexity.

- Explicit pronoun preferred with complex verb phrases ← **need to disambiguate between participants**.

Diverging effects between spoken and subtitle data include verb type, person, clause type. →

- Spoken language shows preference for explicit pronouns with communicative, perception, and mental verbs, 1st person, and subordinate clauses ← **speaker-inclusive participants more salient**; **stress patterns**; (semi-)**automatized sequences** (e.g., *I don't know*, *I think*, *I believe*...).

Answering the research questions

2. When does redundant marking happen?

Factors related to **grammar and semantics** show more diverse effects (both with regard to significance and direction of the effects). Mode of communication > language family.

Similar effects between spoken and subtitle data are found with verb phrase complexity.

- Explicit pronoun preferred with complex verb phrases ← **need to disambiguate between participants**.

Diverging effects between spoken and subtitle data include verb type, person, clause type. →

- Spoken language shows preference for explicit pronouns with communicative, perception, and mental verbs, 1st person, and subordinate clauses ← **speaker-inclusive participants more salient**; **stress patterns**; (semi-)**automatized sequences** (e.g., *I don't know*, *I think*, *I believe*...).

Effects significant only in spoken data include number and tense.

- Explicit pronoun preferred with singular and past tense ← **speech situation participants more salient**; **no other grammatical person coding** in past tense (RUS).

Answering the research questions

2. When does redundant marking happen?

Factors related to **grammar and semantics** show more diverse effects (both with regard to significance and direction of the effects). Mode of communication > language family.

Similar effects between spoken and subtitle data are found with verb phrase complexity.

- Explicit pronoun preferred with complex verb phrases ← **need to disambiguate between participants**.

Diverging effects between spoken and subtitle data include verb type, person, clause type. →

- Spoken language shows preference for explicit pronouns with communicative, perception, and mental verbs, 1st person, and subordinate clauses ← **speaker-inclusive participants more salient**; **stress patterns**; (semi-)**automatized sequences** (e.g., *I don't know*, *I think*, *I believe*...).

Effects significant only in spoken data include number and tense.

- Explicit pronoun preferred with singular and past tense ← **speech situation participants more salient**; **no other grammatical person coding** in past tense (RUS).

Effects significant only in subtitle data are found with verb complexity.

- Explicit pronoun preferred with finite main verbs ← **shorter verb forms leave more space for pronouns**.

Discussion

Is subject expression redundant?

For producing and processing written language: probably yes.

For producing and processing online spontaneous speech: probably no.

→ **Experimental data needed!**

(Me) täname!

(We) thank you!

References

Berdicevskis,Aleksandrs, Karsten Schmidtke-Bode & Ilja Seržant. 2020. Subjects tend to be coded only once: Corpus-based and grammar-based evidence for an efficiency-driven trade-off. Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories, pages 79–92, Düsseldorf, Germany. Association for Computational Linguistics.

Chiari, Isabella. 2007. Redundancy Elimination: The Case of Artificial Languages. Journal of Universal Language 8(2). 7–38. https://doi.org/10.22425/jul.2007.8.2.7.

Duvallon, Outi & Antoine Chalvin. 2004. La réalisation zéro du pronom sujet de première et de deuxième personne du singulier en finnois et en estonien parlés. Linguistica Uralica 40(4). 270–286.

Erker, Daniel & Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. Language 88(3). 526–557.

Helasvuo, Marja-Liisa. 2014a. Searching for motivations for grammatical patternings. Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA) 24(3). 453–476.

Helasvuo, Marja-Liisa. 2014b. “Jotta suomalaiset voisivat puhua enemmän”. Puhetilanteen osallistujat tekstiviestikeskustelussa. In Marja-Liisa Helasvuo, Marjut Johannson & Sanna-Kaisa Tanskanen (eds.), Kieli verkossa. Näkökulmia digitaaliseen vuorovaikutukseen, 29–49. Helsinki: Suomalaisen Kirjallisuuden Seura.

Helasvuo, Marja-Liisa & Aki-Juhani Kyröläinen. 2016. Choosing between zero and pronominal subject: modeling subject expression in the 1st person singular in Finnish conversation. Corpus Linguistics and Linguistic Theory 12(2).

Laury, Ritva, Marja-Liisa Helasvuo & Janica Rauma. 2020. When an expression becomes fixed: mä ajattelin että ‘I thought that’ in spoken Finnish. In Ritva Laury & Tsuyoshi Ono (eds.), 133–166. Amsterdam: John Benjamins.

References

Leufkens, Sterre. 2020. A functionalist typology of redundancy. Revista da Abralin 19(3). 79–103. https://doi.org/10.25189/rabralin.v19i3.1722.

Leufkens, Sterre. 2022. Measuring redundancy: the relation between concord and complexity. Linguistics Vanguard 0(0). https://doi.org/10.1515/lingvan-2020-0143.

Levshina, Natalia. 2022. Communicative Efficiency: Language Structure and Use. Cambridge University Press.

Lindström, Liina, Mervi Kalmus, Anneliis Klaus, Liisi Bakhoff & Karl Pajusalu. 2009. Ainsuse 1. isikule viitamine eesti murretes. Emakeele Seltsi aastaraamat (54 (2008)). 159–185.

Lippus, Pärtel, Kätlin Aare, Anton Malmi, Tuuli Tuisk & Pire Teras. 2023. Phonetic Corpus of Estonian Spontaneous Speech v1.3. Institute of Estonian and General Linguistics, University of Tartu. https://doi.org/10.23673/RE-438.

Lison, Pierre & Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). 923–929.

Nagy, Naomi. 2015. A sociolinguistic view of null subjects and VOT in Toronto heritage languages. Lingua 164. 309–327.

Orozco, Rafael & Luz Marcela Hurtado. 2021. A Variationist Study of Subject Pronoun Expression in Medellín, Colombia. Languages 6(1). 5.

Posio, Pekka. 2014. Subject expression in grammaticalizing constructions: The case of creo and acho ‘I think’ in Spanish and Portuguese. Journal of Pragmatics 63. 5–18.

Savchuk, S., T. Arhangelskij, A. Bonch-Osmolovskaja, O. Donina, Ju. Kuznecova, O. Ljashevskaja, B. Orehov, M. Podrjadchikova. 2024. Nacionalnyj korpus russkogo jazyka 2.0: novye vozmozhnosti i perspektivy razvitija. Voprosy jazykoznanija 2. 7–34.

References

Shiffrin, Richard M. & Walter Schneider. 1977. Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychological Review 84(2). 127–190.

Torres Cacoullos, Rena & Catherine E. Travis. 2014. Prosody, priming and particular constructions: The patterning of English first-person singular subject expression in conversation. Journal of Pragmatics 63(3). 19–34.

Torres Cacoullos, Rena & Catherine E. Travis. 2016. Two languages, one effect: Structural priming in spontaneous code-switching. Bilingualism: Language and Cognition 19(4). 733–753.

Torres Cacoullos, Rena & Catherine E. Travis. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. Linguistics 57(3). 653–692.

Trudgill, Peter. 2011. Sociolinguistic typology: social determinants of linguistic complexity (Oxford Linguistics). Oxford; New York: Oxford University Press.

Väänänen, Milja. 2016. Subjektin ilmaiseminen yksikön ensimmäisessä persoonassa. Tutkimus suomen vanhoista murteista. Turku: Turun yliopisto.

Wagner, Susanne. 2018. Never saw one – first-person null subjects in spoken English. English Language and Linguistics 22(01). 1–34.

Wit, Ernst-Jan C. & Marie Gillette. 1999. What is Linguistic Redundancy? Technical Report. The University of Chicago. What is Linguistic Redundancy? (9 April, 2023).

Results of the log-linear models: Estonian

CLAUSE_TYPE:REFFORM_SAMESPEAKER + REFERENT_SWITCH:NUMBER:PERSON

REFFORM_SAMESPEAKER	no	**pronoun**	verb
	0.003	0.32	-0.32
CLAUSE_TYPE		**subord**	main
		0.22	-0.22
REFERENT_SWITCH		**switch**	same
		0.20	-0.20
NUMBER		**SG**	PL
		0.09	-0.09
PERSON		**1**	2
		0.05	-0.05

Results of the log-linear models: Estonian

CLAUSE_TYPE:REFFORM_SAMESPEAKER + REFERENT_SWITCH:NUMBER:PERSON

		REFFORM_SAMESPEAKER		
		no	pronoun	verb
CLAUSE_TYPE	main	0.15	-0.10	-0.05
	subord	-0.15	0.10	0.05

Results of the log-linear models: Estonian

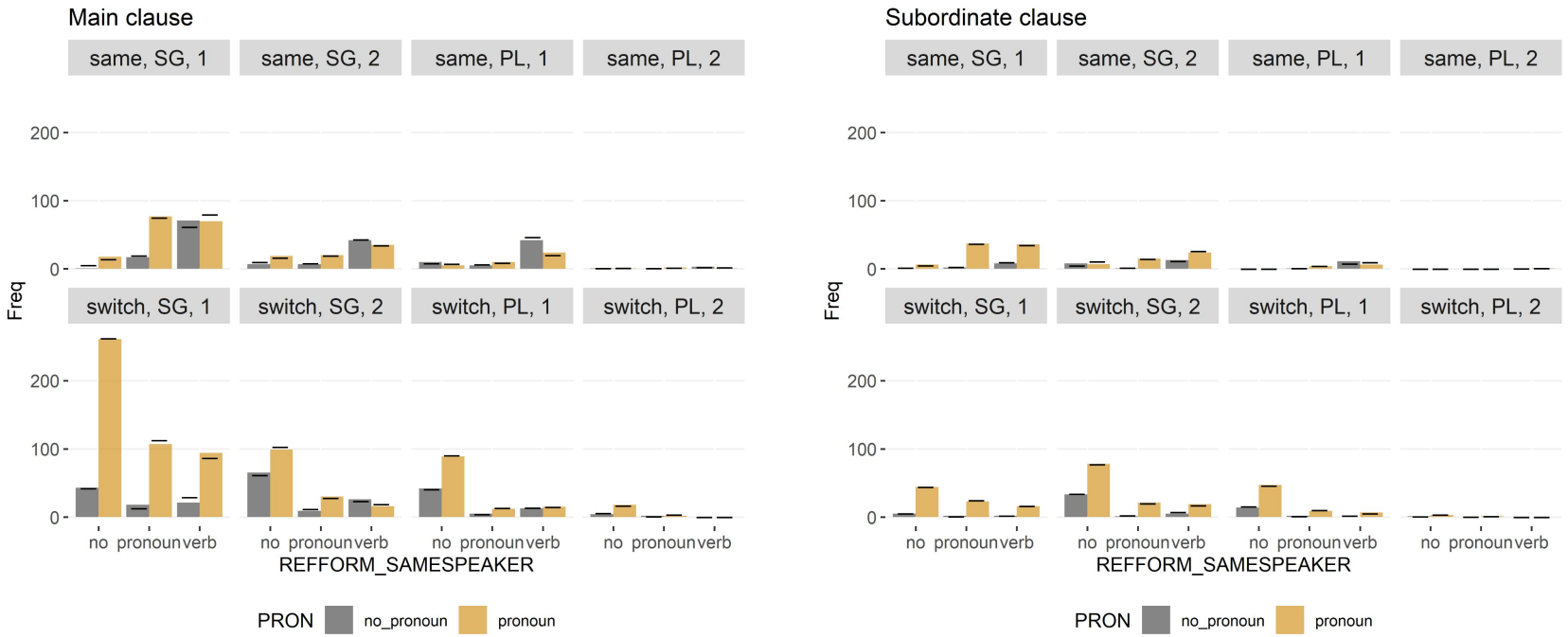
CLAUSE_TYPE:REFFORM_SAMESPEAKER + **REFERENT_SWITCH:NUMBER:PERSON**

		REFFORM_SAMESPEAKER		
		no	pronoun	verb
CLAUSE_TYPE	main	0.15	-0.10	-0.05
	subord	-0.15	0.10	0.05

		PERSON	REFERENT_SWITCH	
			same	switch
NUMBER	SG	1	-0.04	0.04
		2	0.04	-0.04
	PL	1	0.04	-0.04
		2	-0.04	0.04

Results of the log-linear models: Estonian

CLAUSE_TYPE:REFFORM_SAMESPEAKER + REFERENT_SWITCH:NUMBER:PERSON



Results of the log-linear models: Russian

TENSE + REFERENT_SWITCH + VP_COMPLEXITY:VERB_TYPE + REFFORM_SAMESPEAKER

TENSE	present	**past**	future
	-0.03	0.25	-0.22
REFFORM_SAMESPEAKER	no	**pronoun**	verb
	0.04	0.19	-0.23
VP_COMPLEXITY		**compl**	no_compl
		0.17	-0.17
REFERENT_SWITCH		**switch**	same
		0.16	-0.16
VERB_TYPE		**comm_ment_perc**	other
		0.07	-0.07

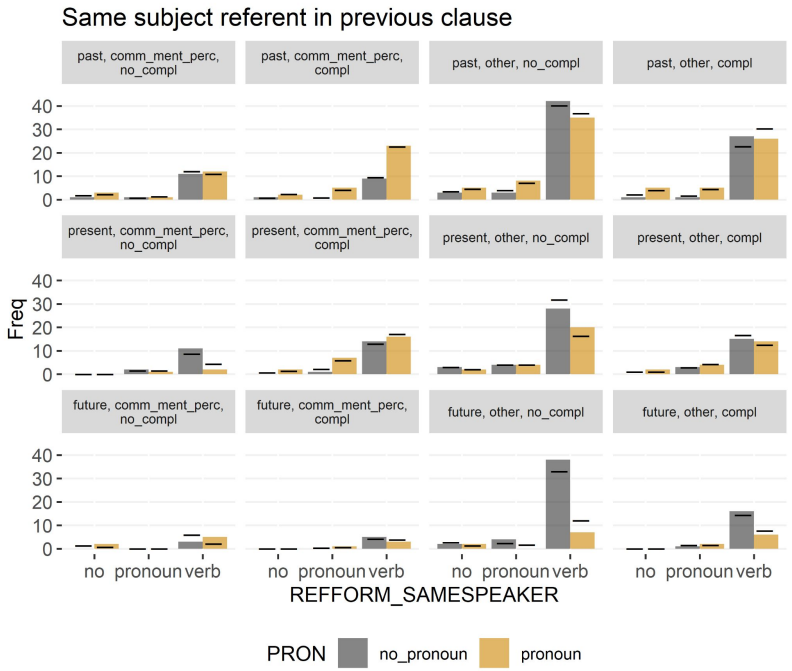
Results of the log-linear models: Russian

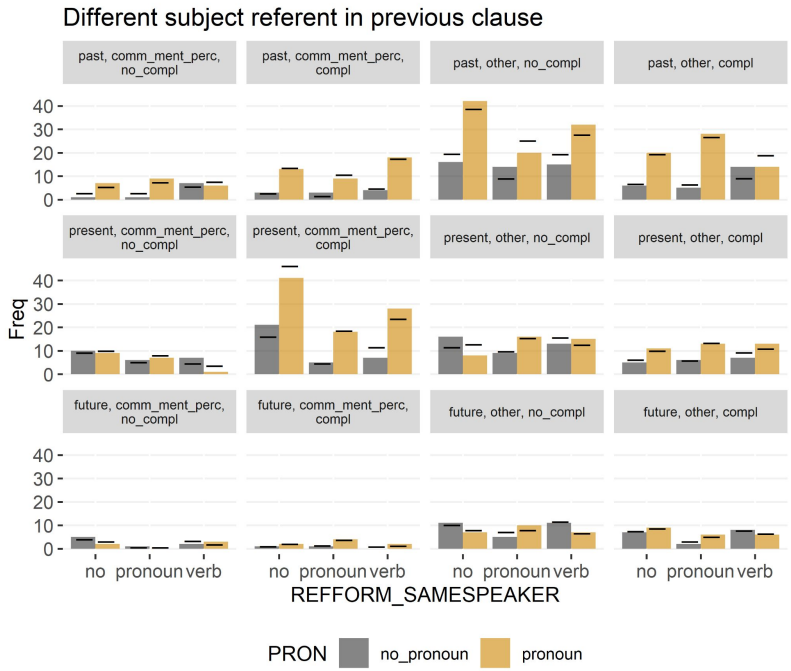
TENSE + REFERENT_SWITCH + **VP_COMPLEXITY:VERB_TYPE** + REFFORM_SAMESPEAKER

		VP_COMPLEXITY	
		compl	no_compl
VERB_TYPE	comm_ment_perc	0.07	-0.07
	other	-0.07	0.07

Results of the log-linear models: Russian

TENSE + REFERENT_SWITCH + VP_COMPLEXITY:VERB_TYPE + REFFORM_SAMESPEAKER

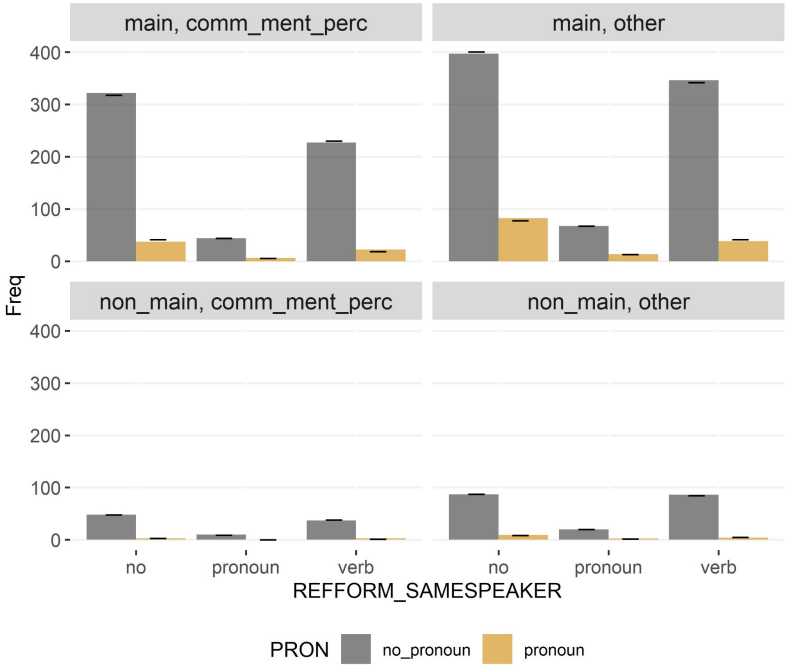




Results of the log-linear models: Finnish

REFFORM_SAMESPEAKER + VERB_COMPLEXITY + VERB_TYPE

VERB_COMPLEXITY		**main**	non-main
		0.17	-0.17
REFFORM_SAMESPEAKER	**no**	**pronoun**	verb
	0.08	0.07	-0.15
VERB_TYPE		**other**	comm_ment_perc
		0.10	-0.10



Results of the log-linear models: Polish

REFFORM_SAMESPEAKER + VERB_COMPLEXITY + CLAUSE_TYPE

VERB_COMPLEXITY		**main**	non-main
		0.39	-0.39
REFFORM_SAMESPEAKER	**no**	**pronoun**	verb
	0.12	0.19	-0.31
CLAUSE_TYPE		**main**	subord
		0.38	-0.38