The Empirical Basis of Slavic Intercomprehension

Tania Avgustinova, Andrea Fischer, Klara Jagrova, Dietrich Klakow, Roland Marti, Irina Stenger

REMU International Conference
28–29 May 2015, Joensuu, Finland

Background (e.g. Czech and Polish)

"The basic mission/task	Základním posláním	Podstawowym zadaniem
of the Czech-Polish Forum	Česko-polského fóra	Forum Polsko-Czeskiego
is to support	je podpora rozvoje	jest wspieranie działalności
both current and	stávajících a vzniku	istniejących oraz powstania,
new common initiatives	nových společných iniciativ	nowych, wspólnych inicjatyw
within the civil societies	nevládních subjektů	wśród społeczeństw obywatelskich
of both countries."	obou zemí.	obydwu państw.

fully understandable *still intelligible* *unintelligible*

- Well-known factors determining similarity of written texts in closely related languages:
  - Orthographic distance (orthographic correspondences in cognate sets)
  - Morphological distance (similarity of forms; correspondences in grammar)
  - Lexical distance (cognates: positive, partial, negative; similarity of closed word classes)
  - Syntactic distance (aggregate linguistic measure: linear order, complexity of constructions)

Approaching intercomprehension

… as processing "noisy code" (→ an information-theoretic view)

Consider a blended text sample constructed by using information chunks in Czech and Polish interchangeably:

Základním posláním Forum Polsko-Czeskiego je podpora rozvoje istniejących oraz powstania nových společných iniciativ wśród społeczeństw obywatelskich obou zemí.

"The basic mission/task of the Czech-Polish Forum is to support both current and new common initiatives within the civil societies of both countries."

It is expected to be intelligible to speakers of these languages, without conforming to the respective encoding systems.

A newly established interdisciplinary Collaborative Research Centre

Language Use

Languages offers a wide range of options of how to encode a message.

Linguistic Variation

Variation is an inherent property of the linguistic system.

- **Central hypothesis**
  - Language processing relies on predictability in context (in a broader sense)
  - Contextually determined predictability is appropriately indexed by Shanon's notion of information

Information Density (Surprisal)

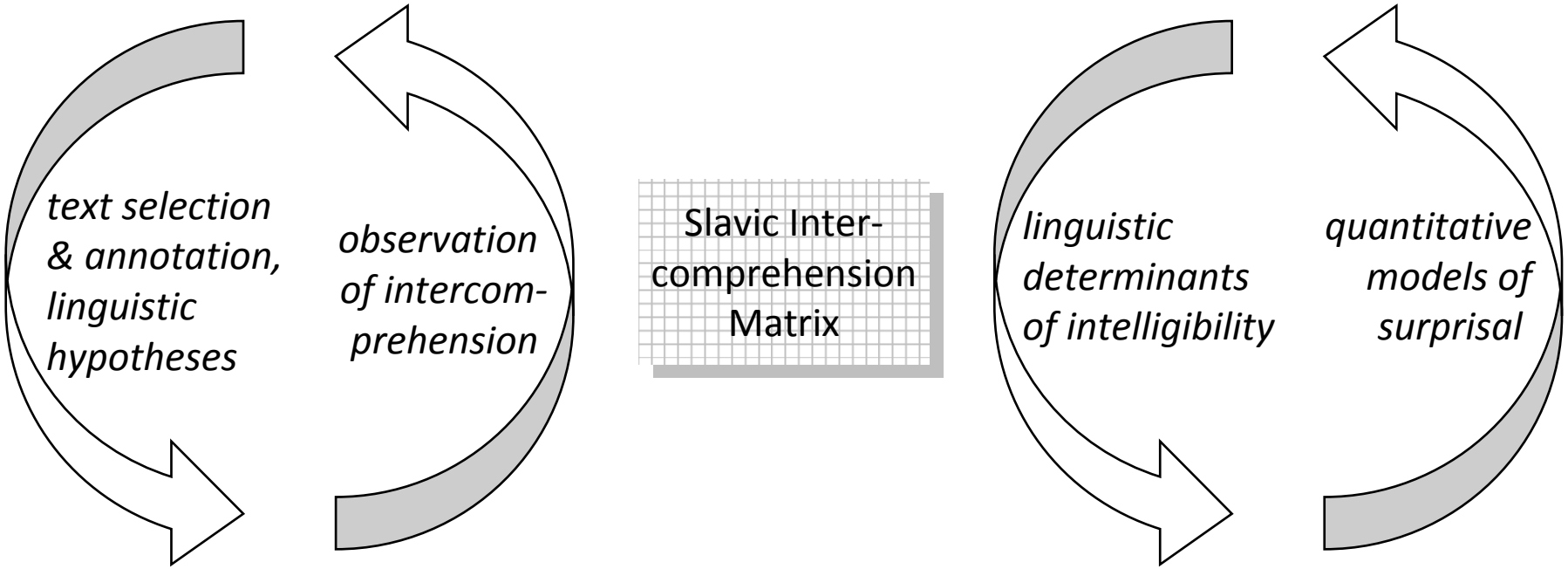
$$Surprisal\,(unit) = \log_2 \frac{1}{P(unit \mid \text{Context})} = -\log_2 P(unit \mid \text{Context})$$

- **Long-term research programme:** information theory for linguistic inquiry
  - Project: Mutual intelligibility and surprisal in Slavic intercomprehension (INCOMSLAV)

Research rationale

- The **reading intercomprehension** scenario reveals
  - inter-lingual tolerance to unfamiliar linguistic encoding
  - asymmetries with regard to intelligibility (depending on the language pair)
- **Goal**: identify mechanisms by which languages encode and decode information
  - (the degree of) similarity between Slavic languages provides the basis for (varying) expectations about the linguistic encoding
  - find statistical evidence of mutual intelligibility
- With **meaningful units of language** we expect
  - diminished intelligibility through missing units
  - confusion through misrecognition of units
- **General idea**: surprisal of language models correlates with intelligibility
  - adapt N-gram LMs for cross-language use via latent space and similarity
  - analyse information-theoretical results with linguistic knowledge

Encoding; linguistic phenomena; meaningful units of language; intelligible information chunks (cognates, paraphrases, fragments); shared grammar



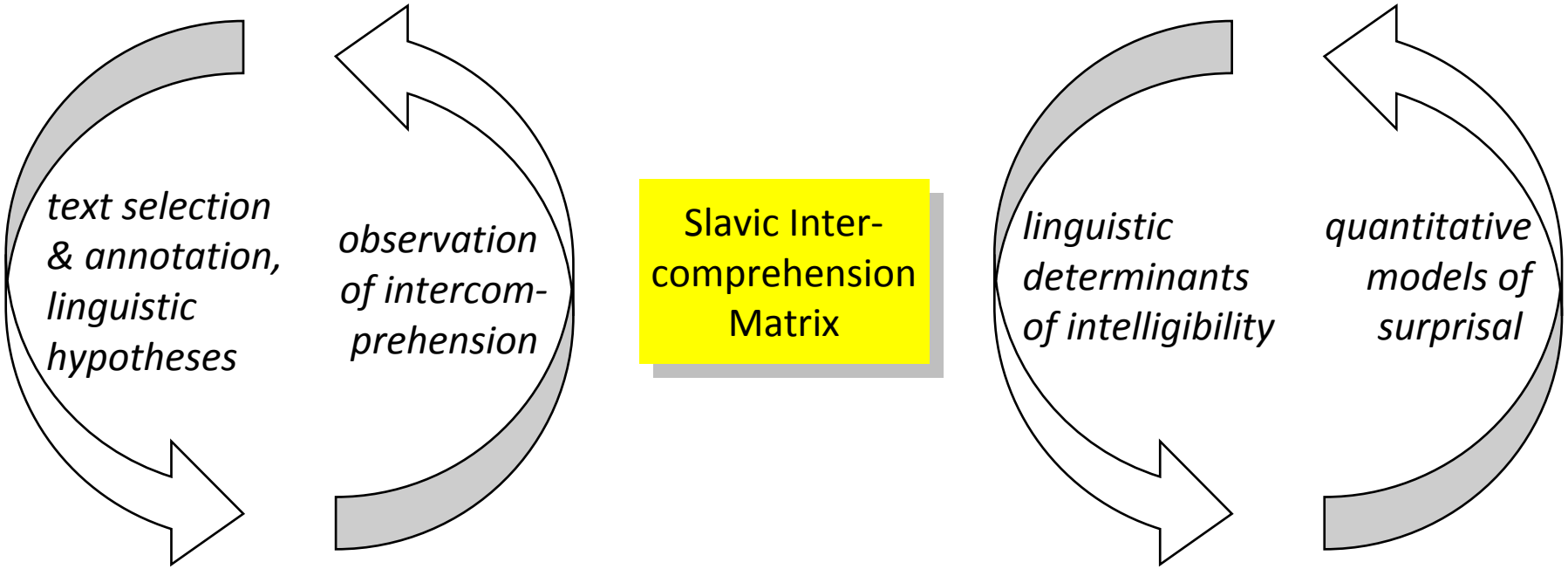
Experiments: variably close language pairs; synchronic and diachronic perspective

surprisal measure

intelligibility validation

Modelling: linguistic and statistical models of surprisal; large-scale corpus studies

Encoding; linguistic phenomena; meaningful units of language; intelligible information chunks (cognates, paraphrases, fragments); shared grammar



Experiments: variably close language pairs; synchronic and diachronic perspective

surprisal measure

intelligibility validation

Modelling: linguistic and statistical models of surprisal; large-scale corpus studies

Slavic intercomprehension matrix

SUB-GROUPS	East Slavic			West Slavic					West South Slavic				East South Slavic	
	Russ	Ruth		Sorb		Lech	Cz-Slk		SCB			Slv		
ISO-code	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**	**12.**	**13.**	**14.**
1. Russian	rus	1(2)	1(3)											
2. Ukrainian	2(1)	ukr	2(3)											
3. Belorusian	3(1)	3(2)	bel											
4. Upper Sorbian				hsb	4(5)	4(6)	4(7)	4(8)						
5. Lower Sorbian				5(4)	dsb	5(6)	5(7)	5(8)						
6. Polish				6(4)	6(5)	pol	6(7)	6(8)						
7. Czech				7(4)	7(5)	7(6)	ces	7(8)						
8. Slovak				8(4)	8(5)	8(6)	8(7)	slk						
9. Bosnian									bos	9(10)	9(11)	9(12)		
1o. Croatian									10(9)	hrv	10(11)	10(12)		
11. Serbian									11(9)	11(10)	srp	11(12)		
12. Slovene									12(9)	12(10)	12(11)	slv		
13. Macedonian													mkd	13(14)
14. Bulgarian													14(13)	bul

Slavic intercomprehension matrix

SUB-GROUPS	East Slavic			West Slavic					West South Slavic				East South Slavic	
	Russ	Ruth		Sorb		Lech	Cz-Slk		SCB			Slv		
ISO-code	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**	**12.**	**13.**	**14.**
1. Russian	rus	1(2)	1(3)											1(14)
2. Ukrainian	2(1)	ukr	2(3)											
3. Belorusian	3(1)	3(2)	bel											
4. Upper Sorbian				hsb	4(5)	4(6)	4(7)							
5. Lower Sorbian				5(4)	dsb	5(6)	5(7)	5(8)						
6. Polish				6(4)	6(5)	pol	6(7)	6(8)						
7. Czech				7(4)		7(6)	ces	7(8)						
8. Slovak				8(4)	8(5)	8(6)	8(7)	slk						
9. Bosnian									bos	9(10)	9(11)	9(		
1o. Croatian									10(9)	hrv	10(11)	10(12)		
11. Serbian									11(9)	11(10)	srp	11(12)		
12. Slovene									(9)	12(10)	12(11)	slv		
13. Macedonian													mkd	13(14)
14. Bulgarian	14(1)												14(13)	bul

Czech through Polish

How can a Russian understand Bulgarian?

Polish through Czech

Serbian Croatian

How can a Bulgarian understand Russian?

Croatian Serbian

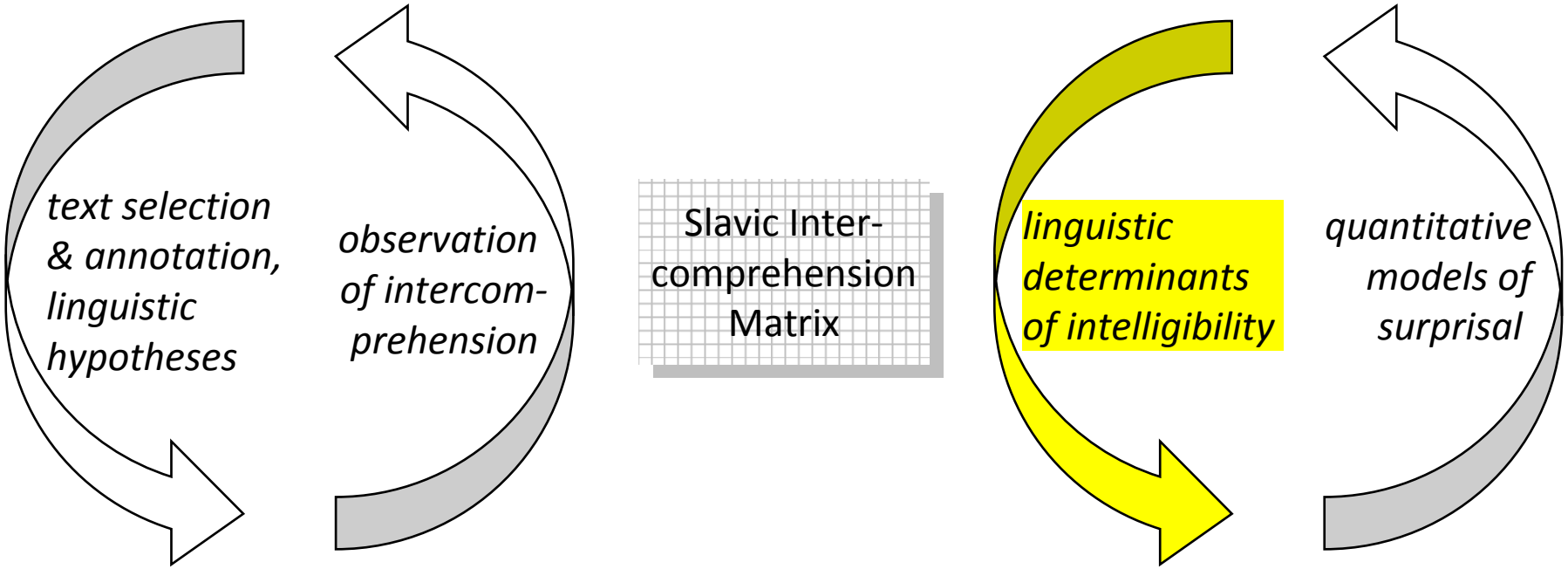
Slavic intercomprehension matrix

1+6+14 (7)

SUB-GROUPS	East Slavic			West Slavic					South Slavic				East South Slavic	
	Russ	Ruth		Sorb		Lech	Cz-Slk							
ISO-code	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
1. Russian	rus	1(2)	1(3)				1(7)							
2. Ukrainian	2(1)	ukr	2(3)											
3. Belorusian	3(1)	3(2)	bel											
4. Upper Sorbian				hsb	4(5)	4(6)	4(7)	4(8)						
5. Lower Sorbian				5(4)	dsb	5(6)	5(7)	5(8)						
6. Polish				6(4)	6(5)	pol	6(7)	6(8)						
7. Czech				7(4)	7(5)	7(6)	ces	7(8)						
8. Slovak				8(4)	8(5)	8(6)	8(7)	slk						
9. Bosnian									bos	9(10)	9(11)	9(12)		
1o. Croatian									10(9)	hrv	10(11)	10(12)		
11. Serbian									11(9)	11(10)	srp	11(12)		
12. Slovene									12(9)	12(10)	12(11)	slv		
13. Macedonian													mkd	13(14)
14. Bulgarian							14(7)						14(13)	bul

Processing Czech, based on knowledge of Russian, Polish and Bulgarian

Encoding; linguistic phenomena; meaningful units of language; intelligible information chunks (cognates, paraphrases, fragments); shared grammar



Experiments: variably close language pairs; synchronic and diachronic perspective

surprisal measure

intelligibility validation

Modelling: linguistic and statistical models of surprisal; large-scale corpus studies

Work in progress and first results

- Investigating the use of Levenshtein distance
  - for projecting the units of a source language into the vocabulary of a target language
- Modelling varying levels of linguistic knowledge of a hypothetical reader
  - via different transformation costs (e.g. Czech-Polish *v=w* for zero cost).
- Assessing the projected unit representations using a language model
  - which allows us to identify the most informative features
  - and to estimate their impact on overall surprisal.
- Each individual word is an agglomerate of meaningful units:
  - list of features, with each feature contributing individually to the word's identity

→ Technical details in the poster session!

Empirical basis for measuring orthographic distance

- **Levenshtein algorithm** for calculating basic differences

a. Czech p l n ý
Polish p e ł n y
(0 + 1 + 0.5 + 0 + 0.5) / 5 → 40%

b. Czech m o ř e
Polish m o r z e
(0 + 0 + 0.5 + 1 + 0) / 5 → 30%

c. Czech t ě l o
Polish c i a ł o
(1 + 1 + 1 + 0.5 + 0) / 5 → 70%

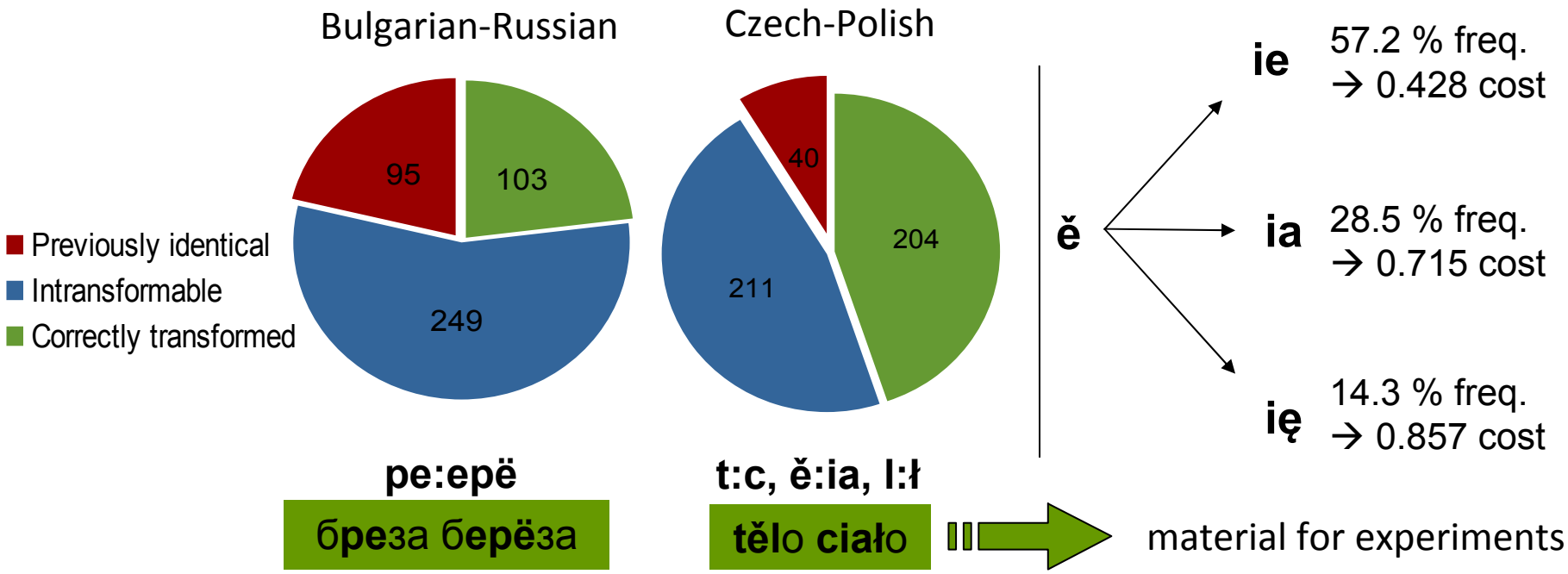
Empirical basis for measuring orthographic distance

- **Levenshtein algorithm** with awareness of regular orthographic correspondences, including diachronic motivation

	Czech	Polish	Croatian	Bulgarian	Russian	Ukrainian	Belorussian
'horse'	kůň	koń	konj	кон	конь	кінь	конь
'body'	tělo	ciało	tijelo	тяло	тело	тіло	цела
'sea'	moře	morze	more	море	море	море	мора
'brush'	štětka	szczotka	četka	четка	щётка	щітка	шчотка
'head'	hlava	głowa	glava	глава	голова	голова	галава
'cow'	kráva	krowa	krava	крава	корова	корова	карова
'before'	před	przed	pred	пред	перед	перед	перад
'voice'	hlas	głos	glas	глас	голос	голос	голас
'long'	dlouhý	długi	dug	дълъг	долгий	довгий	доўгі
'mill'	mlýn	młyn	mlin	мелница	мельница	млин	млын
'full'	plný	pełny	pun	пълен	полный	повний	поўны
'yellow'	žlutý	żołty	žut	жълт	жёлтый	жовтий	жоўты
'wolf'	vlk	wilk	vuk	вълк	волк	вовк	воўк

Empirical basis for measuring orthographic distance

- Automatic application of **diachronically based orthographic transformation rules** between language pairs on **cognate** sets
  - **ranking** of correspondence rules according to their frequency
  - deriving a **weighted Levenshtein distance**: cost = 1 - transformation frequency]
- Example*:* Pan-Slavic vocabulary



Approaching mutual intelligibility of inflectional morphology

- Noun declension (e.g.'winter')

	Czech	Polish	Bulgarian	Russian
N	zima	zima	*зима**	зима
G	zimy	zimy	-	зимы
D	zimě	zimie	-	зиме
A	zimu	zimę	-	зиму
I	zimou	zimą	-	зимой
L	zimě	zimie	-	зиме
V	zimo!	zimo!	*зимо!*	-

- Present tense conjugation (e.g. 'write')

	Czech	Polish	Bulgarian	Russian
1sg	píšu / píši	piszę	пиша	пишу
2sg	píšeš	piszesz	пишеш	пишешь
3sg	píše	pisze	пише	пишет
1pl	píšeme	piszemy	пишем	пишем
2pl	píšete	piszecie	пишете	пишете
3pl	píšou / píši	piszą	пишат	пишут

- Similarity of morphosyntactic forms
  - How have grammatical elements developed in the individual languages?
  - Parallel lists of prefixes and suffixes allow for working out the meaning of complex words by separating affixed elements from roots.
  - Application of morphology processing tools, e.g. Morfessor

Accounting for mutual intelligibility of lexis

- Availability of lexical alternatives leads to asymmetric intelligibility
- Look into cognates: positive (vs. non-cognates), partial (?), negative ("false friends")
- Word sets to use:
  - international and common Slavic vocabulary,
  - closed classes (numerals, prepositions, conjunctions, function words, etc.),
  - named entities, …
- Goal: measuring linguistic distance based on, e.g.
  - the percentage of cognate words (vs. non-cognate words)
  - the degree of lexical relatedness (are cognates easily discernible as related words?)
  - the degree of semantic relatedness (do cognates mean roughly similar things?)

Estimating mutual intelligibility in syntax

- Communicatively determined linearisation on clausal level vs. differences in sub-clausal domain (e.g. NP)

	Subject	Verb	Object
Czech	Student	píše	dopis.
Polish	Student	pisze	list.
Bulgarian	Студент	пише	писмо.
Russian	Студент	пишет	письмо.

Czech	Student	čte	knihu.
Polish	Student	czyta	książkę.
Bulgarian	Студент	чете	книга.
Russian	Студент	читает	книгу.

	nouns & adjectives
Czech	komplikovaný polský **jazyk**
Polish	skomplikowany **język** polski
Czech	Varšavská **univerzita**
Polish	**Uniwersytet** Warszawski
Czech	současné polské **malířství**
Polish	współczesne **malarstwo** polskie
Czech	botanická **zahrada**
Polish	**ogród** botaniczny
Czech	dramatické **divadlo**
Polish	**teatr** dramatyczny

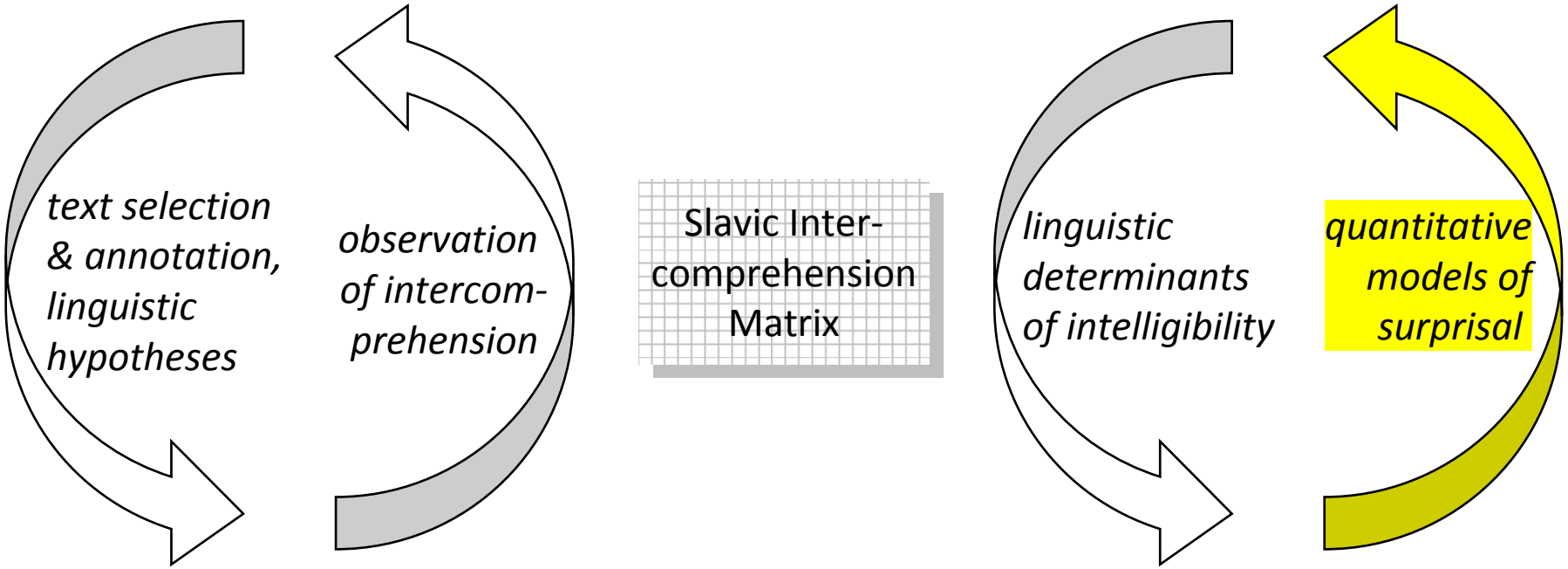
Estimating mutual intelligibility in syntax

- Observed parallels

  w.r.t. diathesis alternations, nominalisations, relatives, conditionals, interrogatives, coordination, apposition etc.

- Syntactic measures have to consider
  - **sentence length**, as longer sentences are on average more likely to consist of more complex syntactic structures than short sentences
  - **type of constituents**, e.g. the mean number of clauses per sentence, dependent clauses per clause, coordinate phrases per clause, complex nominals per clause, modifications to a word, etc.
  - **positional correspondences** in word order variation and collocations can be measured using statistical machine translation models and in particular by analysing the alignment models.

Encoding; linguistic phenomena; meaningful units of language; intelligible information chunks (cognates, paraphrases, fragments); shared grammar



Experiments: variably close language pairs; synchronic and diachronic perspective

surprisal measure

intelligibility validation

Modelling: linguistic and statistical models of surprisal; large-scale corpus studies

Quantitative models of surprisal (e.g. Polish through Czech)

- Surprisal (or "informativeness" of an item)
  - **The model predicting Polish words in Polish context** $P(w^P|h^P)$ measures the surprisal of a Polish item w^P, given a Polish history of preceding items h^P.
  - **The model predicting Czech words in Czech context** $P(w^C|h^C)$ measures the surprisal of a Cuech item w^C, given a Czech history of preceding items h^C.

- We want to derive
  - a model that allows us to estimate $P(w^P|h^P)$ given $P(w^C|h^C)$, i.e.
  - **what expectations a Czech reader might have being exposed to a Polish text**.

- To do this, two additional model components are needed:
  - $P(h^P|h^C)$ mapping **from the Polish history to the Czech history**
  - $P(w^P|w^C)$ mapping **the predicted Czech word to the predicted Polish word**

Quantitative models of surprisal (e.g. Polish through Czech)

- In general there is some **uncertainty about the word to word correspondence**.
- We have two possibilities to account for that.

1. In the first one we are summing over all possible alternatives:

$$P(w^P \mid h^P) = \sum_{w^C} \sum_{h^C} P(w^P \mid w^C) P(w^C \mid h^C) P(h^P \mid h^C)$$

2. In the second one we assume that the knowledge about the correspondence of word and context is very close to certainty.

$$P(w^P \mid h^P) \approx \operatorname{argmax}_{w^C} \operatorname{argmax}_{h^C} P(w^P \mid w^C) P(w^C \mid h^C) P(h^P \mid h^C)$$

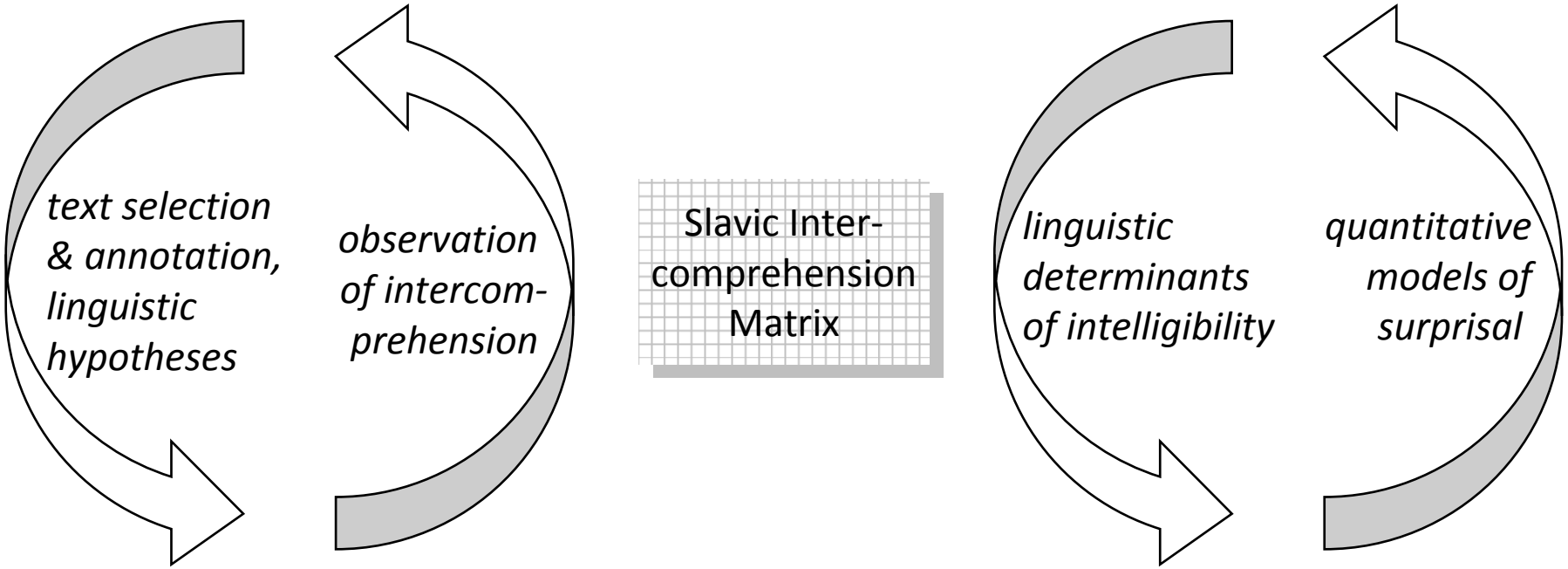
This could be when the correspondence is obvious, e.g. due to the closeness of the languages (i.e. the Czech speaker will make a hard pick).

Summary

- Reading intercomprehension
  - approached as **adaptation between statistical language models**
  - use of **techniques from machine translation**
  - **no extra-linguistic information** used in modeling (yet)

- Current status:
  - primary focus on Czech, Polish, Russian and Bulgarian
  - analyzing the orthographic level
  - reviewing the (historically developed) orthographic correspondences
  - assessing the extent to which these correspondences are attested in large parallel corpora, and whether the data point to further correspondences

Encoding; linguistic phenomena; meaningful units of language; intelligible information chunks (cognates, paraphrases, fragments); shared grammar



Experiments: variably close language pairs; synchronic and diachronic perspective

surprisal measure

intelligibility validation

Modelling: linguistic and statistical models of surprisal; large-scale corpus studies

Important related work

- EuroComSlav: The Seven Sieves

  (1) International vocabulary; (2) Pan-Slavic vocabulary;
  (3) Sound correspondences; (4) Spelling and pronunciation;
  (5) Pan-Slavic syntactic structures; (6) Morphosyntactic elements;
  (7) Prefixes and suffixes

  - All these resources are systematically (re-)considered in our work

- MICReLa: Mutual intelligibility of closely related language in Europe: linguistic and non-linguistic determinants
  - data collected from web experiments
  - possible extensions of the on-line system (?)
  - theoretical findings and models of intercomprehension

- EuroMatrixPlus (http://www.euromatrixplus.net/matrix/)
  - Language Technology aspects

PhD research in INCOMSLAV

- Scientific context
  - Developing a surprisal-based model of intercomprehension combining large-scale corpus studies and psycholinguistic experimental work.
  - Establishing a Slavic intercomprehension matrix

- PhD Projects

	Working Title	Supervisors
1. Irina Stenger	On the role of orthography in Slavic intercomprehension with special attention to the Cyrillic script	Avgustinova Marti
2. Klara Jagrova	Linguistic determinants of successful intercomprehension in Slavic languages	Avgustinova Marti
3. Andrea Fischer	Differences in information en- and decoding between Slavic languages	Klakow Avgustinova