Формални аспекти на взаимната разбираемост между славянските езици

Таня Августинова
avgustinova@coli.uni-saarland.de

Международна научна конференция
Филологическият проект – кризи и перспективи
СУ „Св. Климент Охридски“
Факултет по славянски филологии
24- 25 април 2015 г.

Информационно-теоретична парадигма в лингвистиката

- Езикът предоставя не само необходимата за общуване функционална експресивност, но предлага и разнообразие от варианти при избора на изразните средства за кодиране на дадено информационно съобщение — от подбор на думите и структуриране на синтактичните елементи до аранжиране на изреченията в дискурс.
- Очевидно е, че лекотата, с която се възприема езиковият материал, зависи от неговата **предвидимост** в даден контекст, докато информативността на лигвистичното кодиране корелира с понятието **surprisal** в смисъл на "неочакваност, непредвидимост, изненада", заимствано от теорията на информацията (Shannon 1948).
- Пример: Еднакво предсказуема ли е фразата "пощенска марка" в следните контексти?

1. Оказа се, че Ани е изпратила писмото без _____
2. Оказа се, че Ани е ходила до магазина да купи _____

→ Продължението в (2) е по-непредвидимо, следователно по-информативно.

Информационно-теоретична парадигма в лингвистиката

- Вариантността в езика както и самата езикова употреба могат да бъдат по-добре разбрани, ако се съпоставят с целта на говорещия да модулира количеството информация, предавано чрез определено изказване.
- В съответствие с комуникативната ситуация оптимално разпределение на информацията се постига с помощта на различни изразни средства: фонетични редукции, елиптични конструкции, подчинени изречения, обособяване, прономинализация, словоред, фрагменти и т.н.
- Няколко примера, касаещи една и съща ситуация:

(3) Ще каним ли компанията, с която бяхме на екскурзионно летуване?

(4) Ще каним ли компанията от екскурзионното летуване?

(5) Компанията от екскурзионното ще я каним ли?

(6) От екскурзионното ще ги каним ли?

(7) Ами тия от екскурзионното летуване?

Моделиране на взаимната разбираемост на сродни езици

- Доколко разбираемо е дадено информационно съобщение, формулирано на сроден език, зависи от лингвистични и екстралингвистични фактори.
- Обективните лингвистични детерминанти на разбираемостта включват всички нива на езиковата система.
- Когато носителят на L1 възприема езиков материал на непознат за него, но типологически близък език L2, той може да го интерпретира по отношение на собствения си език като сигнал със смущения (noisy input).
  - При сродни езици граматичното знание за L1 поражда определени очаквания относно формата и структура на съобщението, формулирано на L2 и осъзнаването на възможните интерференциите между L1 и L2 оптимизира този процес
  - Подобен сценарий на възприемане на информацията разкрива своеобразен ефект на междуезикова толерантност към непознатото лингвистично кодиране и същевременно предполага асиметрична разбираемост в зависимост от конкретната езикова двойка.

"Сигнал със смущения": конструиран полско-чешки текст

- Примерът съдържа редуващи се фрагменти от двата езика:

  Základním zadaniem Česko-polského fóra jest podpora działalności stávajících oraz vzniku nowych, wspólnych inicjatyw nevládních subjektů obydwu zemí. Forum navazuje do współpracy nezávislých skupin działających před rokem 1989, której vyvrcholením była činnost Solidarności Polsko-Czesko-Słowackiej.

  (Изходните едноезичните текстове са представени подравнено на следващия слайд.)

- Този езиков материал е частично разбираем за носителите на всеки един от езиците, без обаче да съответства нито на полската, нито на чешката кодираща езикова система.

- Какво е разпознаваемо например за един чех четящ полски текст?
  - Как ортографията затруднява възприемането на съобщението?
  - Каква е ролята на "верните" и "лъжливите" приятели в лексиката?
  - Доколко транспарентно е морфологичното маркиране на падежите?
  - Влияят ли на скоростта на четене различните словоредни варианти? ...

http://www.mzv.cz/cesko-polske_forum/

чешки		полски
Základním posláním		Podstawowym zadaniem
Česko-polského fóra		Forum Polsko-Czeskiego
je podpora rozvoje		jest wspieranie działalności
stávajících a vzniku		istniejących oraz powstania
nových společných iniciativ		nowych, wspólnych inicjatyw
nevládních subjektů		wśród społeczeństw obywatelskich
obou zemí.		obydwu państw.
Fórum navazuje na spolupráci		Forum nawiązuje do współpracy
nezávislých skupin v období nesvobody		niezależnych grup opozycyjnych, działających
před rokem 1989,		przed 1989 rokiem,
jejímž vyvrcholením byla činnost		której ukoronowaniem była działalność
Polsko-Česko-Slovenské Solidavrity.		Solidarności Polsko-Czesko-Słowackiej.

Емпиричната основа на разбираемостта при сродни езици

- Степента на близост на два славянски езика поражда при четене различни очаквания относно лингвистичното кодиране.
  - Между чешки и български, например, очакванията са по-малко, предвидимостта е по-ниска отколкото между чешки и полски.
  - Фактът, че при падежното маркиране в чешкия има по-голям синкретизъм, може да доведе до асиметрична разбираемост.
- В следната таблица (**Slavic Intercomprehension Matrix**) е направен опит за визуализция на възможните езикови комбинмации.
  - Цветовият код отразява традиционно приетото групиране на славянските езици.
  - Различните оттенъци илюстрират общоприетата степен на близост в рамките на южно-славянската, източно-славянската и западно-славянската група.
  - Номерацията служи за индексиране по схемата **L1(L2)**, при което L1 отговаря на изходния език, а L2 на възприемания.
  - При подобен експериментален дизайн е напълно възможно да се моделират и интерференцията при няколко изходни езика.

Slavic intercomprehension matrix

L2→ / ↓L1	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)		
10. 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

L2→ / ↓L1	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(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(		
10. 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

Croatian view of Serbian

How can a Bulgarian understand Russian?

Serbian view of Croatian

Slavic intercomprehension matrix

L2→ / ↓L1	East Slavic			West Slavic					West South Slavic				East South Slavic	
	Russ	Ruth		Sorb		Lech	Cz-Slk		B			Slv		
ISO-code	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.			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)		
10. 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

1*6*14 (7)

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

Как да се моделира разбираемостта при сродни езици?

- Рецептивното многоезичие (receptive multilingualism) при сродни езици е специфична форма на езикова употреба. — В известен смисъл тук става дума за "положителната страна" на междуезиковата интерференция.
- Феноменът на междуезиково разбиране (intercomprehension) е интуитивно понятен: носителят на езика L1 е в състояние в една или друга степен да разбира езика L2 без да го владее, осланяйки се предимно на езикова система на L1.
- В този сценарий трябва да се отчита и наблюдаваната асиметричната разбираемост (asymmetric intelligibility) при определени двойки близко-родствени езици, която се обуславя не само от обективни типологически характеристики, но и от извън-езикови социолингвистични фактори.
- В случай на близко-родствени езици L1 и L2 хипотезата е, че взаимната разбираемост съответства на лингвистичната отдалеченост (linguistic distance) между разглежданите езици, която се измерва на различни нива: фонетика, ортография, лексика, морфология, синтаксис и т.н.

Привличане на класическа славистична експертиза

- В настоящата студия се разработва тезата, че още с първите си стъпки компютърното лингвистично моделиране на обстоятелство, че славянските езици са в различна степен взаимно разбираеми (mutually intelligible), предполага систематично и целенасочено привличане на класическа славистична експертиза в областта както на сравнително-историческата граматика, така и на традиционното съпоставително описание на различните езикови двойки.
- В сравнително-исторически план общославянската основа дава възможност за формулиране на абстрактна репрезентация на фонетични, морфосинтактични и лексикални явления, обуславящи езиковата близост.
- В същото време съпоставително-синхронната перспектива ни предоставя конкретна база за установяване на лингвистичната дистанция между всеки два славянски езика на графемично, граматично и лексикално ниво.

Пример 1: чисто формална близост на ниво ортография

- Ако се приложи директно алгоритмът на Levenshtein за определяне на лингвистична дистанция, изчисляващ цената за вмъкване, изтриване или замяна на символ, получаваме следните различия в проценти:

a. чешки **p l n ý**
полски **p e ł n y**
(цена) **(0 + 1 + 0.5 + 0 + 0.5) / 5** → 40%

b. чешки **m o ř e**
полски **m o r z e**
(цена) **(0 + 0 + 0.5 + 1 + 0) / 5** → 30%

c. чешки **t ě l o**
полски **c i a ł o**
(цена) **(1 + 1 + 1 + 0.5 + 0) / 5** → 70%

Пример 2: славистична близост на ниво ортография

- Ако обаче привлечем **славистична експертиза** при определянето на съответствията, бихме получили значително по-интересен резултат:

	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

- На тази база **директните еквивалентите** (маркирани цветово) може да обединят няколко трансформации, включително транслитерация.

Лингвистично моделиране

- Отдавайки дължимото на катализираща роля на езиковите технологии в съвременните славистични проучвания (корпусни методи, статистическо моделиране, психолингвистични експерименти), настоящото изследване прави опит за осмисляне как от своя страна славянското езикознание би обогатило компютърно-лингвистичното моделиране.
- Една от основните цели на представения тук проект е да верифицира наложилите се в славянската филология схващания за групирането по близост на славянските езици като разработи обективна емпирична база за типологичната им класификация.

Общата рамка

- Научно-изследователски център (2014-2026)
  Information Density and Linguistic Encoding
  http://www.sfb1102.uni-saarland.de/

- Проект INCOMSLAV (2014-2018)
  Mutual intelligibility and surprisal in Slavic intercomprehension
  http://www.coli.uni-saarland.de/~tania/incomslav.html

  - **Principal investigators:** Prof. Dr. Tania Avgustinova
    Prof. Dr. Roland Marti
    Prof. Dr. Dietrich Klakow

  - **Doctoral students:** Andrea Fischer
    Klára Jágrová
    Irina Stenger