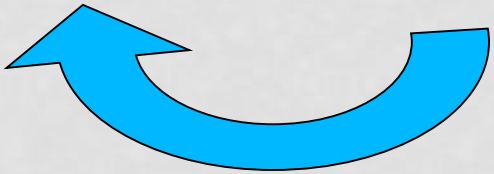
Кореферентные отношения в тексте. Сравнительный анализ размеченных данных

Анна Недолужко
Карлов Университет, Прага

РЕФЕРЕНЦИЯ - КОРЕФЕРЕНЦИЯ

Лена спросила папу, где он был.



ЧТО МЫ РАЗМЕЧАЕМ

- Грамматическая кореференция:

человек, который пьет

- Прономинальная и именная кореференция:

Лена спросила папу, где он был. Папа промолчал.

- Ассоциативная анафора (bridging)

Лена вошла в дом. С потолка капала вода.

ТИПОЛОГИЯ ПРОНОМИНАЛЬНОЙ И ИМЕННОЙ КОРЕФЕРЕНЦИИ

отношения между конкретнореферентными и родовыми ИГ

- **тип 0** (конкретнореферентные ИГ)

напр.: *Елена — она — девушка — Ø — дочь*

- **тип GEN** (родовые ИГ)

напр. *Женщины часто боятся, что их обманут. […] На самом деле женщины просто не понимают своего счастья.*

+ почти все абстрактные имена

ТИПОЛОГИЯ ОТНОШЕНИЙ-BRIDGING

- **часть - целое**

напр.: *Бавария — Германия*

- **множество – подмножество/элемент множества**

напр.: *студенты – три студента*

- **объект – его функция/позиция**

напр.: *школа — учитель*

- **эксплицитная анафора без кореференции**

напр.: *учителя — такие же учителя*

- **отношение дискурсивного контраста**

напр.: *Люди не жуют, жуют только коровы.*

- **остальное**

напр.: *Германия – немец, дед – внук, спор – спорщик* и т.д.

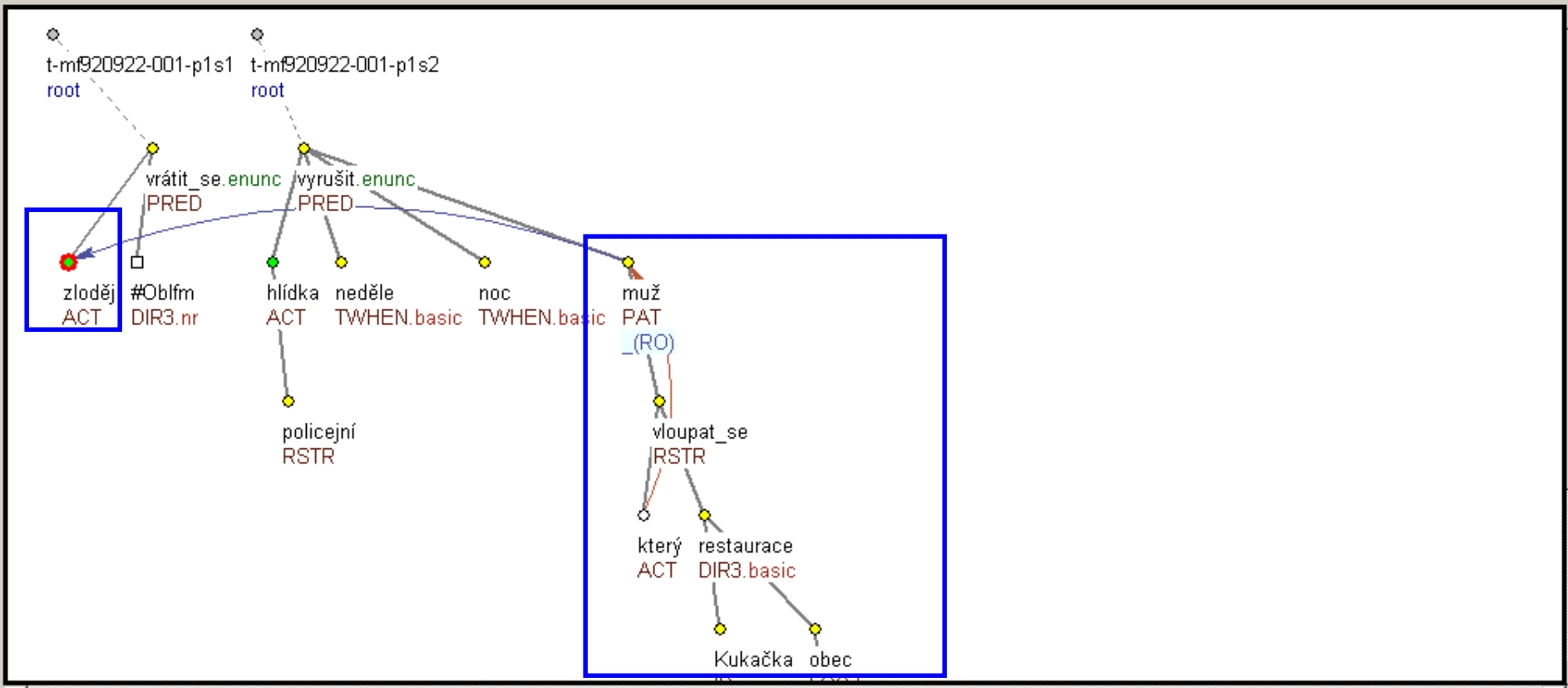
TRee EDitor Default(4/4): D:\1_UFAL\shoda\20091223\joint_JP_RO\mf920922_001.t.gz

File Node Tree View Macros Setup Help Mode: PML_T_Bridging

Style: PML_T_Bridging

--> ZLODĚJ SE VRÁTIL.
Policejní hlídka vyrušila v neděli v noci muže, který se vloupal do restaurace Kukačka v obci Horní Životice.
Podařilo se mu zmizet, přestože policisté použili varovného výstřelu a vypustili služebního psa.
Ještě téže noci se zloděj na místo činu vrátil.
S policisty se tam setkal podruhé.
Tentokrát ho zadrželi.
Jedná se o několikrát trestaného M. K. z Ostravy.
NEDOKONČIL LUP.
Nedaleko skladu tabákových výrobků v Zábřehu v okrese Šumperk bylo nalezeno deset balíků cigaret Marlboro za čtvrt miliónu korun.
Připravil si je tady zatím neznámý zloděj, který je nestačil odnést.

1/10



a: ZLODĚJ

neigh_trees neigh_sent Scale: 100%

Start D:\1_UFAL\Hej... diser_text.odt ... TrEd TRee EDitor ... Microsoft Powe... untitled - Paint EN 11:35

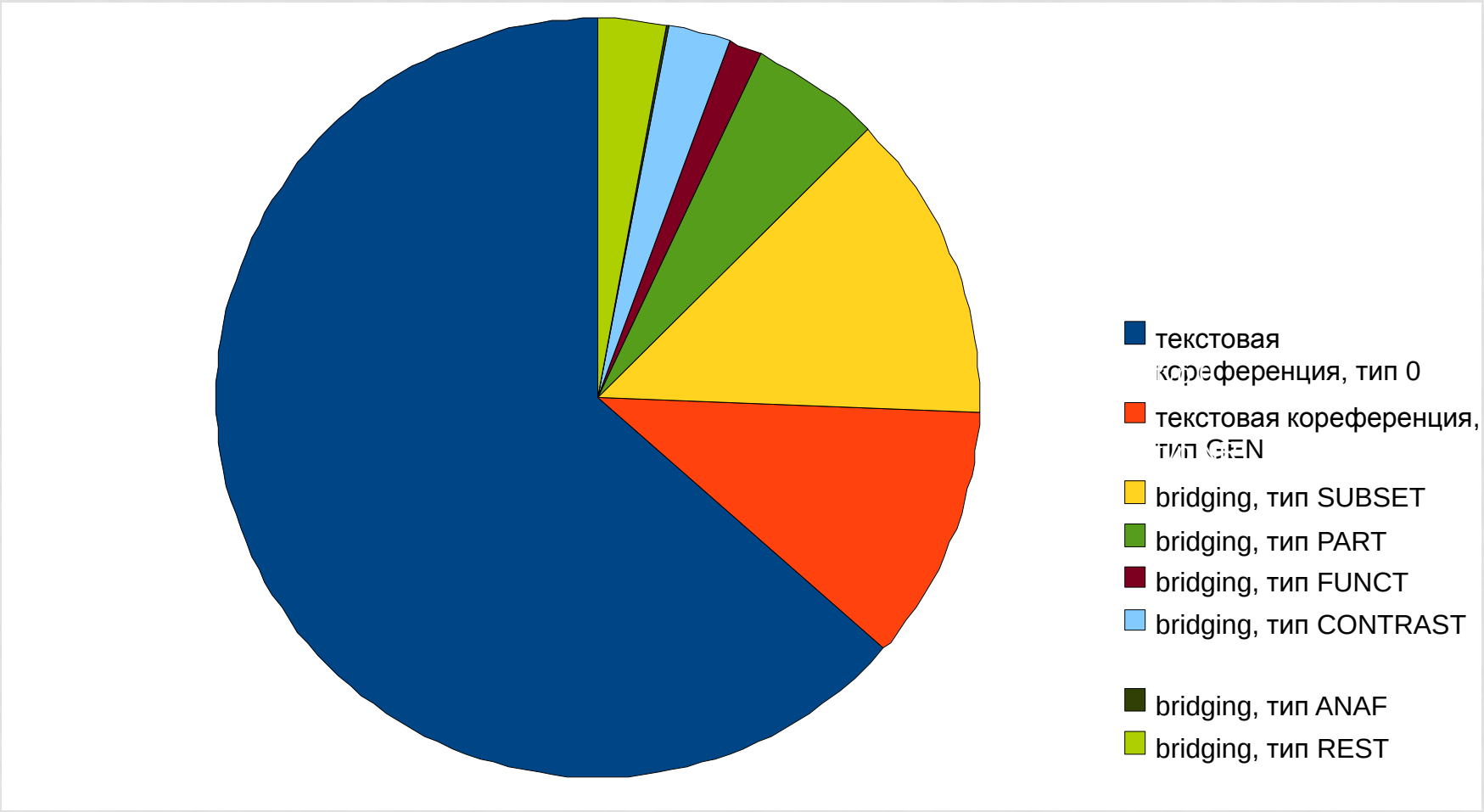
КТО? ГДЕ? КОГДА?

- КТО: 2 разметчика-лингвиста
- ГДЕ: Институт формальной и прикладной лингвистики, Карлов Университет, Прага
- НАЧАЛО: январь 2009
- ПЛАНИРУЕТСЯ ДО конца 2010.

КОЛИЧЕСТВЕННЫЙ АНАЛИЗ РАЗМЕЧЕНЫХ ДАННЫХ

кол-во размеченных файлов	1580
кол-во предложений	23891
кол-во слов	403990
кол-во узлов глубинно-синтаксического уровня	327380
кол-во новых кореферентных связей (именная текстовая кореференция и bridging)	45726
кол-во исходных кореферентных связей (прономинальная и грамматическая кореференция)	10747
кол-во всех кореферентных связей (грамматическая + текстовая + bridging)	55744
узлы ГСУ связанные кореф.	17.00%
всего размечено	**50.00% PDT**

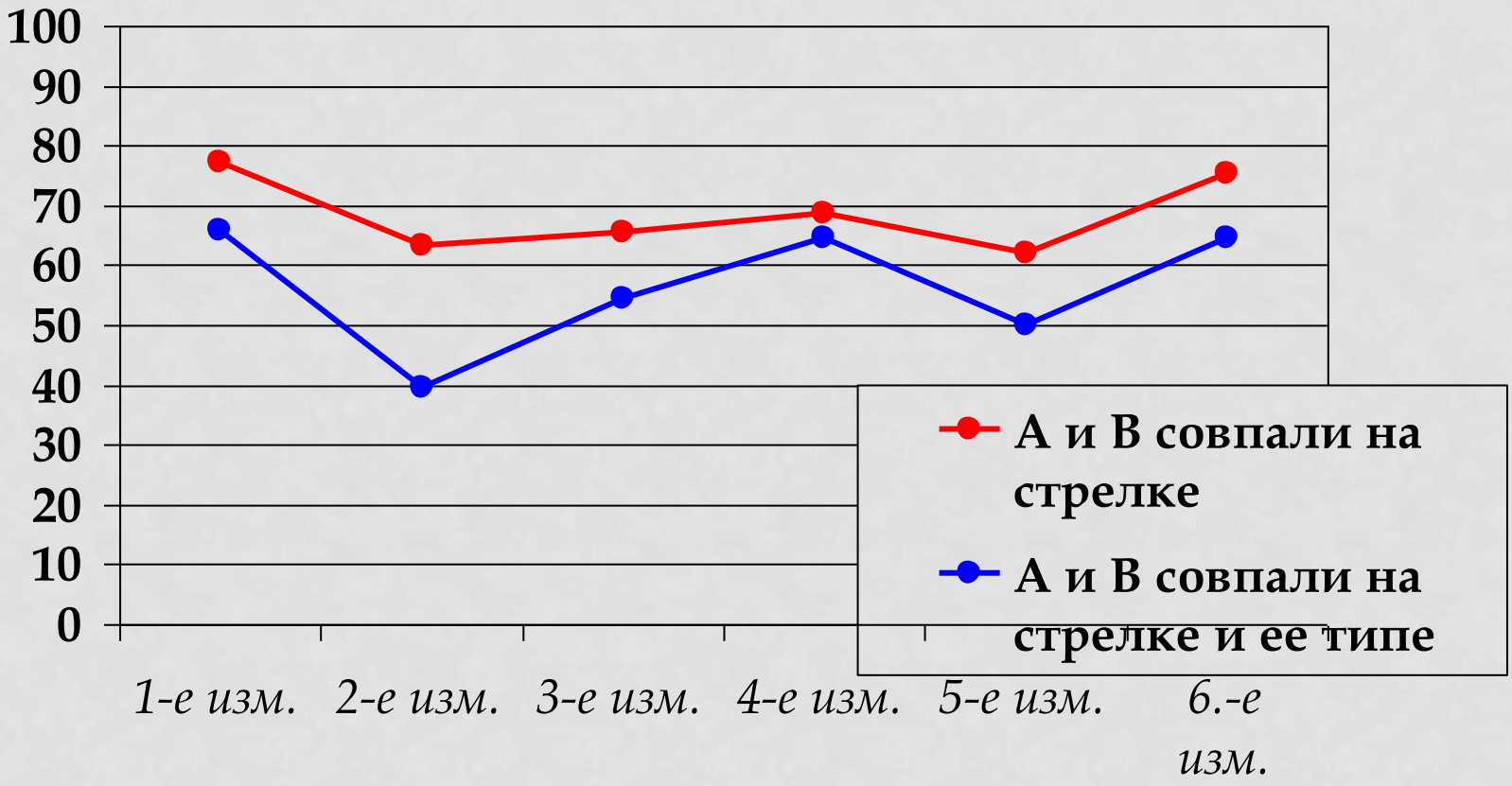
СООТНОШЕНИЕ ТИПОВ КОРЕФЕРЕНТНЫХ ОТНОШЕНИЙ



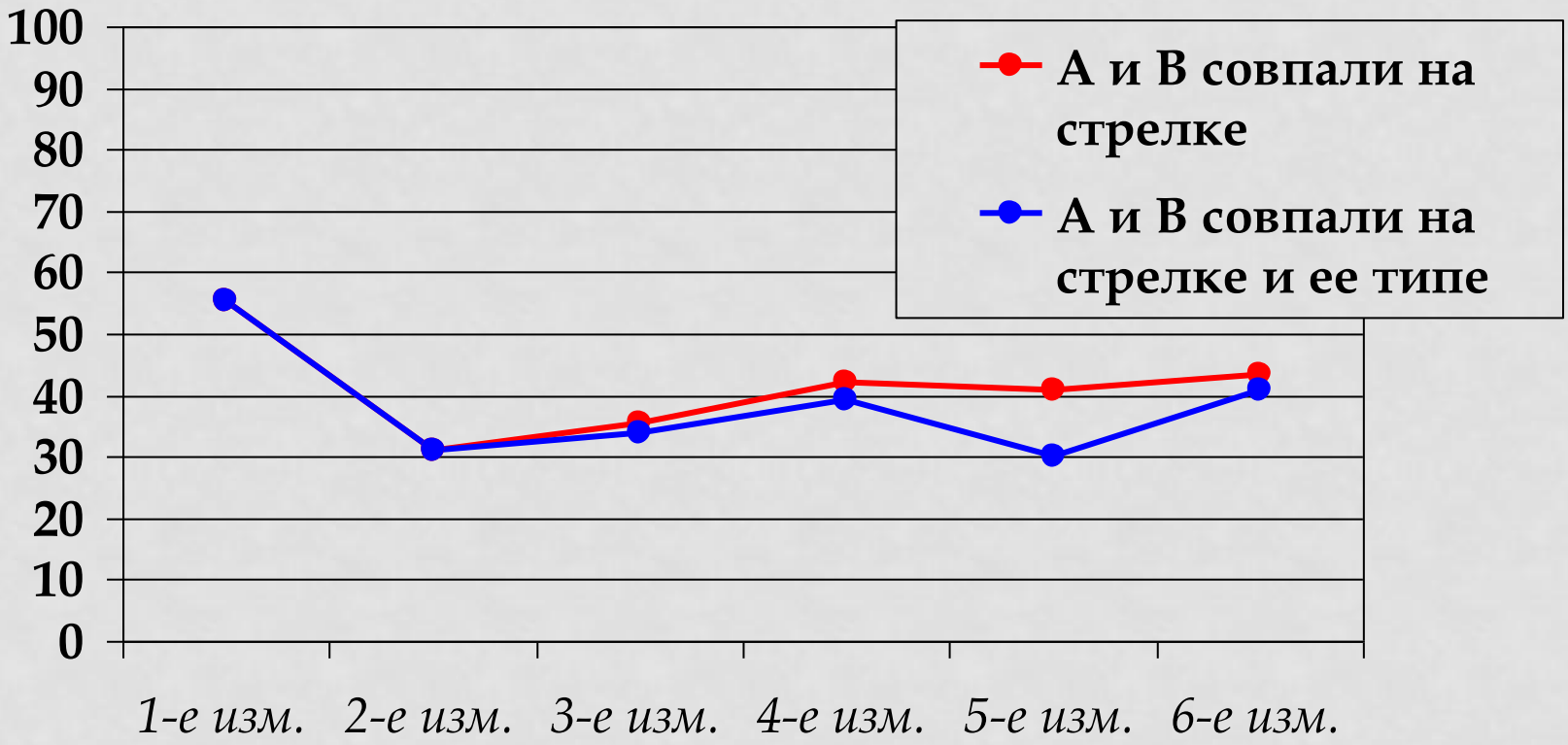
ИЗМЕРЕНИЯ СООТВЕТСТВИЙ МЕЖДУ РАЗМЕТЧИКАМИ

	кол-во файлов	**кол-во предложений**
1	3	41
2	1	40
3	1	101
4	2	106
5	3	100
6	8	211

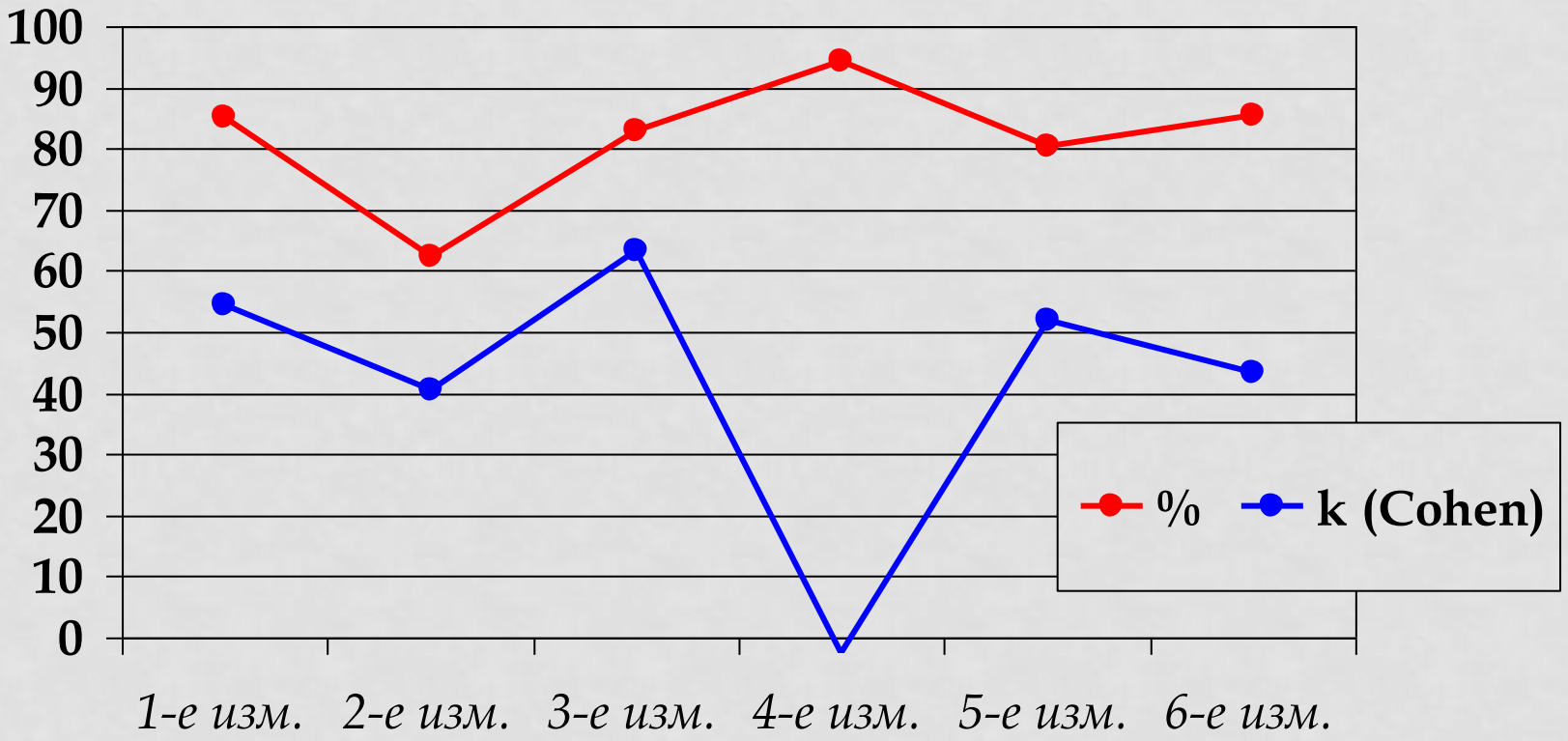
СООТВЕТСТВИЕ МЕЖДУ РАЗМЕТЧИКАМИ – ТЕКСТОВАЯ КОРЕФЕРЕНЦИЯ (F-мера)



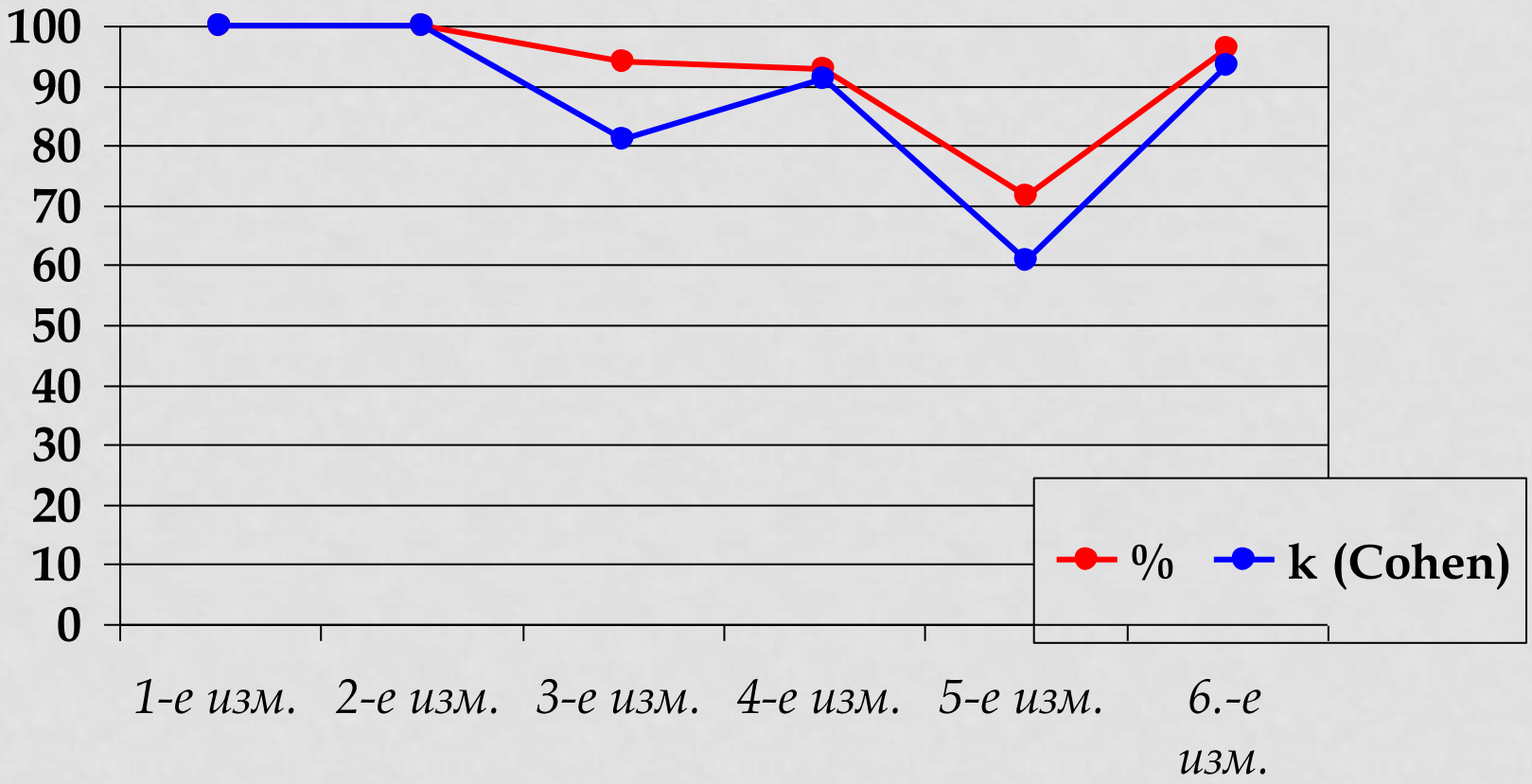
СООТВЕТСТВИЕ МЕЖДУ РАЗМЕТЧИКАМИ – BRIDGING (F-мера)



СООТВЕТСТВИЕ ПО ТИПАМ СВЯЗЕЙ – ТЕКСТОВАЯ КОРЕФЕРЕНЦИЯ



СООТВЕТСТВИЕ ПО ТИПАМ СВЯЗЕЙ – BRIDGIGNG



СООТВЕТСТВИЕ МЕЖДУ РАЗМЕТЧИКАМИ ...

... зависит от уровня сложности текста и его длины

ПРИМЕР ТЕКСТА СО 100%-М СООТВЕТСТВИЕМ

(1) ZLODĚJ SE VRÁTIL.

(2) Policejní hlídka vyrušila v neděli muže, který se vloupal do restaurace Kukačka v obci Horní Životice.

(3) Podařilo se mu zmizet, přestože policisté použili varovného výstřelu a vypustili služebního psa.

(4) Ještě téže noci se zloděj na místo činu vrátil.

(5) S policisty se tam Ø setkal podruhé.

(6) Tentokrát ho Ø zadrželi.

(7) Jedná se o několikrát trestaného M. K. z Ostravy.

(1) ВОР ВЕРНУЛСЯ.

(2) В воскресенье вечером полиция задержала мужчину, который вломился в ресторан «Кукушка» в деревне Горни Животице.

(3) Ему удалось скрыться, несмотря на то, что полиция использовала предупредительный выстрел и выпустила собак.

(4) Но в эту же ночь вор вернулся на место преступления.

(5) Там он встретился с полицией во второй раз.

(6) На этот раз он был задержан Ø .

(7) Речь идет о неоднократно судимом М.К. из г. Остравы.

ПРИМЕР ОЧЕНЬ СЛОЖНОГО ТЕКСТА

(11) Ваша книга описывает различные проблемы – от неизлечимых болезней ребенка до легких дисфункций и влияния развода родителей на психику ребенка.

(12) Из всех описанных проблем конкретную семью может интересовать максимум пять, в худшем случае десять глав.

(13) З.М.: Изначально книга была предназначена для медицинских работников, прежде всего для врачей, которые находятся в непосредственном контакте с проблемными семьями.

(14) Однако выяснилось, что эта тема интересна и для педагогов и воспитателей.

(15) Ведь они постоянно находятся в контакте с проблемными и истязаемыми детьми.

(16) А когда книга была написана, выяснилась, что она небесполезна и для родителей.

(17) Естественно, не любая глава касается любого родителя.

(18) З.Д: Если бы одной семьи касались сразу 10 глав нашей книги, это была бы невероятно несчастная семья.

(19) Хватит и одной, но обычно их бывает больше.

(20) Вот, например, разводы – тридцать тысяч в год по стране, то есть почти тридцати тысяч детей это каким-то образом касается.

(21) В этой книге описывается, как дети переносят развод, как они на него реагируют, и как должны вести себя родители, чтобы их дети меньше страдали.

(22) Или, например, легкая мозговая дисфункция, которой по результатам нашего исследования страдает около пяти процентов детей.

(23) Это заболевания трудно распознаются.

(24) Ребенок малоподвижен, беспокоен, рассеян, но при этом часто очень умен.

(25) Родители считают его лентяем, ругают за плохие оценки, тем самым еще более осложняя его отношение к учебе.

(26)И об этом родители должны знать.

(27) А также педагоги, и в книге содержатся инструкции, как вести себя в подобных ситуациях.

(28)З.М.: Мы рассматриваем также проблемы, о которых часто забывают.

(29) Например, смерть ребенка или рождение больного ребенка.

(30) Причем речь не только о родителях, но и о том, как вести себя окружающим

ТИПЫ НЕСООТВЕТСТВИЙ МЕЖДУ РАЗМЕТЧИКАМИ

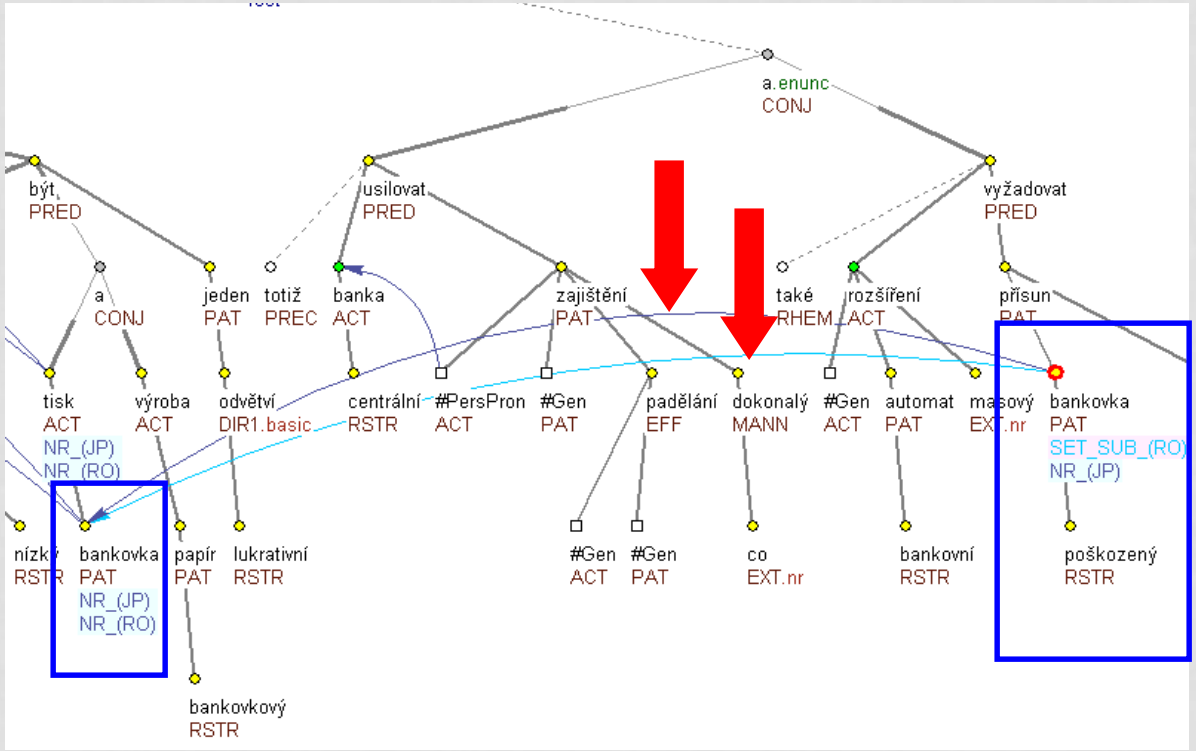
- Разметчик А отметил отношение кореференции (или ассоциативной анафоры) там, где разметчик Б его не увидел
- Разметчик А отметил отношение кореференции там, где разметчик Б отметил отношение ассоциативной анафоры
- Различный выбор первого или второго члена отношения

ПРИМЕР 1: разметчик А обозначил связь там, где разметчик В ее не увидел

чеш. *Na této stránce vám budeme v průběhu 2. vlny kuponové privatizace představovat jednotlivé obory národního hospodářství. Bylo to v době, kdy se nebývale zvýšil zájem zahraničních turistů a podnikatelů o návštěvu České republiky.*

рус. *На этом сайте будут представлены отдельные отрасли национальной экономики. […] Это было в тот период, когда небывало возрос интерес иностранных туристов и предпринимателей к посещению Чешской Республики.*

ПРИМЕР 2: разметчик A отметил кореференци, размечик B - bridging



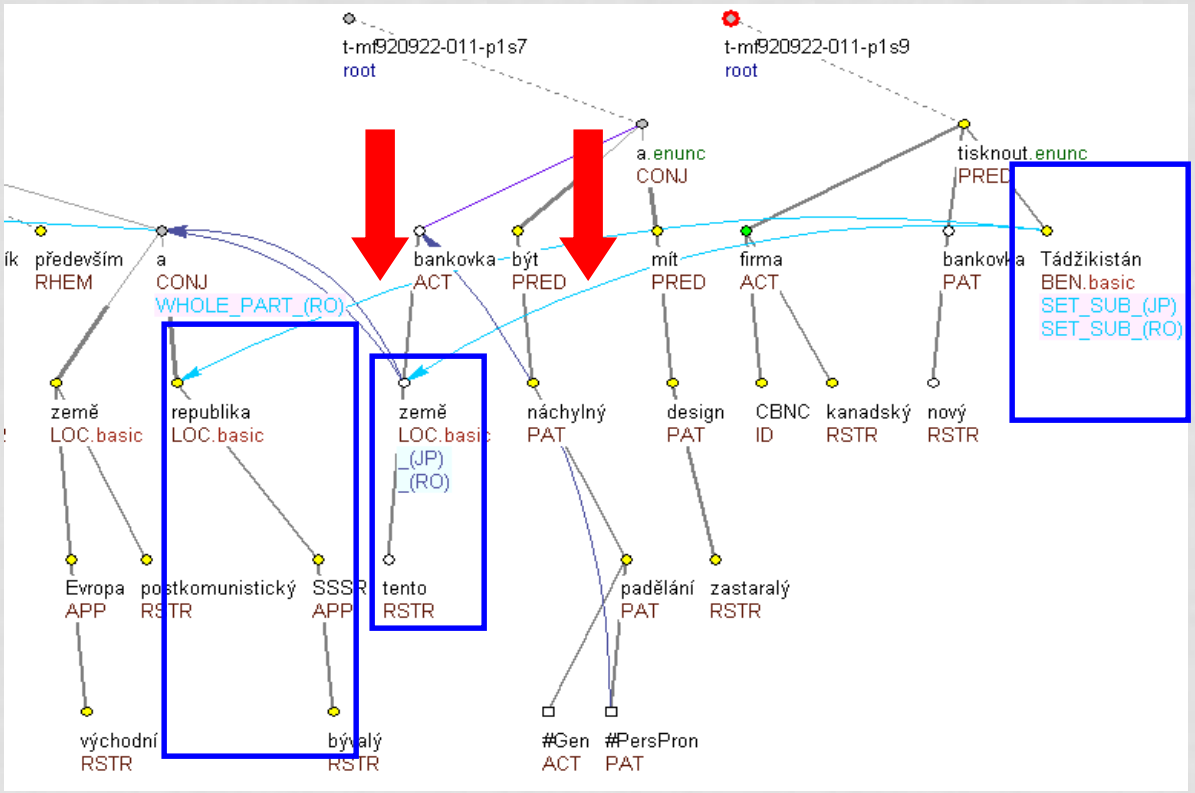
чеш. *I přes klesající inflaci ve světě, a tedy nižší potřebu peněz v oběhu, je tisk bankovek a výroba bankovkového papíru jedním z nejlukrativnějších odvětví. […] … Rozšíření bankovních automatů vyžaduje neustálý přísun nepoškozených bankovek.*

рус. *Несмотря на снижение инфляции в мире, и соответственно меньшую потребность в оборотных денежных средствах, печать банкнот и производство специальной бумаги является одной из наиболее доходных отраслей. […] … В связи с расширением сети банкоматов требуется постоянное пополнение неповрежденных банкнот.*

ПРИМЕР 3: различный выбор антецедента

чеш. *Tiskárny bankovek mají i nové zákazníky, především v postkomunistických zemích východní Evropy a republikách bývalého SSSR. Bankovky v těchto zemích jsou náchylné na padělání a mají zastaralý design. Kanadská firma CBNC bude tisknout nové bankovky pro Tádžikistán*

русрус. *У монетных дворов есть и другие клиенты, прежде всего в посткоммунистических государствах Восточной Европы и в республиках бывшего СССР. Банкноты в этих странах легко подделать, и у них устаревший дизайн. Канадская фирма CBNC будет печатать новые банкноты для Таджикистана.*



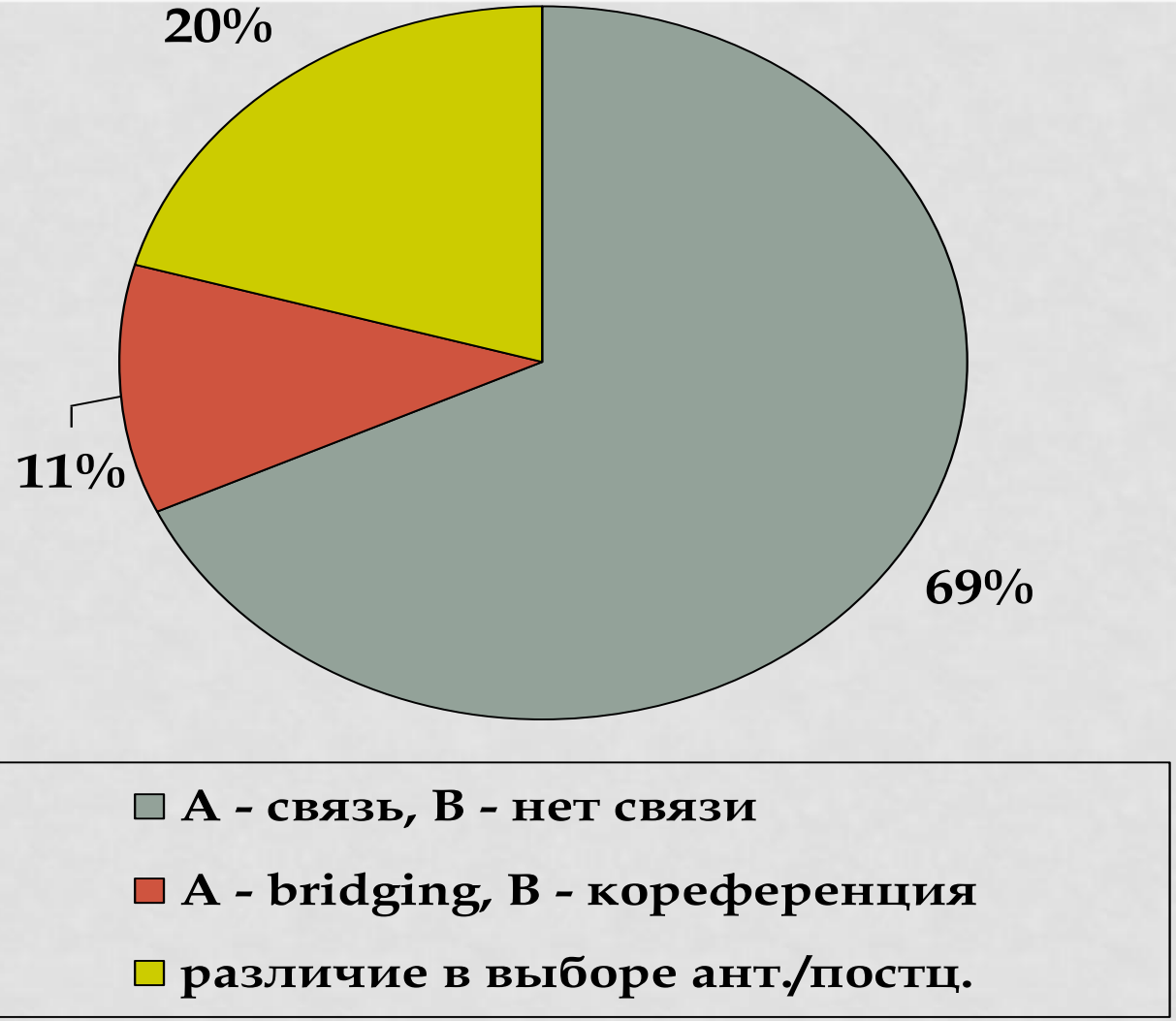
ПРИМЕР 4: РАЗЛИЧИЯ В ГЛУБИНЕ ИНТЕРПРЕТАЦИИ

- чеш. *Méně výnosný cestovní ruch. Hotelových kapacit je mnohem víc než současná poptávka.* [...] *Bylo to v době, kdy se nebývale zvýšil zájem zahraničních turistů a podnikatelů o návštěvu České republiky, především Prahy.*

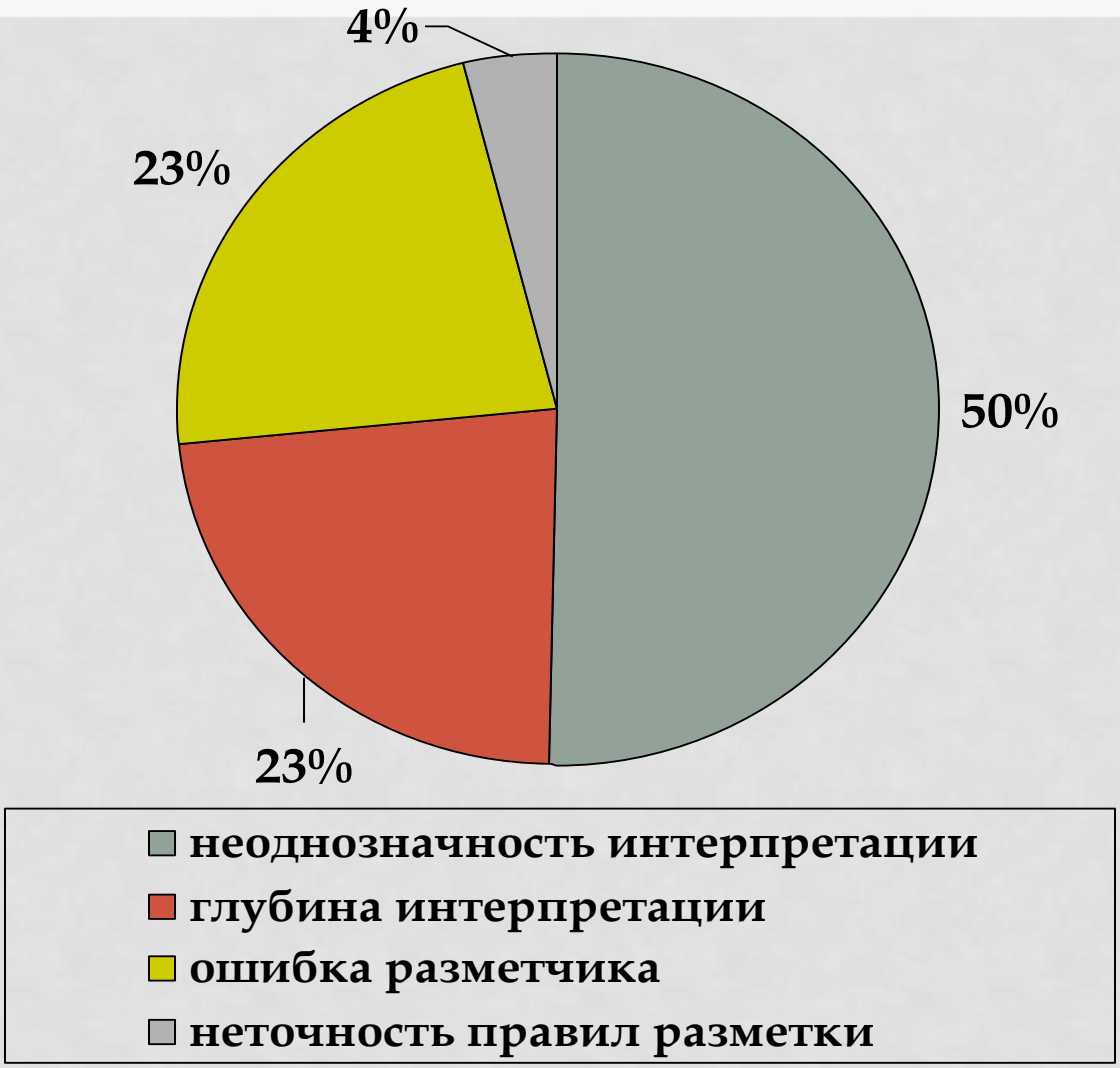
  рус. *Менее доходным является туризм. Количество мест в гостиницах существенно превышает современный спрос.* […] *Это было в тот период, когда небывало возрос интерес иностранных туристов и предпринимателей к посещению Чешской Республики.*

ТУРИЗМ = {МЕСТА В ГОСТИНИЦАХ, ИНОСТРАННЫЕ ТУРИСТЫ, ...}

ТИПОЛОГИЯ НЕСООТВЕТСТВИЙ МЕЖДУ РАЗМЕТЧИКАМИ



вероятные причины несоответствий между разметчиками



причины несоответствий классифицированные по типам

тип несоответствия	вероятная причина	%
установил связь – не увидел связи	неоднозначность	32
	ошибка разметчика	31
	глубина	34
	техн. ошибка	2
	ошибка правил	1
bridging vs. кореференция	неоднозначность	100
различный выбор антецедента/постцедента	неоднозначность	71
	ошибка разметчика	7
	не знаю	7
	ошибка правил	15

СПАСИБО ЗА ВНИМАНИЕ