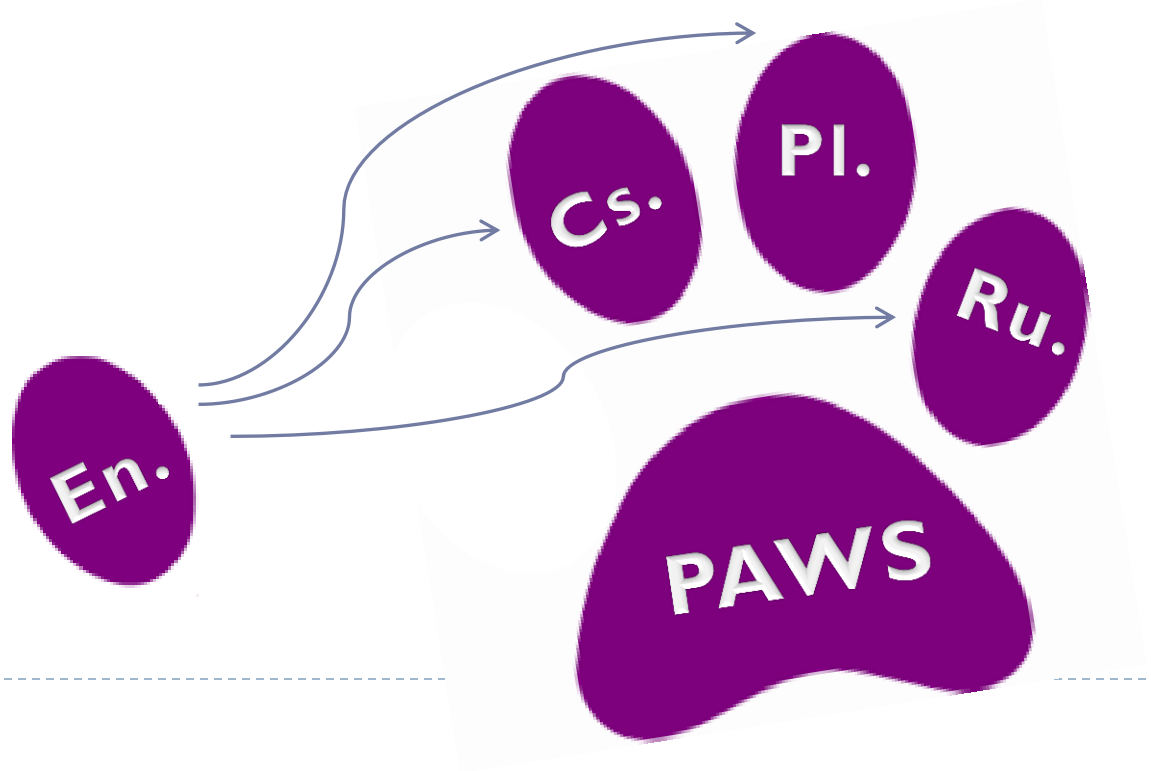
Comparable coreference expressions in parallel Czech, English, Polish and Russian data

Anja Nedoluzhko, Maciej Ogrodniczuk, Michal Novák

Promised in Abstract

- analysis of cohesive devices important for discourse analysis
- three Slavonic languages vs. English based on translated texts



Special points to compare

- finite and infinite constructs:
  - relative clauses,
  - participial constructions,
  - possessive constructions and
  - correlative constructions with a demonstrative pronoun.
- pro-drop qualities of three Slavonic languages in comparison with English.

Data - PAWS

Parallel Anaphoric Wall Street Journal

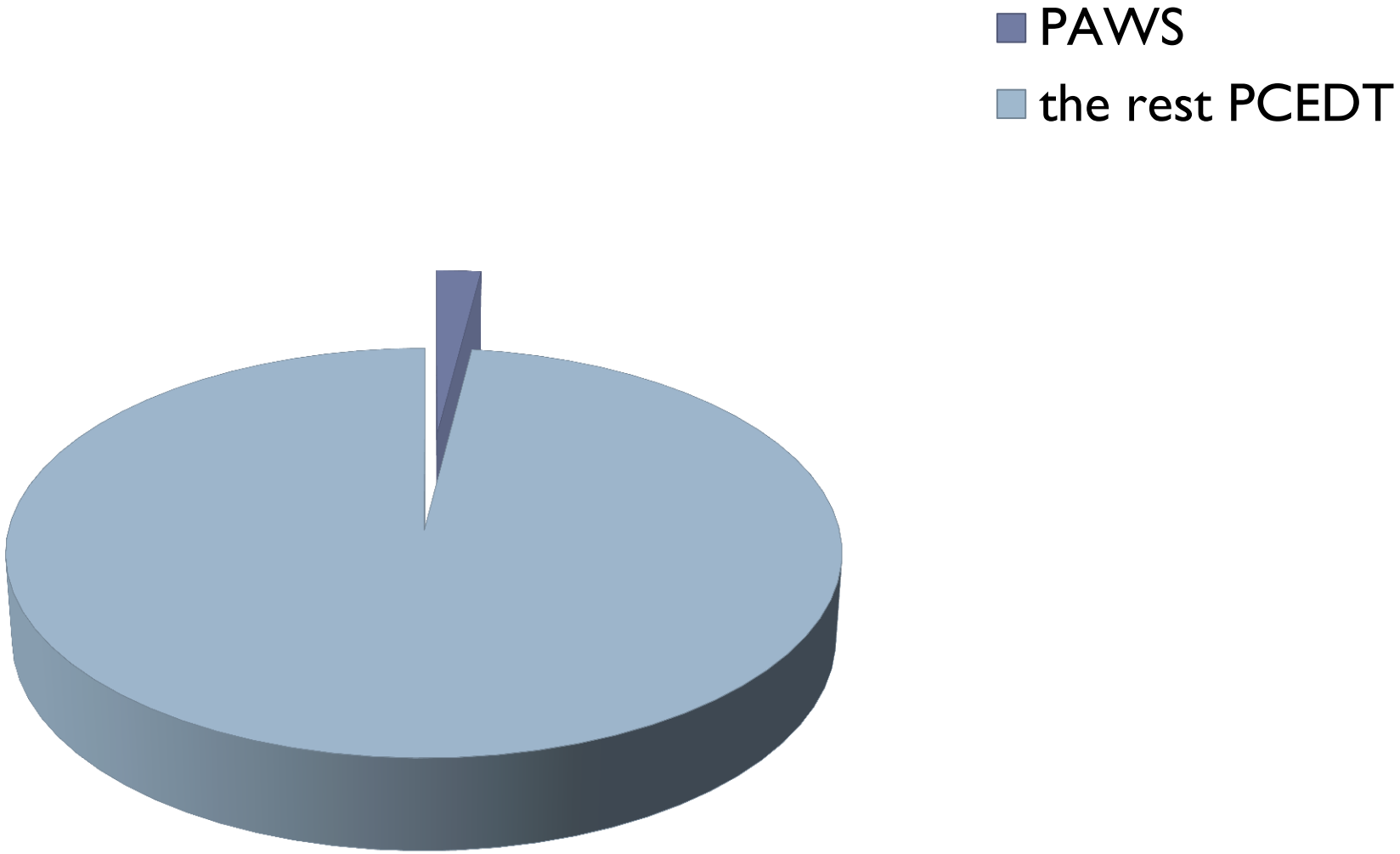
Data - PAWS

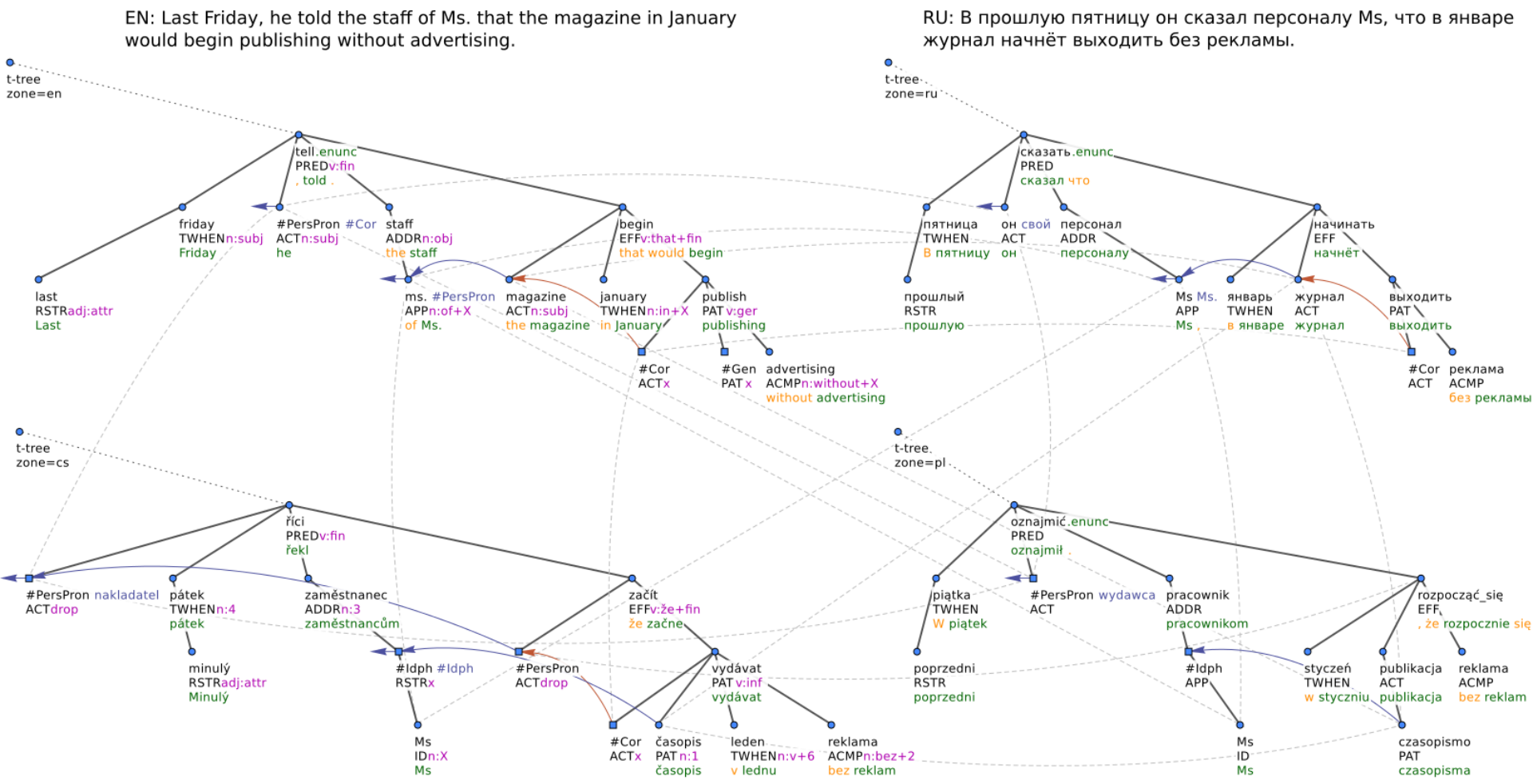
- A first half of the PCEDT section 19, particularly the 50 documents from wsj 1900 to wsj 1949
- Manual annotation of word alignment

PCEDT – Prague Czech-English Dependency Treebank

- Prague Czech-English Dependency Treebank [Hajic et al., 2012]
- English Wall Street journal texts translated to Czech sentence by sentence
- 1.2 million words in almost 50,000 sentences for each language
- annotated on morphological (m-layer), analytical (shallow syntactic, a-layer) and tectogrammatical (deep syntactic, t-layer),
- sentence-aligned, word-aligned
- t-layer includes
  - semantic labeling of content words and coordinating conjunctions
  - argument structure description based on a valency lexicon
  - coreference annotation
  - ellipsis reconstruction

PCEDT and PAWS

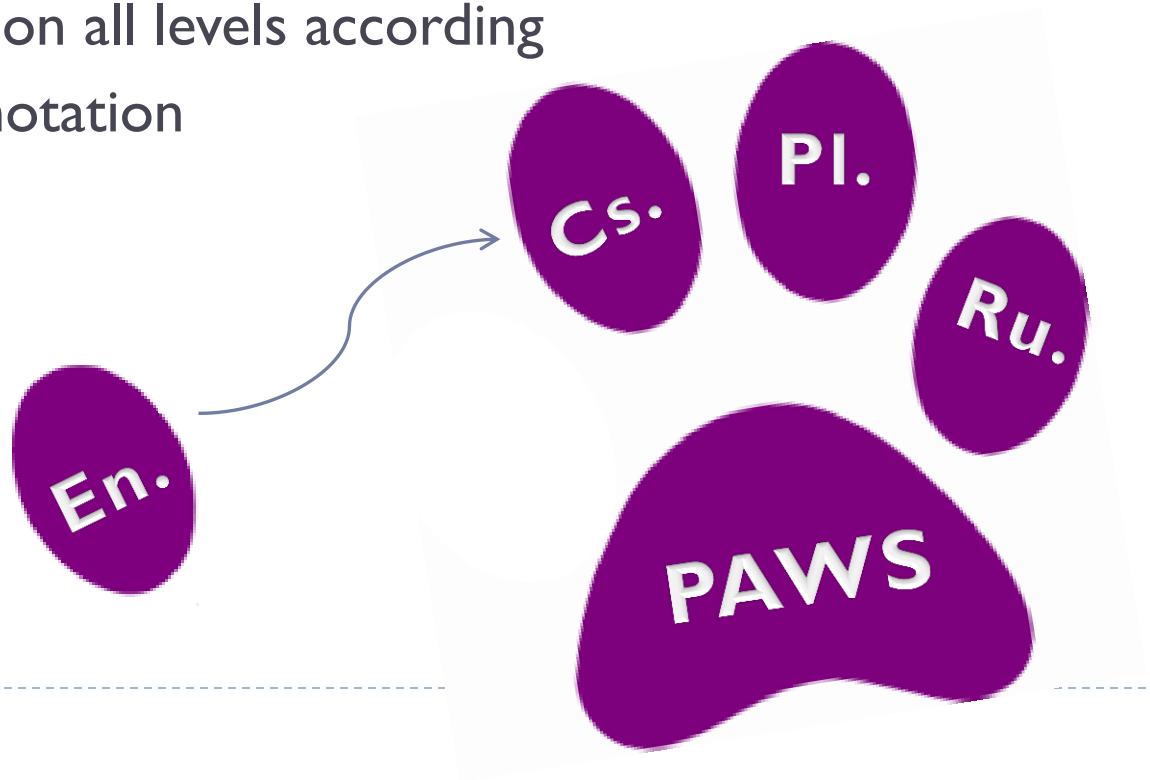




Last Friday, he told the staff of Ms. that the magazine in January would begin publishing without advertising.

What Do We Have So Far?

- All texts are translated from English into Czech, Russian and Polish
- For the **English-Czech** part:
  - alignment,
  - annotation on all levels according to Prague annotation scenario



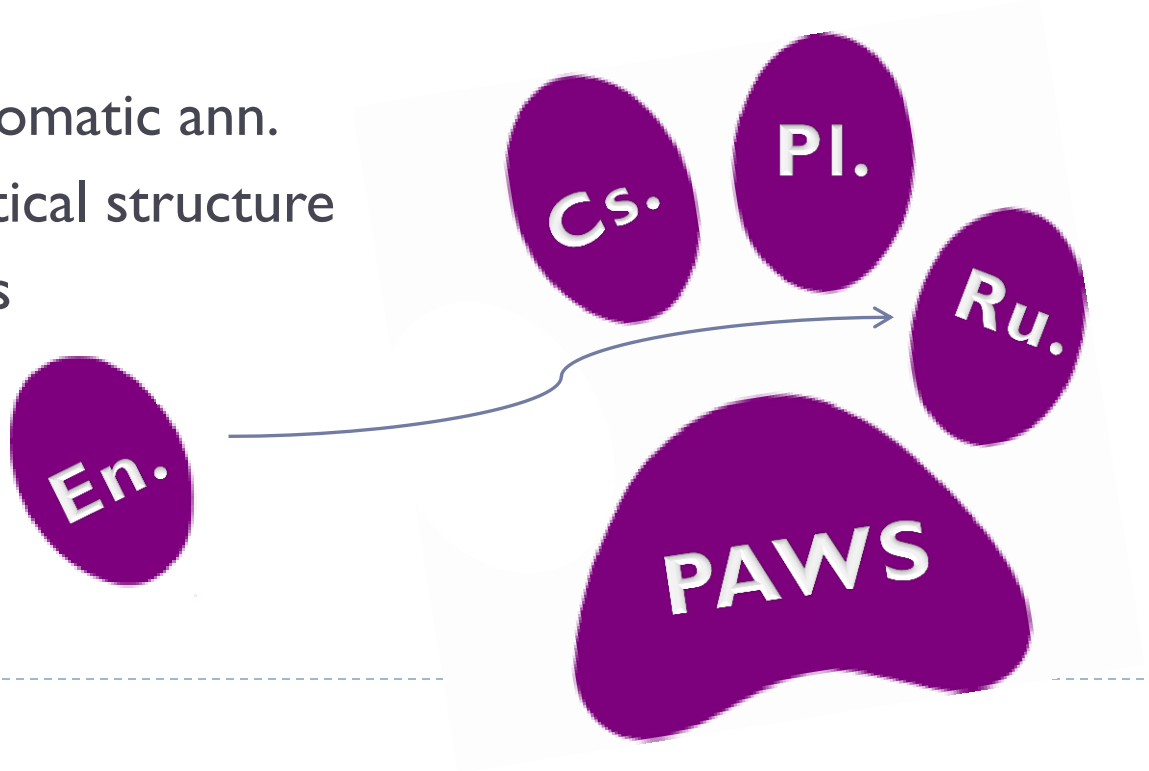
What Do We Have So Far?

- All texts are translated from English into Czech, Russian and Polish
- For the **English-Czech** part,
- For the **English-Russian** part:
  - alignment,
  - parsing, automatic ann.

tectogrammatical structure

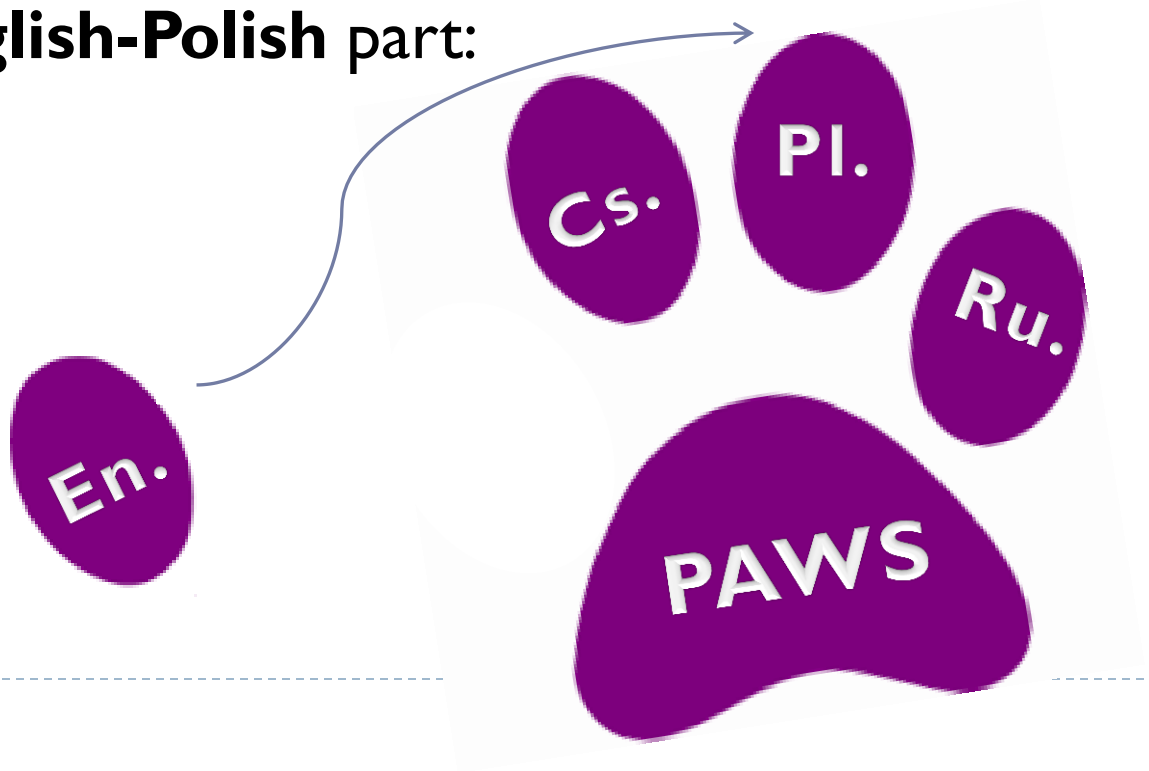
and sem. roles

coreference



What Do We Have So Far?

- All texts are translated from English into Czech, Russian and Polish
- For the **English-Czech** part,
- For the **English-Russian** part,
- For the **English-Polish** part:
  - alignment,
  - parsing
  - ca. 50% of coreference



Translation factor

- very important
- factors of translation: e.g. expliciteness
- multiple variants of translation

BUT

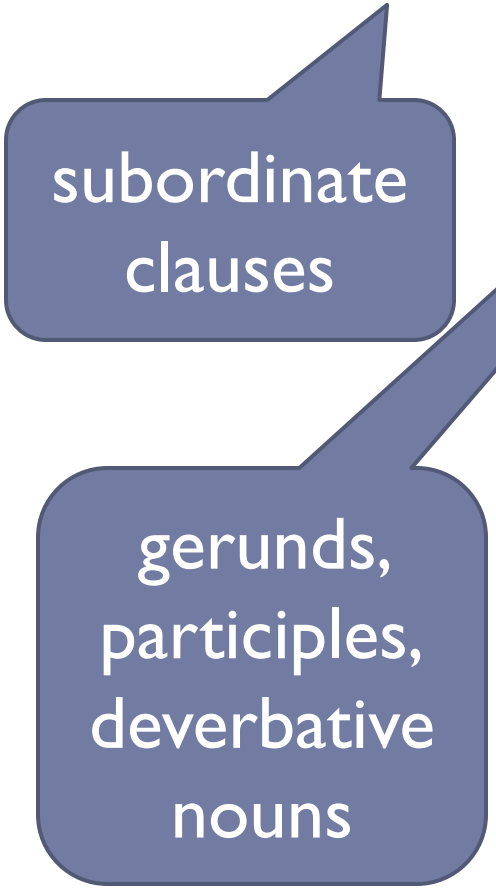
- appropriateness of this very translation variant in the language

Observations on the (being) Annotated Data

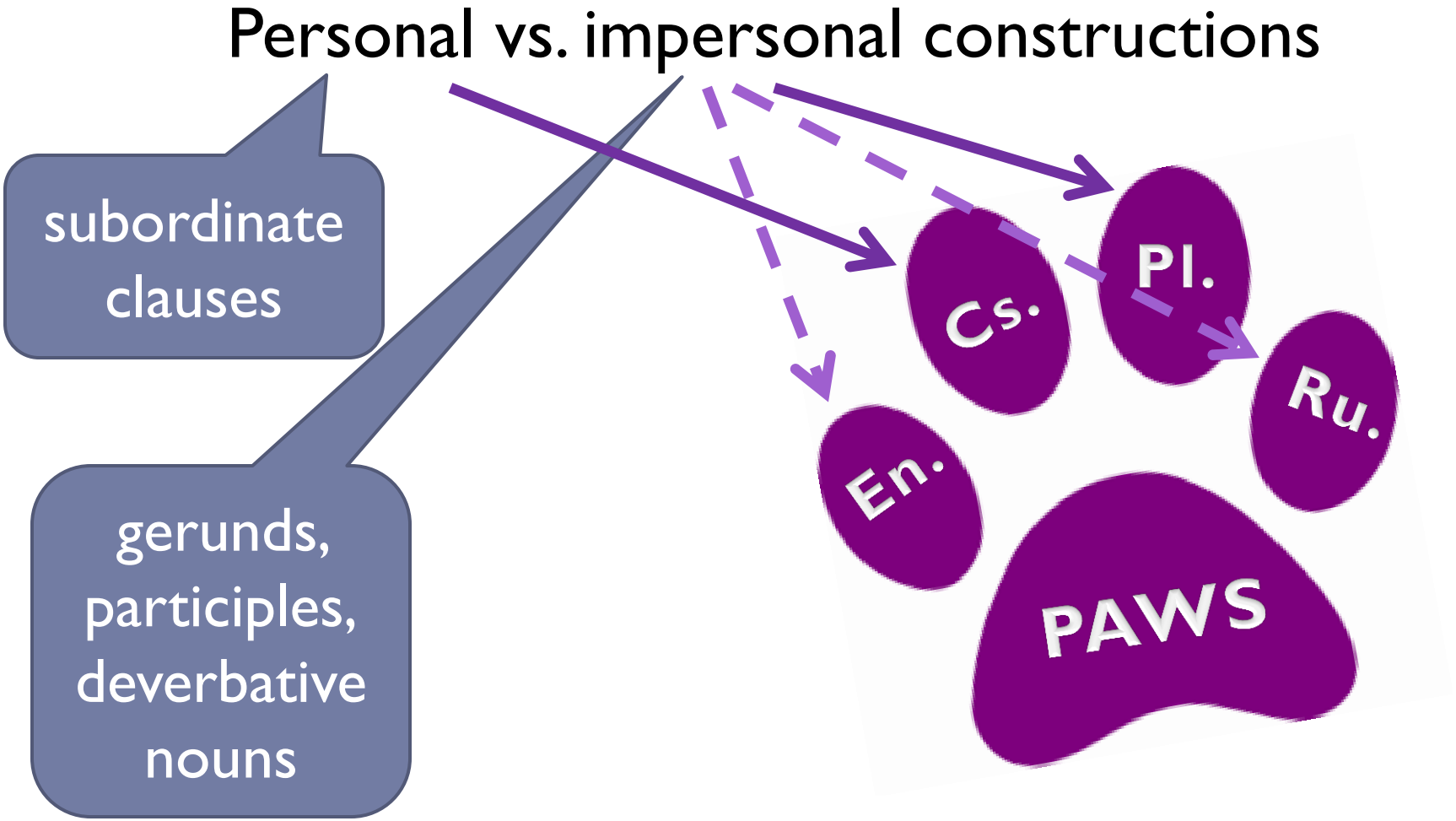
Personal vs. impersonal constructions

Observations on the (being) Annotated Data

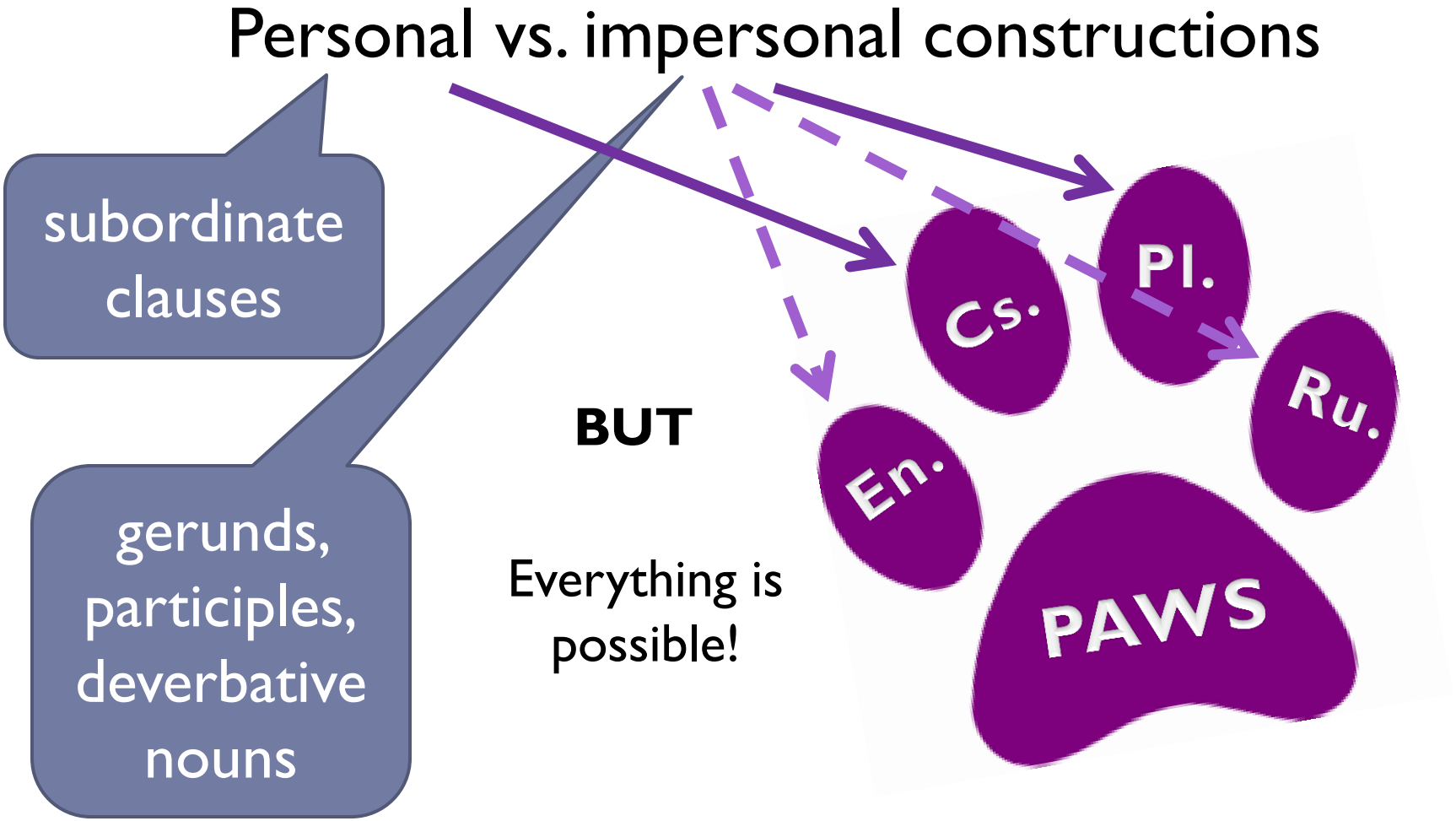
Personal vs. impersonal constructions



Observations on the (being) Annotated Data



Observations on the (being) Annotated Data

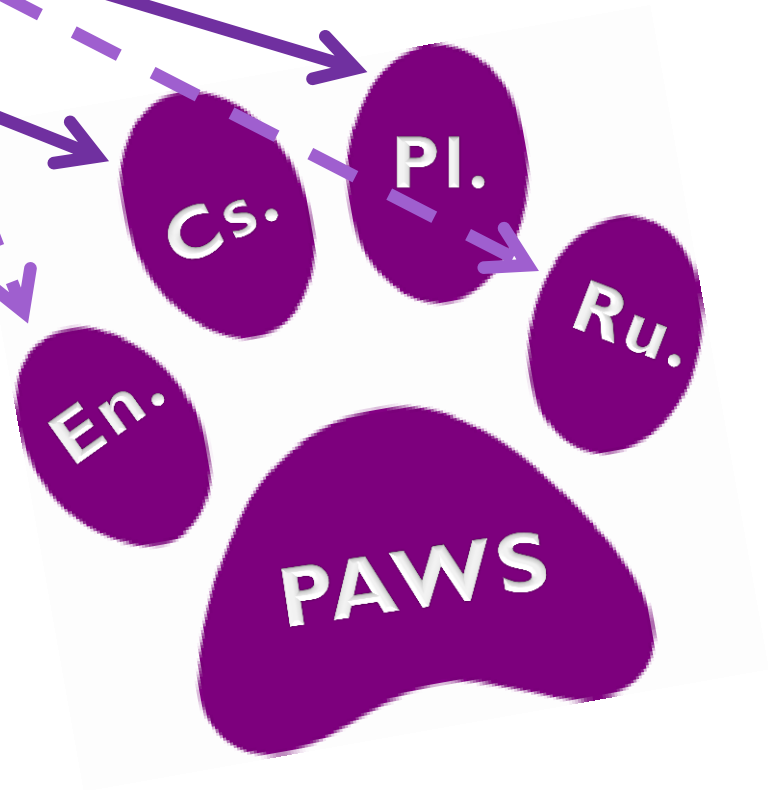


Observations on the (being) Annotated Data

Personal vs. impersonal constructions

subordinate clauses

gerunds, participles, deverbative nouns



BUT

Everything is possible!

Very strong translation factor!

PL: Morrison Knudsen Corp. zaksięgował dochód netto za trzeci kwartał równy 7,9 milionom dolarów, czyli 69 centów za akcję, **kontynuując** odbicie po znacznych zeszłorocznych stratach.

CZ: Společnost Morrison Knudsen Corp. vykázala čistý zisk za třetí čtvrtletí ve výši [7,9]7.9 miliónu dolarů, neboli 69 centů na akcii, **čímž pokračuje** v zotavení z velkých loňských ztrát.

EN: Morrison Knudsen Corp. posted third-quarter net income of $7.9 million, or 69 cents a share, *-1 **continuing** a rebound from steep year-ago losses.

RU: Корпорация Morrison Knudsen опубликовала данные о чистых доходах, составивших $7.9 млн. или 69 центов за акцию, в третьем квартале, **продолжая** восстанавливаться после больших прошлогодних убытков.

Infinitive (EN, RU) – deverbative NP (PL) - subordinate clause (CZ)

PL: W odpowiedzi na szczegółową ofertę, Gary Risley, zastępca prezesa Mesy, powiedział, że zarząd poprosi dyrektorów **o zatrudnienie** konsultanta finansowego w celach doradczych.

CZ: Gary Risley, vicepresident společnosti Mesa, uvedl, že jako odpověď na konkrétní nabídku požádá vedení společnosti představenstvo, **aby použilo služeb finančního poradce**.

EN: In response to the specific offer, Gary Risley, Mesa vice president, said management will ask directors *-1 **to employ** a financial consultant 0 *T*-2 to advise them.

RU: В ответ на конкретное предложение Гэри Рисли, вице-президент Mesa, сказал, что руководство попросит директоров **нанять** финансового советника для получения консультации.

Infinitive (EN, RU) – deverbative NP (PL) - subordinate clause (CZ)

PL: W odpowiedzi na szczegółową ofertę, Gary Risley, zastępca prezesa Mesy, powiedział, że zarząd poprosi dyrektorów **o zatrudnienie** konsultanta finansowego w celach doradczych.

CZ: Gary Risley, vicepresident společnosti Mesa, uvedl, že jako odpověď na konkrétní nabídku požádá vedení společnosti představenstvo, **aby použilo služeb finančního poradce**.

EN: In response to the specific offer, Gary Risley, Mesa vice president, said management will ask directors *-1 **to employ** a financial consultant 0 *T*-2 to advise them.

RU: В ответ на конкретное предложение Гэри Рисли, вице-президент Mesa, сказал, что руководство попросит директоров **нанять** финансового советника для получения консультации.

+ other options

- other options

A step to almost secondary prepositions

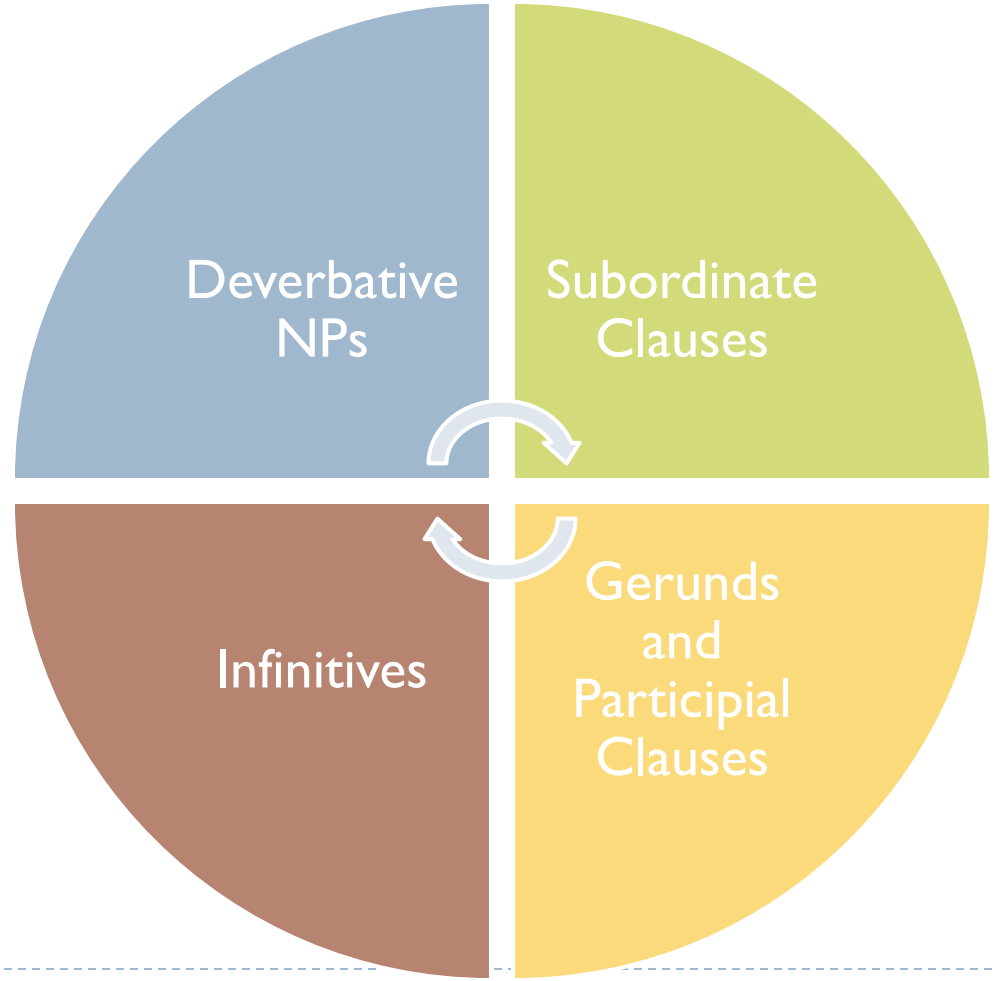
PL: **Bazując** na pozostałej liczbie akcji Mesy, które nie są jeszcze w posiadaniu StatesWest, proponowane przejęcie osiągnęłoby wartość około 15,3 milionów dolarów.

CZ: **Na základě** množství akcií společnosti Messa v oběhu, které společnost StatesWest dosud nevlastní, bude mít navrhované převzetí hodnotu okolo [15,3]15.3 milionu dolarů.

EN: **Based** on the number of Mesa shares outstanding not already owned * by StatesWest, the proposed takeover would have a value of about $15.3 million.

RU: Запланированное поглощение, **исходя из** количества акций Mesa, которые ещё не выкуплены StatesWest, имело бы стоимость почти $15.3 млн. StatesWest владеет 7.25% Mesa.

Interesting to compare



Pronominalisations

PL: W następstwie **trzęsienia** ziemi w Kalifornii, Prezydent Bush i jego pomocnicy wczoraj rano rzucili się w wir związanej z **nim** działalności.

CZ: Po kalifornském **zemětřesení** prezident Bush se svými poradci včera ráno vlétl do víru činností spojených se **zemětřesením**.

EN: In the aftermath of the California earthquake, President Bush and his aides flew into a whirlwind of **earthquake-**related activity yesterday morning.

RU: После калифорнийского землетрясения президент Буш и его помощники вчера утром устремились в ураганную деятельность, связанную с **землетрясением**.

Pronominalisations

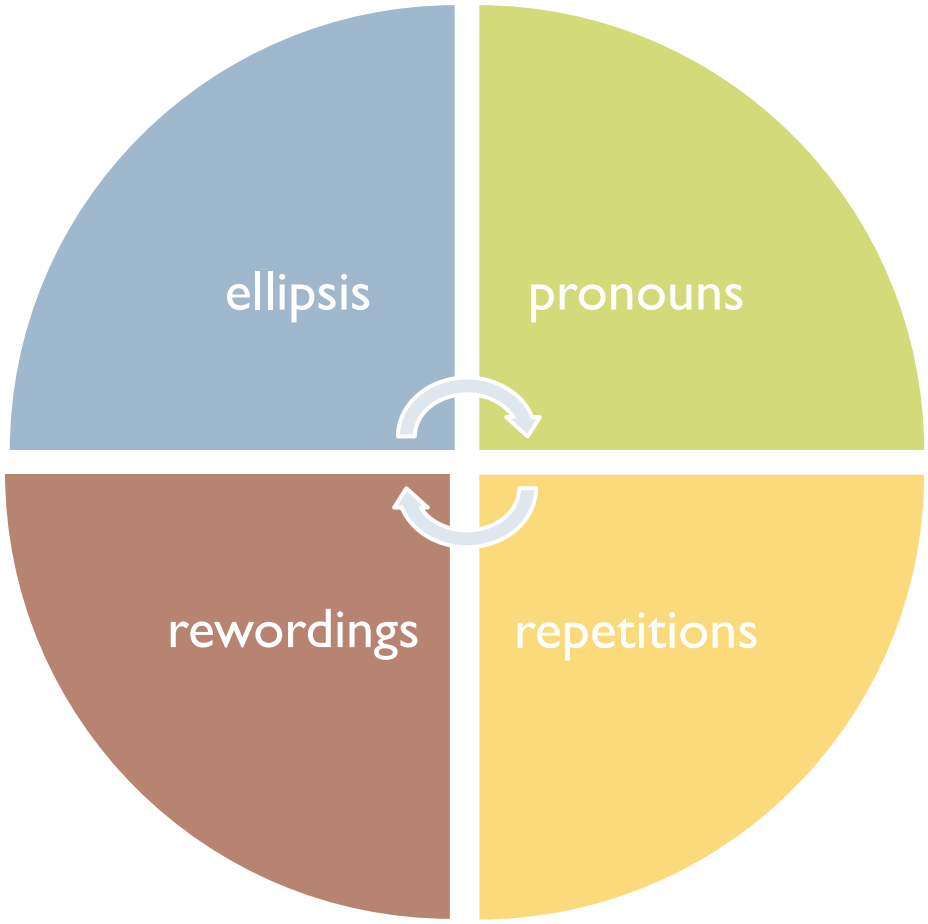
CZ: Dopoledne měl Bush dva telefonáty od viceprezidenta Dana Quayleho, který je v Kalifornii, natočil televizní prohlášení, kde vyjádřil svoji účast, podepsal vyhlášení katastrofy, dostal písemnou zprávu **od Federální agentury pro krizové řízení (FEMA)** a navštívil **vedení agentury FEMA**.

EN: By noon, Mr. Bush had taken two phone calls from Vice President Dan Quayle, who *T*-1 was in California; made a televised statement of concern; signed a disaster proclamation; received a written report from **the Federal Emergency Management Agency**; and visited **FEMA headquarters**.

RU: К полудню г-н Буш получил два телефонных звонка от Вице-президента Дэна Куейла, который находился в Калифорнии; сделал переданное по телевидению заявление о своей озабоченности вопросом; подписал указ об аварийной ситуации; получил письменный отчёт **от Федерального Агентства по чрезвычайным ситуациям**; и посетил **штаб-квартиру FEMA**.

PL: Do południa Bush odbył dwie rozmowy telefoniczne z wiceprezydentem Danem Quaylem przebywającym w Kalifornii; wydał oświadczenie wyrażające zaniepokojenie, transmitowane w telewizji; podpisał ogłoszenie katastrofy; otrzymał pisemny raport **od Federalnej Agencji Zarządzania Kryzysowego** oraz odwiedził **jej** siedzibę.

To be compared within coreference chains



Conclusion

- a new parallel word-aligned corpus is created: English, Czech, Polish, Russian
- coreference is annotated
- analysis of parallel constructions, from the point of view of coreference relations
- original questions wait for statistics (pro-drop, possessive means, etc.)

THANK YOU!