HUMAN CAPITAL
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

Overview

* Annotation of textual phenomena in the Prague Dependency Treebank (April 27)
* **Coreferential expressions in English and Czech (April 28)**
* Coreference in Czech and cross-lingually – ideas and perspectives (April 30)

"made in ÚFAL" Institute of Formal and Applied Linguistics, Charles University in Prague, Faculty of Mathematics and Physics)

HUMAN CAPITAL
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

* Coreferential expressions in English and Czech

Michal Novák, Anna Nedoluzhko

PhD STUDIES

Warsaw, 28.4.2015

Example

EN: ***It*** *switched to a caffeine-free formula* ***Ø*** *using* ***its*** *new Coke in 1985.*
CS: *V roce 1985* ***Ø****.ACT přešla na bezkofeinovou recepturu,* ***kterou*** *používá pro* ***svoji*** *novou kolu.*

- different languages use different means of expressing coreference (or they prefer some more than the others)
- parallel English – Czech annotated corpus
- comparison in terms of syntax and deep syntax

HUMAN CAPITAL
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

Motivation

- NLP tasks
  - anaphor detection – categories helping to improve mention detection
  - bilingual coreference resolution (= treebanking)
  - co-training [Blum and Mitchell, 1998] with language-dependent feature sets – use semi-supervised technique of ML: co-training using two set of features (2 languages), resulting in systems working on monolingual texts
  - MT via deep syntax: TectoMT [Popel and Zabokrtsky, 2010]
- Theoretical
  - linguistic typology
  - grammatical structure of individual languages
  - preference for different types of constructions – translatology, etc.

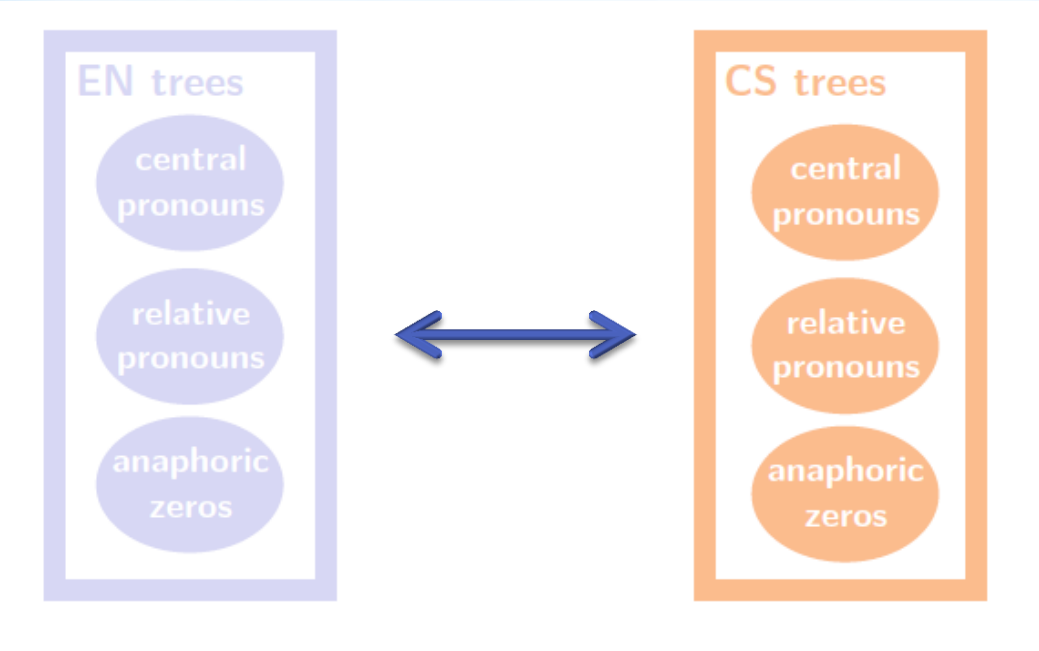
*Workflow

* Parallel treebank

*Workflow

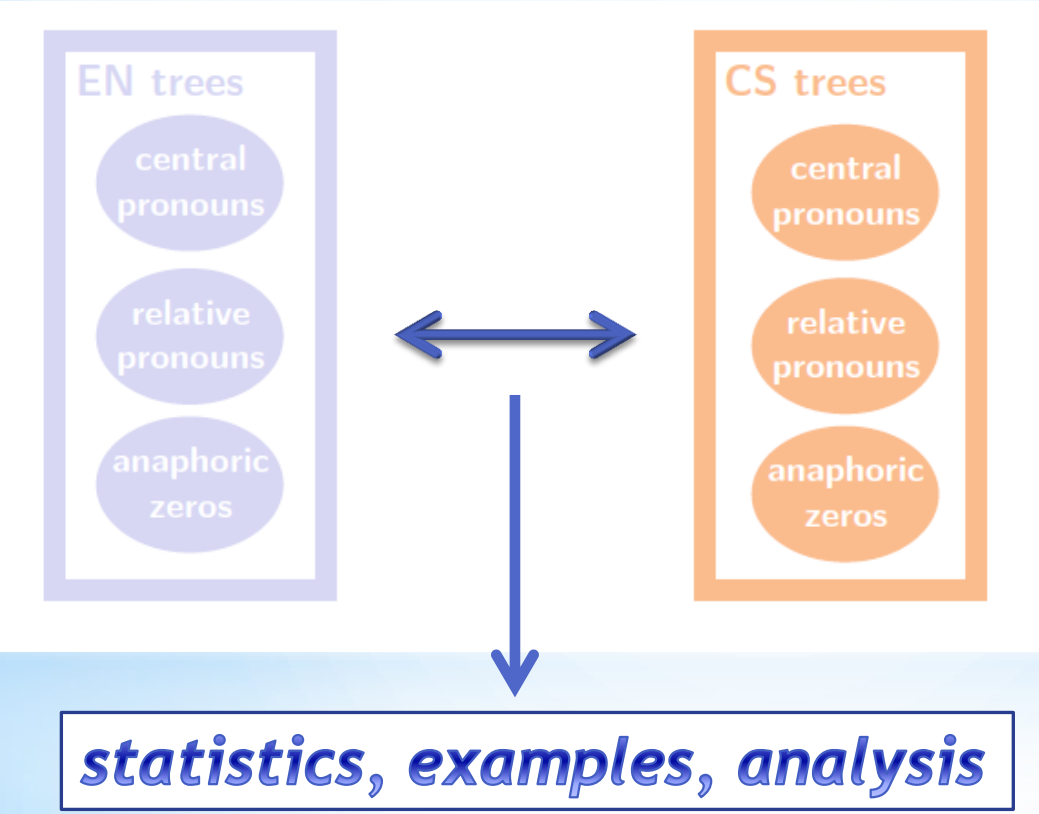
* Parallel treebank
* Classes of nodes

*Workflow



* Parallel treebank
* Classes of nodes
* Alignment

*Workflow



- Parallel treebank
- Classes of nodes
- Alignment
- Statistics of counterparts

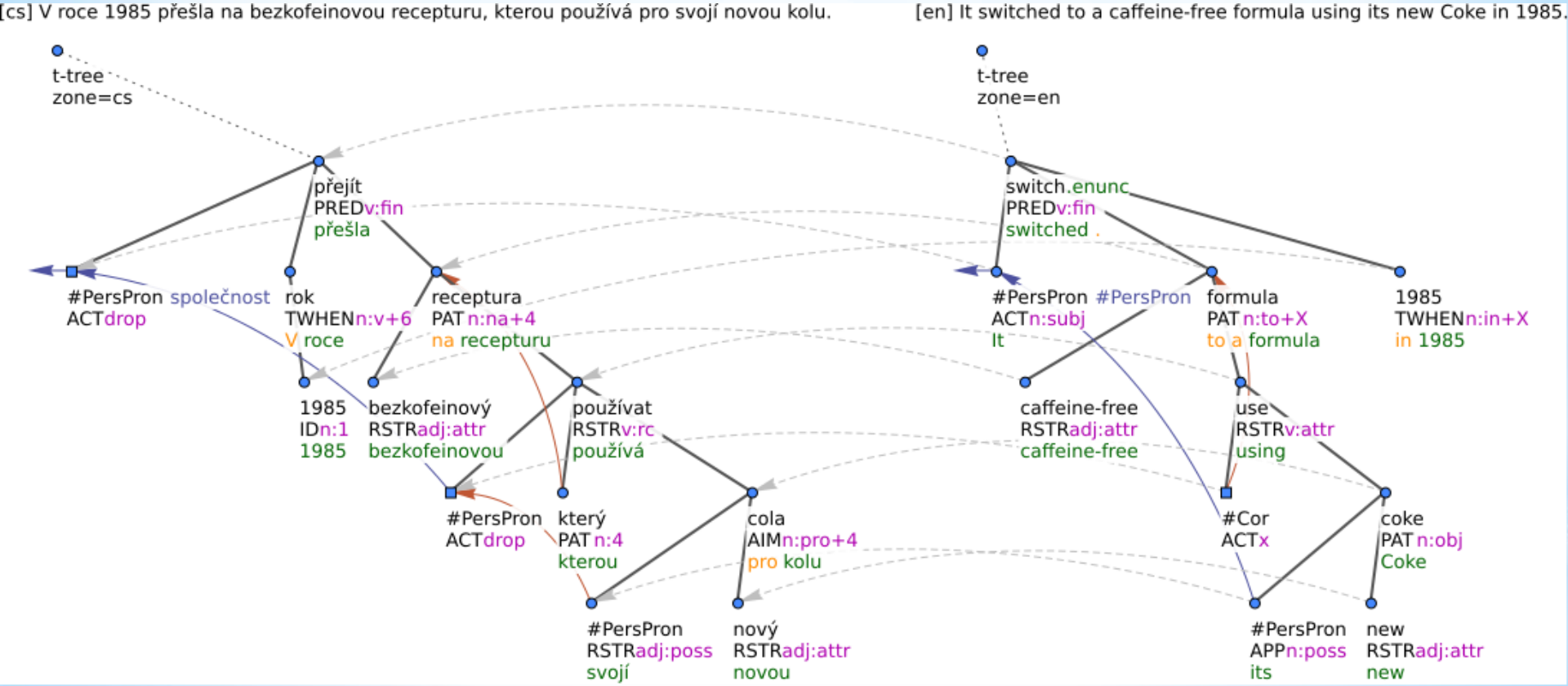
Related Work

* Theoretical linguistics and typology
  * Referent activation theories (theory of topicality in [Givon, 1983]; hierarchy of referential devices in [Ariel, 2015]; neural networks model for pronominal choice in [Kibrik, 1997], [Kibrik, 2011]; saliences in [Hajicova et al., 2006])
  * Analysis of one language vs. linguistic typology
* Corpora: Postolache et al., 2006 [Rom-En], Guillou et al., 2014 [DE-EN ParCor], Bojar et al., 2012 [CzEng]
* Corpus-based studies: (Kunz and Lapshinova-Koltunski, 2015; Zinsmeister et al., 2012)
* NLP applications: coreference projection: de Souza and Orasan, 2011, Postolache et al., 2006, Rahman and Ng, 2012; Ogrodniczuk, 2013
* Czech-English comparison: [Onderkova, 2009], [Veselovska et al., 2012], [Novak et al., 2013], [Novak and Zabokrtsky, 2014])

*PCEDT

* 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 - sample tree and alignment



*PCEDT - Coreference

* Grammatical coreference
  * Coreference with **reflexive pronouns**,

    My daughter likes to dress herself without my help;
  * Coreference with **relative elements** (pronouns and pronominal adverbs),

    Alex is the boy who kissed Mary;
  * Control (with **control verbs**, e.g. *begin*, *let*, *want*, etc.)

    Peter wants to Ø.ACT sleep;
  * Coreference with **verbal modifications that have dual dependency**,

    John saw Mary Ø.ACT run around the lake;
  * Coreference in constructions with **reciprocity**.

    John and Mary kissed Ø.PAT
* Textual coreference

  Helen asked her mother to wait for her but mother did not agree.

*Data subset under analysis

* a first half of the PCEDT section 19, particularly the 50 documents from wsj 1900 to wsj 1949
* manual annotation of word alignment

	English	Czech
Sentences	1078	1078
T-layer nodes	18611	20696
Coreferential	1362 (7.3%)	1440 (6.9%)
Grammatical	763 (56%)	568 (40%)
Textual	599 (44%)	872 (60%)

*Classes of inspected expressions

* Central pronouns
* Relative pronouns
* Anaphoric zeros

*Central pronouns

* Personal pronouns in the third person (*he*, *she*, *him*, *her*, etc.)
* Possessive pronouns (*his*, *her*, *mine*, etc.)
* Reflexive pronouns (*myself*, *themselves*, etc.)
* Reflexive possessive pronoun (*svůj*)
* English central pronouns that do not have their own representation on the t-layer (the pleonastic usage of the pronoun *it*, such as in *It is possible that*. . .)

!!! Central pronouns must be expressed on the surface!

*Relative pronouns

- Relative pronouns (*which*/*který*, etc.)
- Relative *that* with heuristics excluding subordinate conjunctions
- Relative adverbs that act like relative or interrogative pronouns (in English, e.g., *how*, *where*, *why* etc.), in Czech *kde* (*where*) and *kdy* (*when*)
- Relative pronouns which are not represented by its own node on the t-layer
- Numeral *kolik* (*how much*/*many*)

*Classes of inspected expressions

	English		Czech	
	covered	coreferential	covered	coreferential
Central pronouns	578	537 (93%)	286	284 (99%)
Relative pronouns	234	151 (65%)	341	302 (89%)
Anaphoric zeros	702	659 (94%)	850	777 (91%)
Total	1514	1350 (89%)	1477	1363 (92%)

We covered 99% and 95% of coreferential nodes in English and Czech, respectively.

*Original alignment

- *original alignmnet of t-nodes in PCEDT
- *unsupervised: GIZA++ [Och and Ney, 2000] run on a surface text in both directions and then projected onto the t-layer
- *+ rule-based alignment of nodes with already aligned parents sharing the same semantic role
- * this covered most unexpressed sujects
- *still many generated nodes remained uncovered

*The rule-based improvements

* quality of unsupervised alignment for function words and pronouns lower than for content words
* rule-based heuristics exploiting:
  * links between content words
  * gold annotation of t-trees: both structure and attributes
* set of rules designed for:
  * English central pronouns
  * Czech relative pronouns

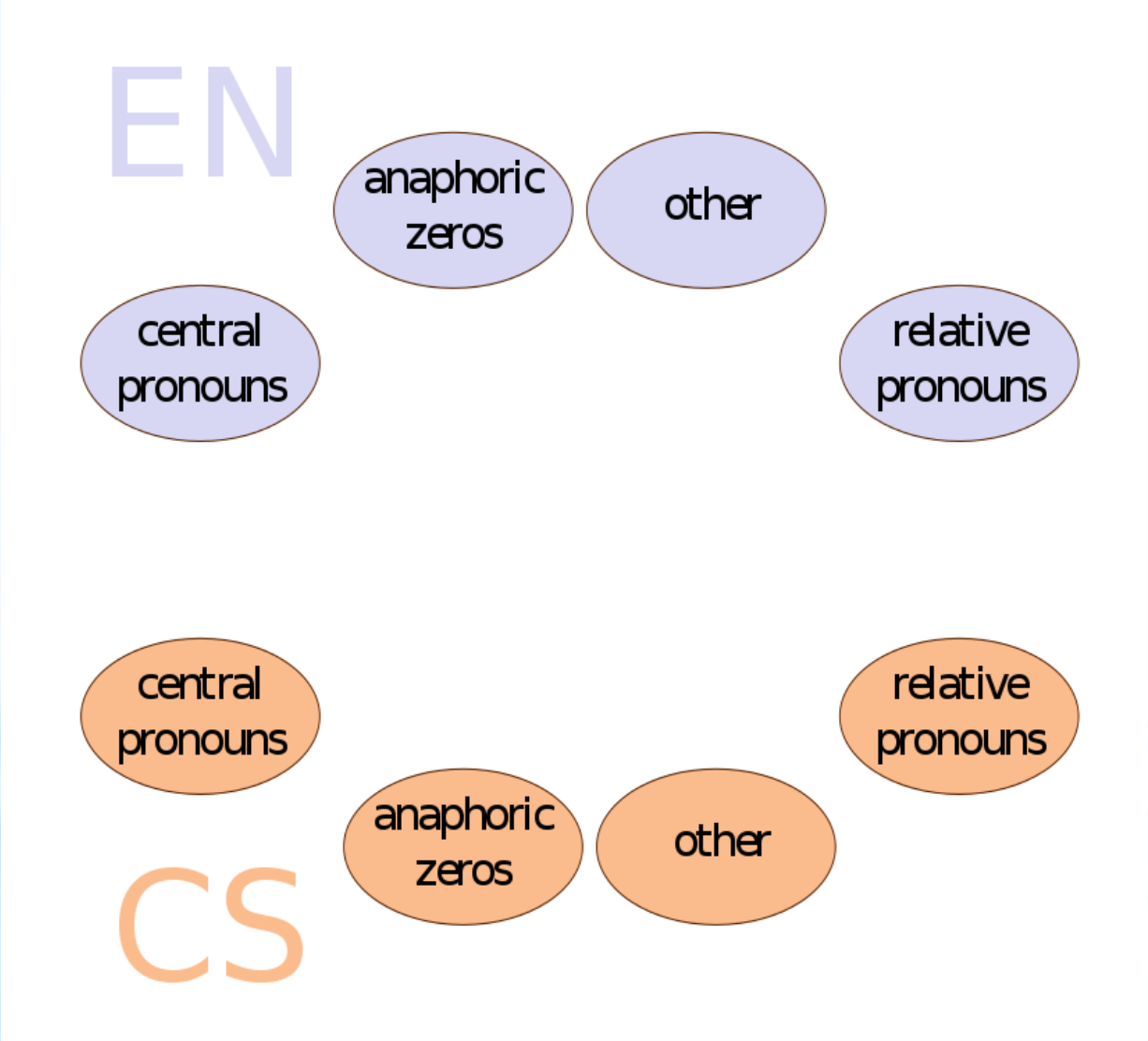
*Manual alignment

* two annotators
* annotated according these rules:
  1. Align with a direct translation of the source expression (direct alignment)
  2. Align with the translation of the source expression's antecedent, if there is no direct translation of the expression and the antecedent appears close enough to the expression (indirect alignment)
  3. Do not align, otherwise.
* sum of all CS and EN instances: 2991; only 2036 pairs necessary to annotate

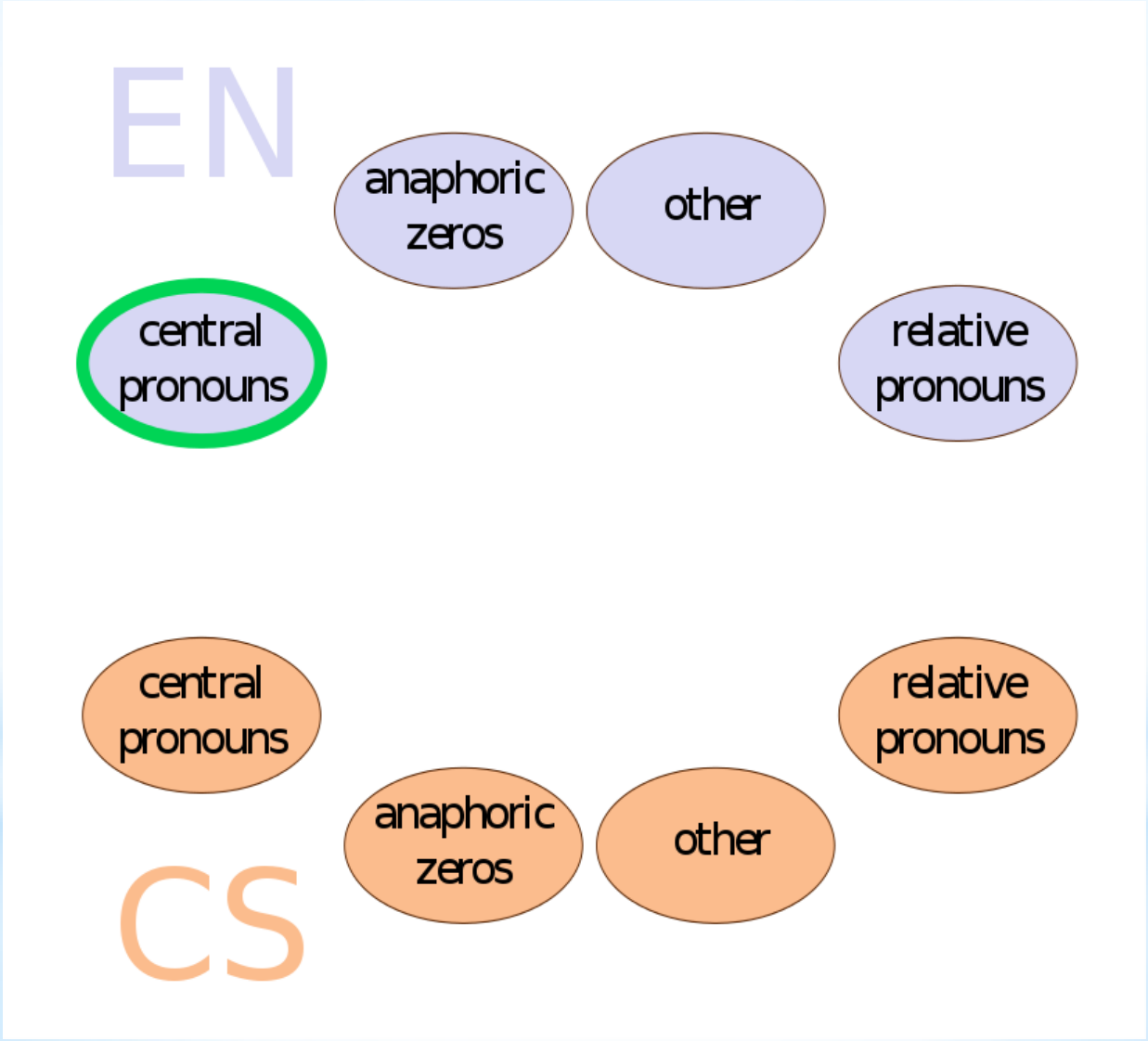
*Alignment quality evaluation

		CS				EN			
		A	P	R	F	A	P	R	F
Central pronouns	orig	88.11	93.80	89.02	91.35	76.47	83.15	80.21	81.65
	rule	89.16	94.26	90.20	92.18	83.74	88.15	88.33	88.24
Relative pronouns	orig	67.16	86.96	66.87	75.60	96.15	96.52	97.00	96.76
	rule	83.87	90.29	84.80	87.46	97.44	98.01	98.50	98.25
Anaphoric zeros	orig	78.71	98.89	71.18	82.78	75.93	98.60	62.58	76.57
	rule	81.76	98.75	75.32	85.46	79.63	99.03	68.37	80.90
Total	orig	77.86	94.40	73.76	82.82	79.26	90.62	76.17	82.77
	rule	83.68	95.16	81.02	87.52	83.95	93.55	82.20	87.51

Counterparts



Counterparts



*English central pronouns

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

English central pronouns: personal to zero

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

57% of English personal pronouns turn into Czech anaphoric zeros, most of them (99%) are in a subject position

EN: *He left a message accusing Mr. Darman of selling out.*
CS: *Ø Zanechal mu zprávu, ve které viní Darmana ze zaprodanosti.*

English central pronouns: personal to personal

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

14% of English personal pronouns turn into Czech personal pronouns, most of them in non-subject position, but still over 30% of them are subjects

EN: *Mr. Bush himself essentially acknowledged that **he** and his aides were trying to head off criticism.*
CS: *Bush sám v podstatě přiznal, že se **on** a jeho poradci snaží odvrátit kritiku.*

English central pronouns: personal to demonstrative

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

Czech demonstrative pronouns (represented solely by *ten*), are aligned with the pronoun *it* (in 99% cases)

EN: *It endorsed the White House strategy, believing it to be the surest way to victory.*
CS: *Ta přijala strategii Bílého domu v domnění, že je to nejjistější cesta k vítězství.*

English central pronouns: pleonastic *it*

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

EN: ***It*** wasn't known to what extent, if any, the facility was damaged.
CS: — Nebylo známo, do jaké míry, a jestli vůbec, bylo zařízení poškozeno.

English central pronouns: possessive to possessive

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

English possessives are mapped to Czech possessives (40%), Czech *svůj* (35%) or nothing (20%)

EN: *Mr. Bush himself essentially acknowledged that he and **his** aides were trying to head off criticism.*
CS: *Bush sám v podstatě přiznal, že se on a **jeho** poradci snaží odvrátit kritiku.*

English central pronouns: reflexive possessive *svůj*

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

EN: *While the book amply justifies **its** subtitle, the title itself is dubious.*
CS: *Zatímco **svůj** podtitul kniha dostatečně ospravedlňuje, samotný název je zavádějící.*

English central pronouns: possessive to nothing

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

EN: *As a result of **their** illness, they lost $1.8 million in wages and earnings.*
CS: *Důsledkem — nemoci, přišli na mzdách a výdělcích o 1.8 milionu dolarů.*

English central pronouns: possessive to dative

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

usually occupying the role of Benefactor or Adressee

EN: *Residents picked **their** way through glass-strewn streets.*
CS: *Obyvatelé města **si** razili cestu ulicemi zasypanými sklem.*

English central pronouns: reflexive basic

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

- **basic** and **emphatic** use of English reflexives [Quirk et al., 1985]
- [Novák et al., 2013]
  - basic – mapped to CS reflexives

EN: *The original is a comedy about Alceste, a man who sees falseness and vanity in everyone except **himself**.*
CS: *Původně to byla komedie o Alcestovi, muži, který vidíı ve všech kromě **sebe** faleš a marnivost.*

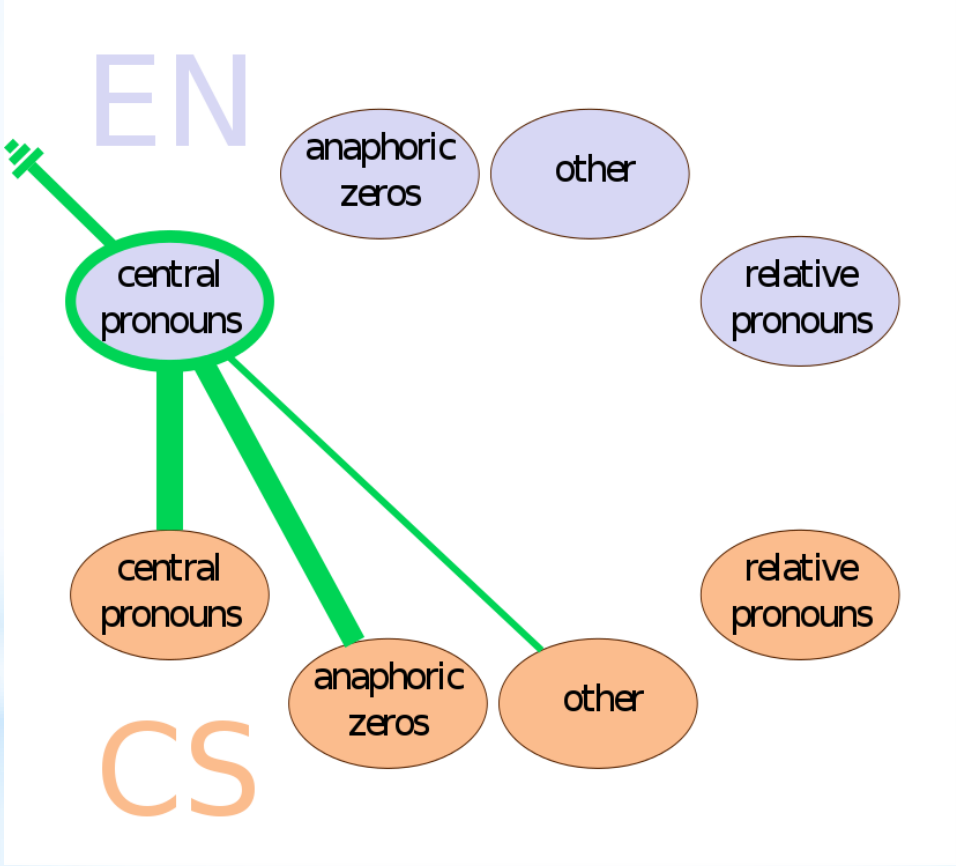
English central pronouns: reflexive basic

EN\CS	Aligned								Not aligned			Total
	pers	zero	poss	refl poss	refl	demon	noun	other	no poss	pleo	reword	
pers	49	190	3		1	21	18	7		29	16	334
poss	2	1	94	80	2		6	1	46		4	236
refl					3			8				11
Total	51	191	97	80	6	21	24	16	46	29	20	581

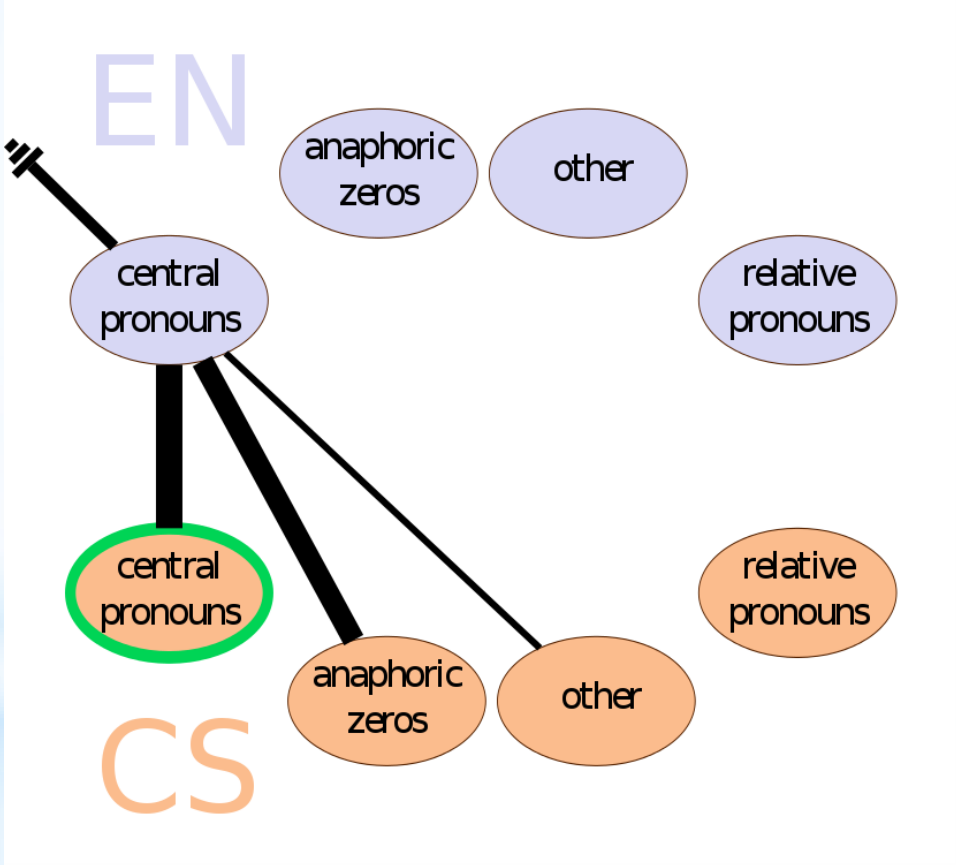
- **basic** and **emphatic** use of English reflexives [Quirk et al., 1985]
- [Novák et al., 2013]
  - basic – mapped to CS reflexives
  - emphatic – other means

EN: *As Mr. Bronner **himself** says, the smell of "raw meat" was in the air.*
CS: *Jak říká **sám** pan Bronner, ve vzduchu byl cítit zápach "syrového masa".*

*English central pronouns



*Czech central pronouns



Czech central pronouns

CS\EN	Aligned						Not aligned	Total
	pers	poss	refl	*the*	zero	other		
pers	49	2			7	2	4	64
poss	3	94		3		4	3	107
refl poss		80		3		3	4	90
refl	1	2	3		1	4	14	25
Total	53	178	3	6	8	13	25	286

Czech central pronouns: personal to zero

CS\EN	Aligned						Not aligned	Total
	pers	poss	refl	*the*	zero	other		
pers	49	2			7	2	4	64
poss	3	94		3		4	3	107
refl poss		80		3		3	4	90
refl	1	2	3		1	4	14	25
Total	53	178	3	6	8	13	25	286

non-finite constructions

CS: *Poslanec Bates prohlásil, že dopisy napíše tak, jak* **mu** *bylo nařízeno.*
EN: *Rep. Bates said he would write the letters as* ∅ *ordered.*

Czech central pronouns: possessives

CS\EN	Aligned						Not aligned	Total
	pers	poss	refl	*the*	zero	other		
pers	49	2			7	2	4	64
poss	3	94		3		4	3	107
refl poss		80		3		3	4	90
refl	1	2	3		1	4	14	25
Total	53	178	3	6	8	13	25	286

- mostly to EN poss (94 of 107 cases)
- definite article

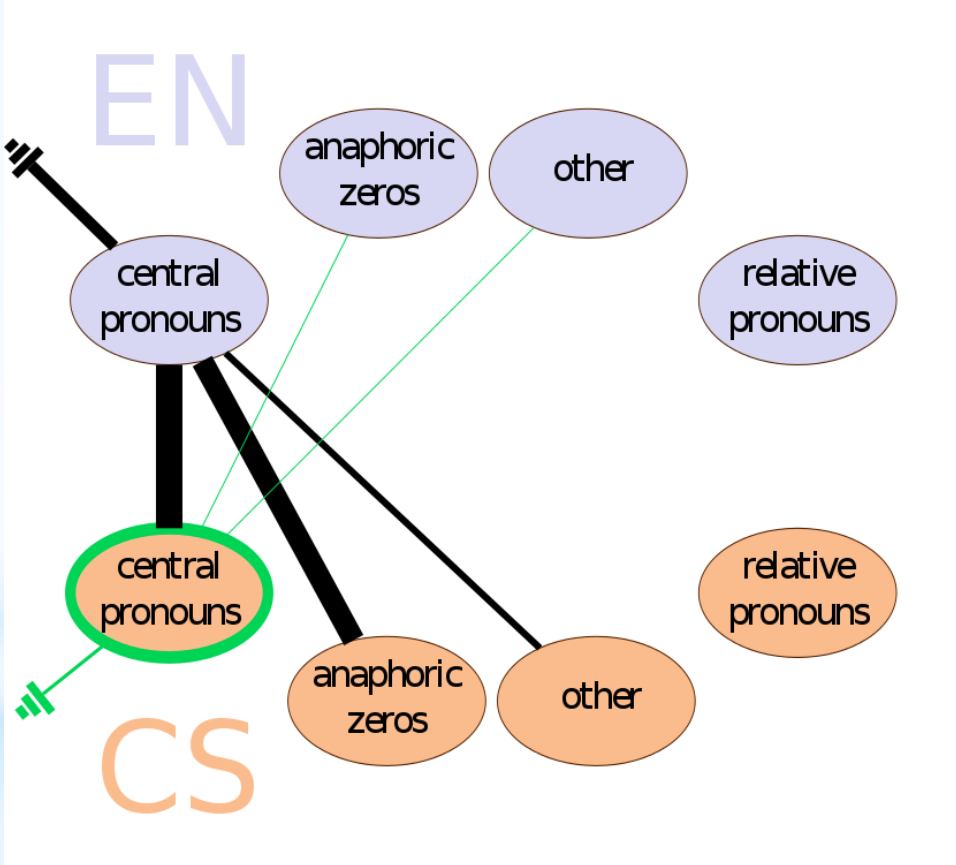
CS: *Tento maloobchodník nebyl schopen najít pro **svoji** budovu kupce.*
EN: *The retailer was unable to find a buyer for **the** building.*

Czech central pronouns: reflexives

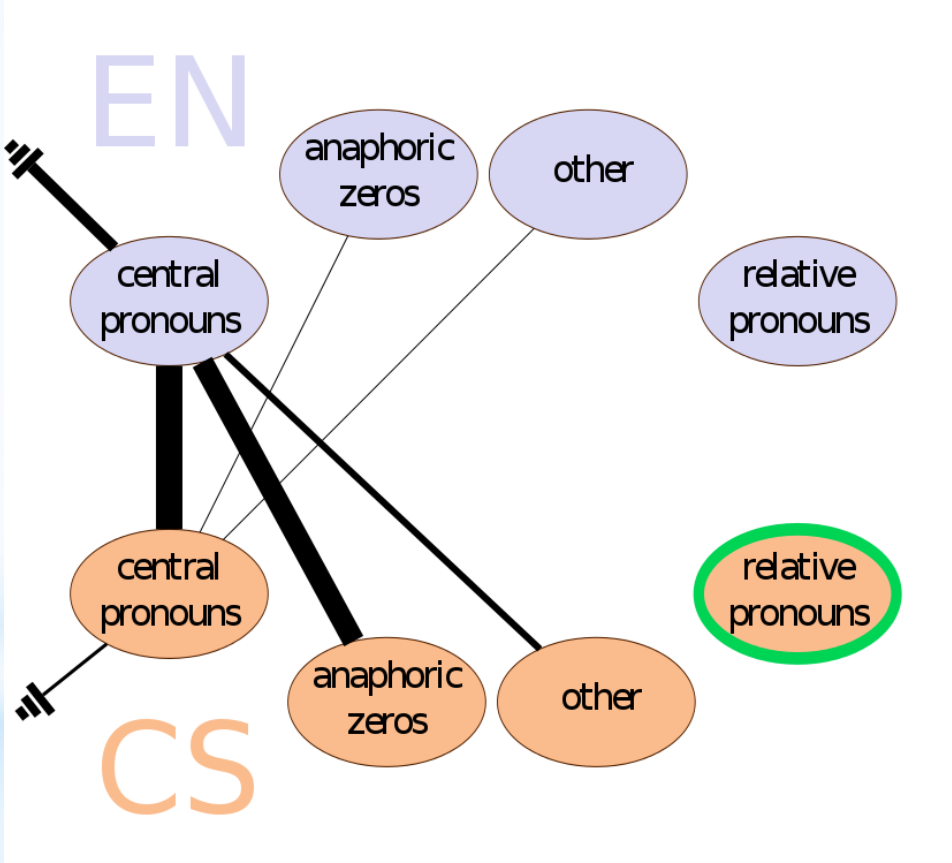
CS\EN	Aligned						Not aligned	Total
	pers	poss	refl	*the*	zero	other		
pers	49	2			7	2	4	64
poss	3	94		3		4	3	107
refl poss		80		3		3	4	90
refl	1	2	3		1	4	14	25
Total	53	178	3	6	8	13	25	286

CS: *Obyvatelé města **si** razili cestu ulicemi zasypanými sklem.*
EN: *Residents picked **their** way through glass-strewn streets.*

Czech central pronouns



Czech relative pronouns



Czech relative pronouns

CS\EN	Aligned						Not aligned			Total
	that	wh-word relat	wh-word inter & fused	zero	appos	pers	NP modif	VP modif	other	
což		7		4	15		2	6		34
other	51	102	23	71	2	1	42		15	307
Total	51	109	23	75	17	1	44	6	15	341

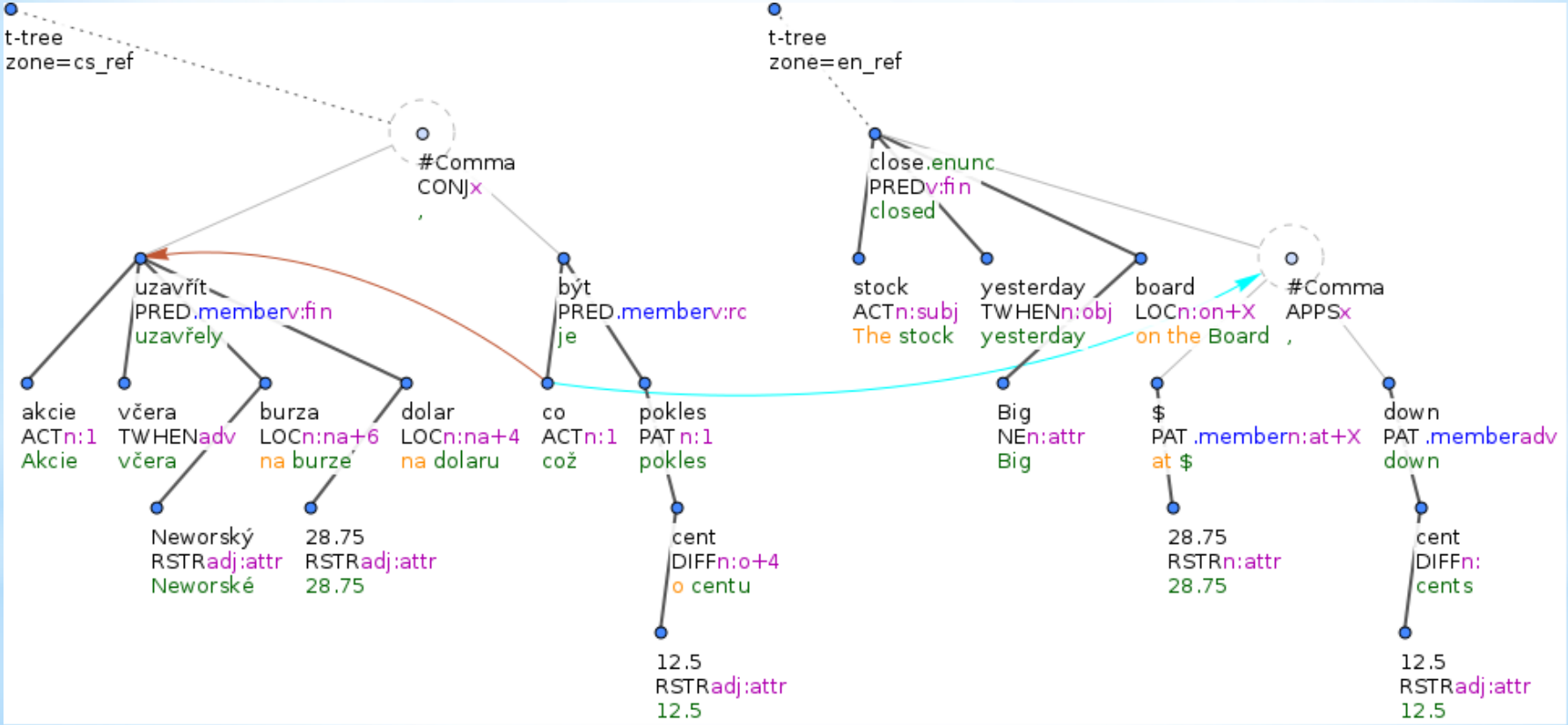
což: sentential relative clause
other: mostly adnominal relative clause

Czech relative pronouns: *což* to apposition

CS\EN	Aligned						Not aligned			Total
	that	wh-word relat	wh-word inter & fused	zero	appos	pers	NP modif	VP modif	other	
což		7		4	15		2	6		34
other	51	102	23	71	2	1	42		15	307
Total	51	109	23	75	17	1	44	6	15	341

CS: *Akcie včera uzavřely na Neworské burze na 28.75 dolaru, což je pokles o 12.5 centu.*
EN: *The stock closed yesterday on the Big Board at $28.75, down 12.5 cents.*

Czech relative pronouns: *což* to apposition



CS: *Akcie včera uzavřely na Neworské burze na 28.75 dolaru, což je pokles o 12.5 centu.*
EN: *The stock closed yesterday on the Big Board at $28.75, down 12.5 cents.*

Czech relative pronouns: relative to relative

CS\EN	Aligned						Not aligned			Total
	that	wh-word relat	wh-word inter & fused	zero	appos	pers	NP modif	VP modif	other	
což		7		4	15		2	6		34
other	51	102	23	71	2	1	42		15	307
Total	51	109	23	75	17	1	44	6	15	341

50% cases

CS: *Mohou se objevit síly, **které** tento scénář pozdrží.*
EN: *There may be forces **that** would delay this scenario.*

Czech relative pronouns: relative to zero

CS\EN	Aligned						Not aligned			Total
	that	wh-word relat	wh-word inter & fused	zero	appos	pers	NP modif	VP modif	other	
což		7		4	15		2	6		34
other	51	102	23	71	2	1	42		15	307
Total	51	109	23	75	17	1	44	6	15	341

23% cases: (a) ”zero relatives” or (b) non-finite clauses

CS: *To je otázka, **na níž** nemůže Východní Německo odpovědět snadno, at už jeho nový představitel udělá cokoli.*
EN: *That’s a question East Germany can’t answer easily, no matter what its new leader does.*

CS: *Zanechal mu zprávu, **ve které** viní Darmana ze zaprodanosti.*
EN: *He left a message ∅ accusing Mr. Darman of selling out.*

Czech relative pronouns: relative not aligned

CS\EN	Aligned						Not aligned			Total
	that	wh-word relat	wh-word inter & fused	zero	appos	pers	NP modif	VP modif	other	
což		7		4	15		2	6		34
other	51	102	23	71	2	1	42		15	307
Total	51	109	23	75	17	1	44	6	15	341

23% cases: (a) ”zero relatives” or (b) non-finite clauses

CS: *Dovoz, **který** tehdy činil šest milionů barelů denně, přicházel především z Venezuely a Kanady.*
EN: *Imports, **then six million barrels a day**, came primarily from Venezuela and Canada.*

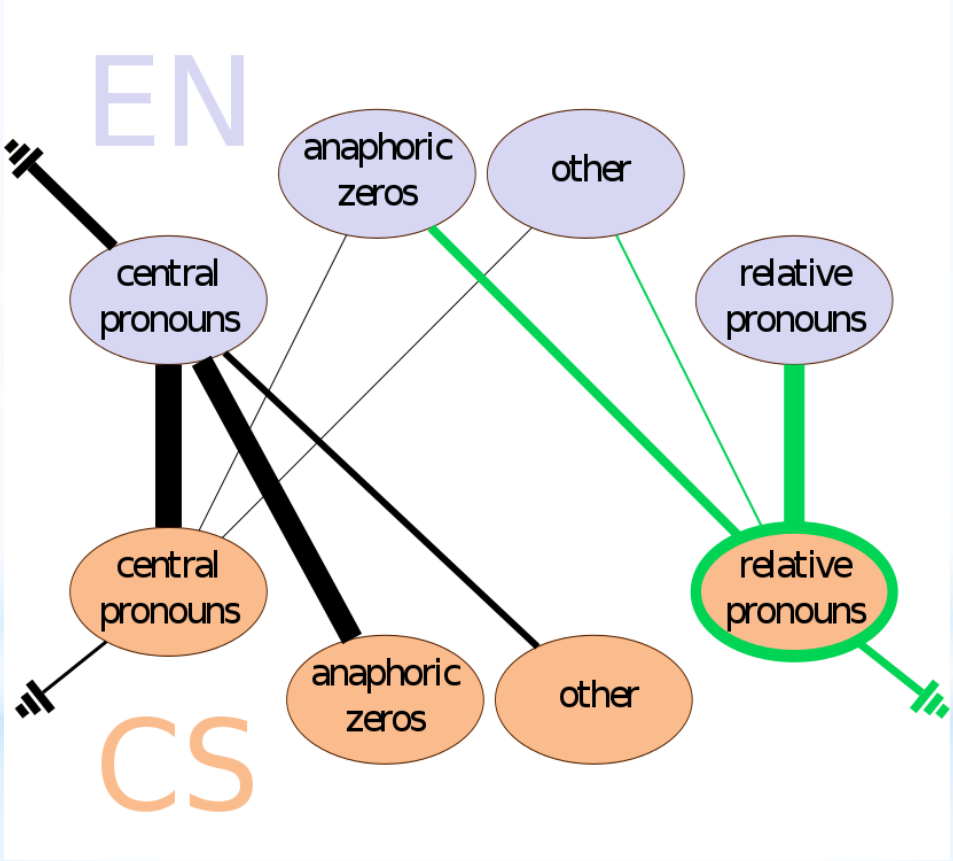
Czech relative pronouns: relative to fused or interrogative

CS\EN	Aligned						Not aligned			Total
	that	wh-word relat	wh-word inter & fused	zero	appos	pers	NP modif	VP modif	other	
což		7		4	15		2	6		34
other	51	102	23	71	2	1	42		15	307
Total	51	109	23	75	17	1	44	6	15	341

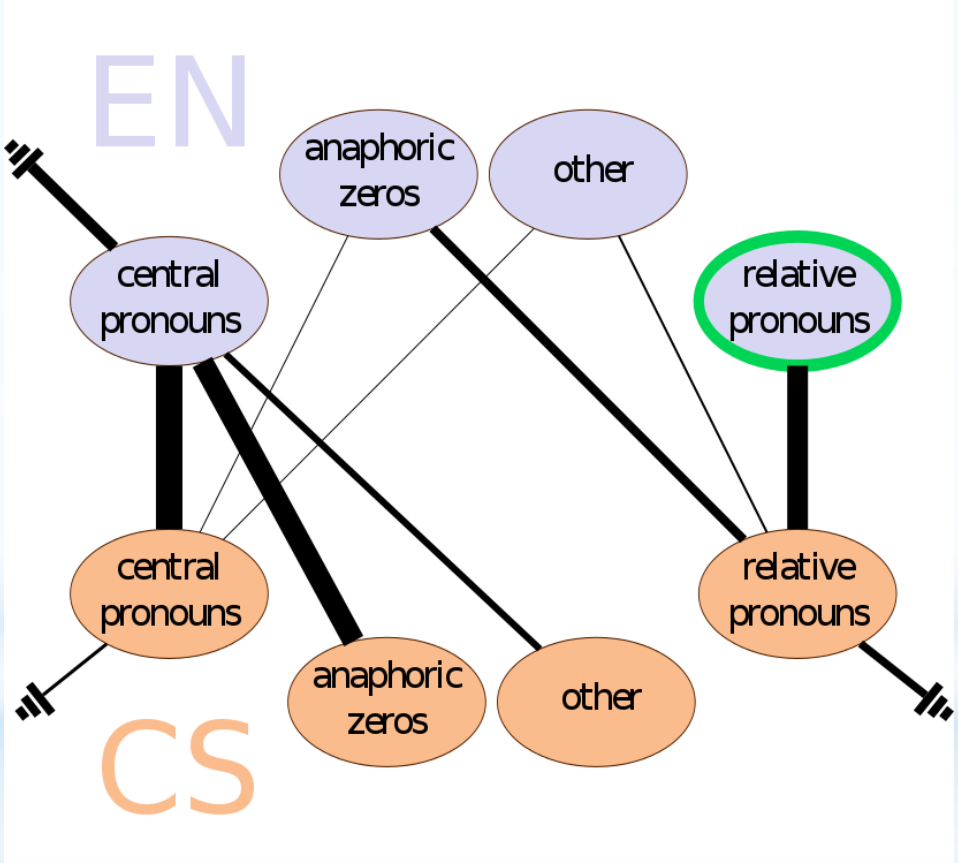
CS: *Na tom, co ∅ máme, je třeba udělat hodně práce.*
EN: *There is plenty of work to be done on what we have.*

CS: *Nebylo jasné, kdy se znovu obnoví normální tempo 750 vozů za den.*
EN: *It wasn't clear when will resume the normal pace 750-car-a-day.*

Czech relative pronouns



English relative pronouns



English relative pronouns

EN\CS	Aligned				Not aligned	Total
	což	other relat	conj	other		
that		49		1	6	56
wh-words relat	7	102	2		7	118
wh-words inter & fused		23		14	6	43
wh-words conj			16		1	17
Total	7	174	18	15	20	234

English relative pronouns: *that* and relative

EN\CS	Aligned				Not aligned	Total
	což	other relat	conj	other		
that		49		1	6	56
wh-words relat	7	102	2		7	118
wh-words inter & fused		23		14	6	43
wh-words conj			16		1	17
Total	7	174	18	15	20	234

alignments between similar categories – 68% of all instances

EN: *There may be forces **that** would delay this scenario.*
CS: *Mohou se objevit síly, **které** tento scénář pozdrží.*

English relative pronouns: interrogative and fused

EN\CS	Aligned				Not aligned	Total
	což	other relat	conj	other		
that		49		1	6	56
wh-words relat	7	102	2		7	118
wh-words inter & fused		23		14	6	43
wh-words conj			16		1	17
Total	7	174	18	15	20	234

- other relat – 43% of them translated with correlative pairs
- other – mostly *proč* and *jak*

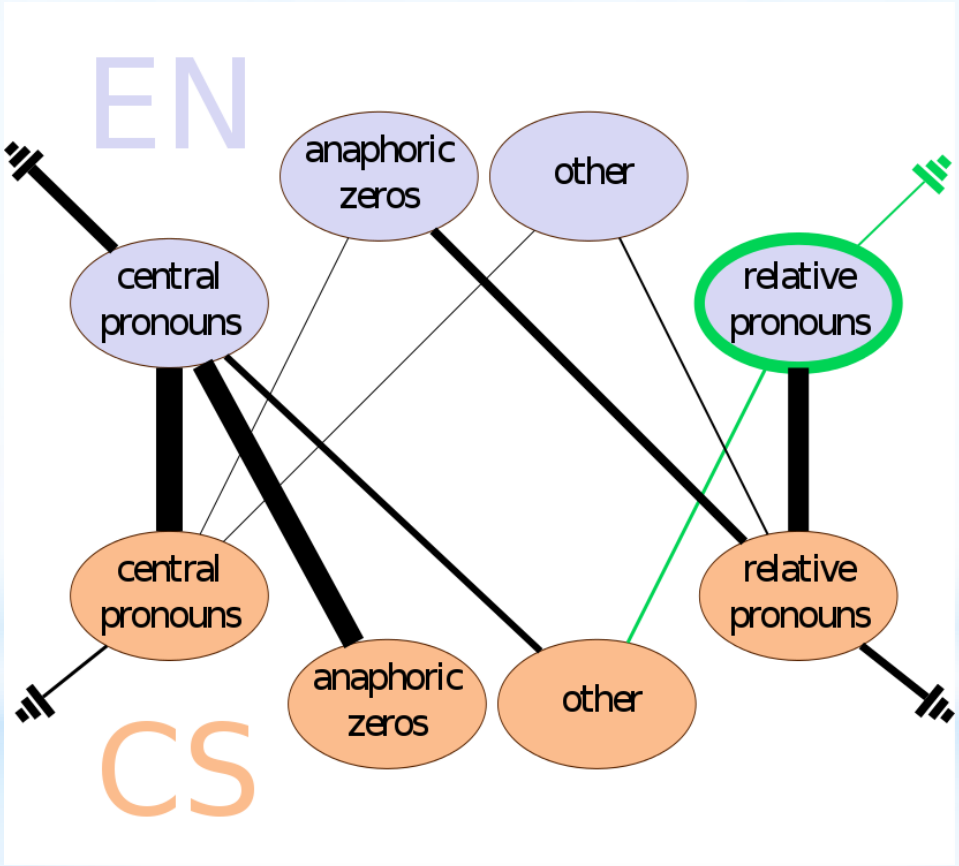
EN: *There is plenty of work to be done on what we have.*
CS: *Na tom, co máme, je třeba udělat hodně práce.*

English relative pronouns: subordinating conjunction

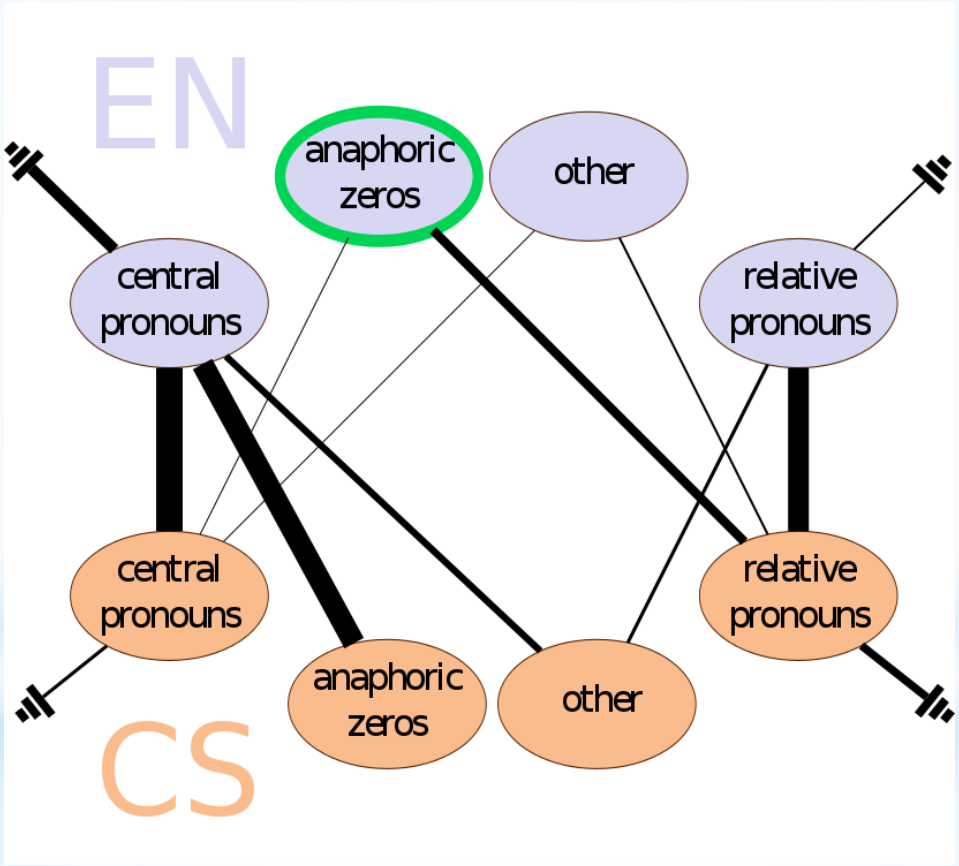
EN\CS	Aligned				Not aligned	Total
	což	other relat	conj	other		
that		49		1	6	56
wh-words relat	7	102	2		7	118
wh-words inter & fused		23		14	6	43
wh-words conj			16		1	17
Total	7	174	18	15	20	234

EN: *In 1956, **when** Britain, France and Israel invaded Egypt, Arab producers cut off supplies to Europe.*
CS: *V roce 1956, **když** Británie, Francie a Izrael napadly Egypt, zastavili arabští výrobci dodávky do Evropy.*

English relative pronouns



English anaphoric zeros



English anaphoric zeros

EN\CS	Aligned				Not aligned	Total
	zero	relat	pers	other		
zero	263	75	7	28	329	702

English anaphoric zeros: zero to zero

EN\CS	Aligned				Not aligned	Total
	zero	relat	pers	other		
zero	263	75	7	28	329	702

EN: *Their reaction was to ø.ACT do nothing and ø.ACT ride it out.*
CS: *Jejich reakcí bylo ø.ACT nedělat nic a ø.ACT nechat to odeznít.*

English anaphoric zeros: zero to relative

EN\CS	Aligned				Not aligned	Total
	zero	relat	pers	other		
zero	263	75	7	28	329	702

About 10% of English anaphoric zeros correspond to Czech relative pronouns

EN: *He left a message ∅ accusing Mr. Darman of selling out.*
CS: *Zanechal mu zprávu, **ve které** viní Darmana ze zaprodanosti.*

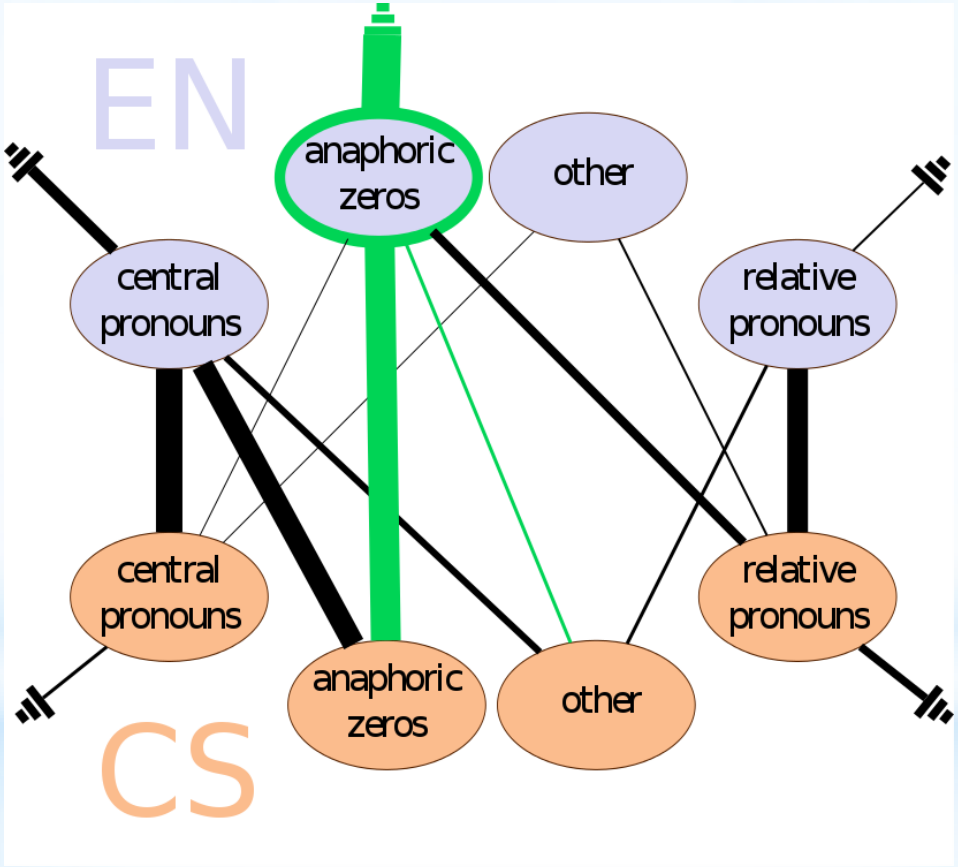
English anaphoric zeros: not aligned

EN\CS	Aligned				Not aligned	Total
	zero	relat	pers	other		
zero	263	75	7	28	329	702

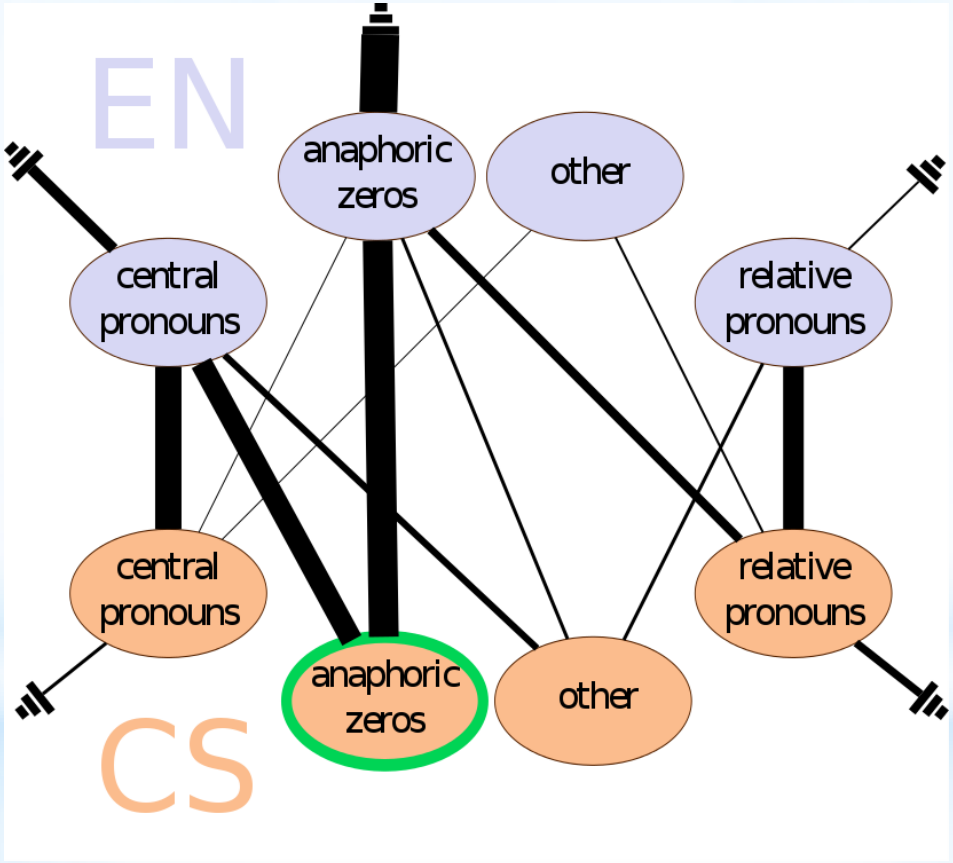
Almost 50% of anaphoric zeros in English have not found their Czech counterparts (rewording, missing argument, technical reasons)

EN: *I want to ø.ACT publish one that succeeds.*
CS: *Já chci vydávat takový, který uspěje.*

English anaphoric zeros



Czech anaphoric zeros



Czech anaphoric zeros

CS\EN	Aligned					Not aligned	Total
	zero	pers	pers 1st & 2nd	poss	other		
zero	263	190	40	1	84	278	856

Czech anaphoric zeros: zeros to personal

CS\EN	Aligned					Not aligned	Total
	zero	pers	pers 1st & 2nd	poss	other		
zero	263	190	40	1	84	278	856

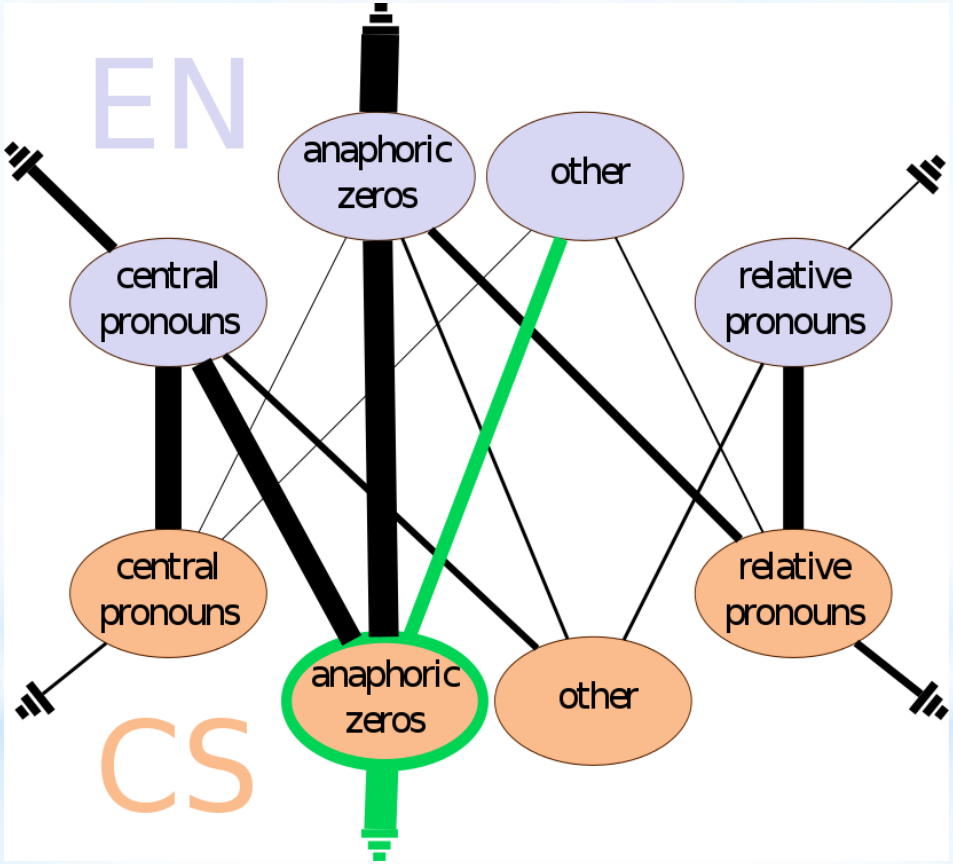
CS: *Ø Zanechal mu zprávu, ve které viní Darmana ze zaprodanosti.*
EN: ***He*** *left a message accusing Mr. Darman of selling out.*

Czech anaphoric zeros: zeros to personal (1st&2nd pers)

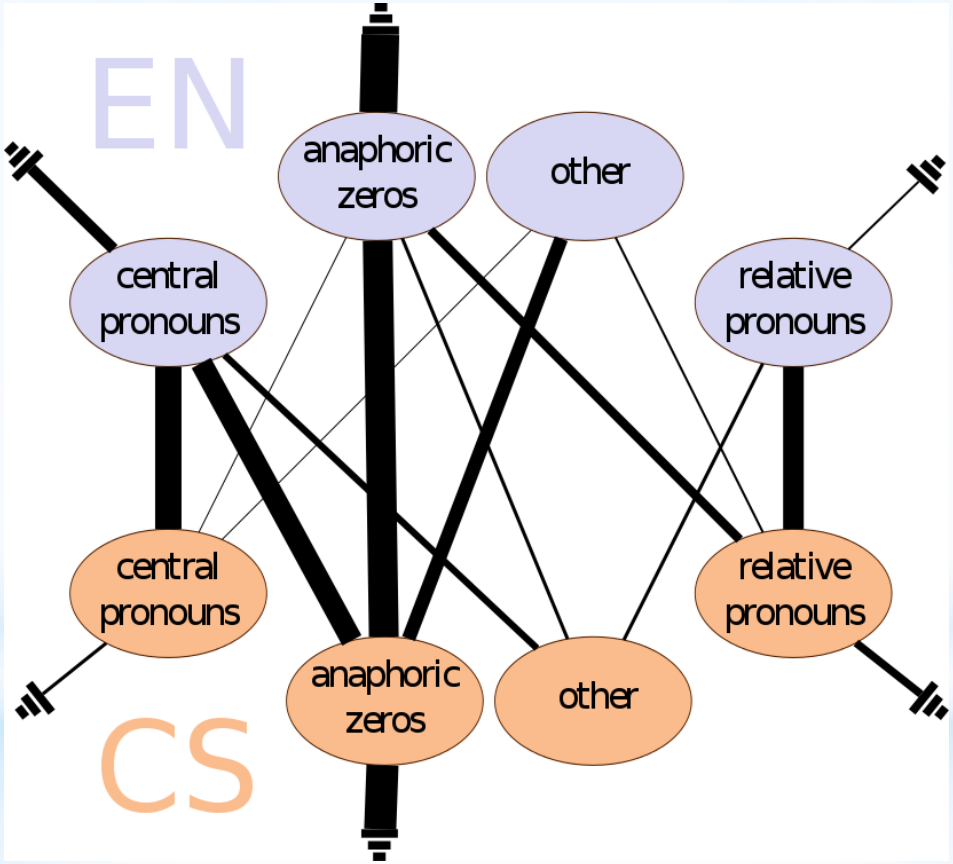
CS\EN	Aligned					Not aligned	Total
	zero	pers	pers 1st & 2nd	poss	other		
zero	263	190	40	1	84	278	856

CS: *Ø Nemáme pasivní čtenáře.*
EN: *We don't have passive readers.*

Czech anaphoric zeros



Counterparts



Discussion

* Possessivity

EN: *As a result of **their** illness, they lost $1.8 million in wages and earnings.*
CS: *Důsledkem (**své**) nemoci, přišli na mzdách a výdělcích o 1.8 milionu dolarů.*

  * dative possessors

CS: *Čeští reformátoři **si** ve **své** zemi mohou ze stejné doby připomenout Wilsonovy ideály.*
CS: *Czech reformers can recall the Wilsonian ideals of the same period in **their** country.*

* Pro-drop character of Czech
* Factor of translated text

HUMAN CAPITAL
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

Conclusion

- improved alignment of coreferential expressions + manual annotation
- comprehensive analysis of mappings between the expressions in Czech and English
- Future work:
  - use the improved alignment in the new version od PCEDT
  - analyze the alignment of antecedents
  - co-training with language-dependent feature sets
  - translation via deep syntax (TectoMT)

HUMAN CAPITAL
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

*Acknowledgements

* The presentation of the results is co-financed by the European Union from resources of the European Social Fund

* The research was supported from the Grant Agency of the Czech Republic (grant P406/12/0658 Coreference, discourse relations and information structure in a contrastive perspective), GAUK 3389/2015, EU (grant FP7-ICT-2013-10-610516 – QTLeap) and SVV project number 260 224. This work has been using language resources developed, stored, and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

*Bibliography I

Ariel, M. (2015). Accessibility theory: An overview. *Text representation*, pages 29–87.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pages 92–100, New York, NY, USA. ACM.

de Souza, J. G. C. and Orăsan, C. (2011). Can projected chains in parallel corpora help coreference resolution? In *Proceedings of the 8th International Conference on Anaphora Processing and Applications*, Berlin, Heidelberg. Springer-Verlag.

Givón, T., editor (1983). *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, volume 3 of *Typological Studies in Language*. John Benjamins Publishing.

Hajičová, E., Hladká, B., and Kučová, L. (2006). An annotated corpus as a test bed for discourse structure analysis. In *Proceedings of the Workshop on Constraints in Discourse*, pages 82–89, Maynooth, Ireland. National University of Ireland, National University of Ireland.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association.

Kibrik, A. A. (1997). Modelling multi-factorial processes: Referential choice in russian discourse. In *Vestnik MGU 1997.4*, pages 94–105.

Kibrik, A. A. (2011). *Reference in Discourse*. Oxford University Press.

Kunz, K. and Lapshinova-Koltunski, E. (2015). Cross-linguistic analysis of discourse variation across registers. *Special Issue of Nordic Journal of English Studies*.

Novák, M. and Žabokrtský, Z. (2014). Cross-lingual coreference resolution of pronouns. In Tsujii, J. and Hajič, J., editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 14–24, Dublin, Ireland. Dublin City University, Dublin City University and Association for Computational Linguistics.

HUMAN CAPITAL
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

*Bibliography II

Novák, M., Žabokrtský, Z., and Nedoluzhko, A. (2013). Two case studies on translating pronouns in a deep syntax framework. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1037–1041, Nagoya, Japan. Asian Federation of Natural Language Processing.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Onderková, K. (2009). Possessive pronouns in english and czech works of fiction, their use with parts of human body and translation. Master's thesis, Masaryk University, Faculty of Arts, Brno.

Popel, M. and Žabokrtský, Z. (2010). TectoMT: Modular NLP framework. In *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233, Berlin / Heidelberg. Springer.

Postolache, O., Cristea, D., and Orasan, C. (2006). Transferring coreference chains through word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy. European Language Resources Association.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Veselovská, K., Nguy, G. L., and Novák, M. (2012). Using Czech-English parallel corpora in automatic identification of "it". In *The Fifth Workshop on Building and Using Comparable Corpora*, İstanbul, Turkey. European Language Resources Association.

Zinsmeister, H., Dipper, S., and Seiss, M. (2012). Abstract pronominal anaphors and label nouns in german and english: selected case studies and quantitative investigations. *Translation: Computation, Corpora, Cognition*, 2(1).