

Contrasting Coreference In Czech And German

From Different Frameworks to Joint Results

or
Russians about the languages they live in

Екатерина Лапинова-Колтунски, Аня Недолужко

Motivation

- compare languages
- area of discourse phenomena
 - coreference
 - coreference chains
- variation of coreference chains reflects cohesion
- take advantage of the existing resources reflecting systemic peculiarities of Czech and German
- use Czech and German data annotated on the basis of two different frameworks

Aims

analyse differences between German and Czech with respect to coreference

GERMAN



CZECH

- ☐ reduced morphology
- ☐ more isolating
- ☐ less flexible word order
- ☐ definite article

- ☐ morphologically rich
- ☐ less isolating
- ☐ flexible word order
- ☐ no definite article

What we have:

- two large-scale annotated corpora

GECCO



- comparable and parallel texts in English and German (ca. 80,000 sentences)
- written and spoken texts
- morpho-syntactic annotation
- cohesive devices annotated:
 - reference
 - lexical cohesion
 - discourse relations
 - connectives
 - ellipsis
 - substitution

PDT 3.0



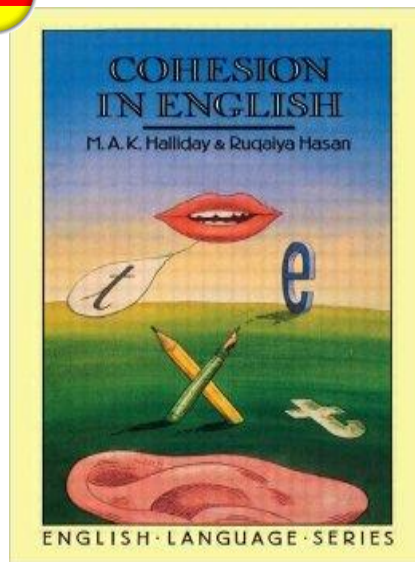
- journalistic written texts in Czech with further genre classification (ca. 50,000 sentences)
- texts in Czech
- multilayer annotation: morphological, analytical and tectogrammatical (dependency trees)
- textual phenomena annotated:
 - coreference
 - bridging relations
 - discourse relations
 - information structure
 - ellipsis

What we have:

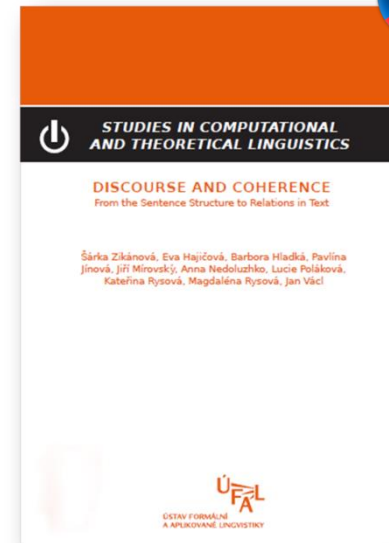
- two large-scale annotated corpora
- two different theoretical approaches

What we have:

- two large-scale annotated corpora
- two different theoretical approaches

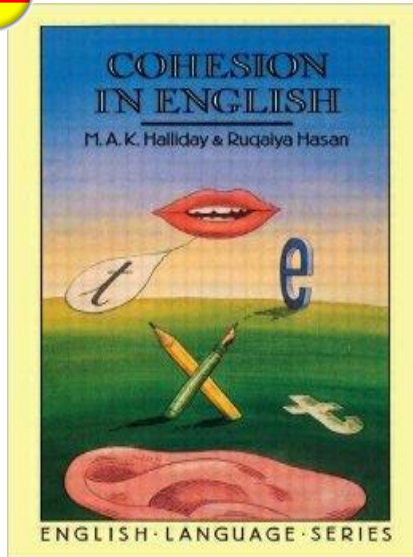


Halliday and Hasan, 1976

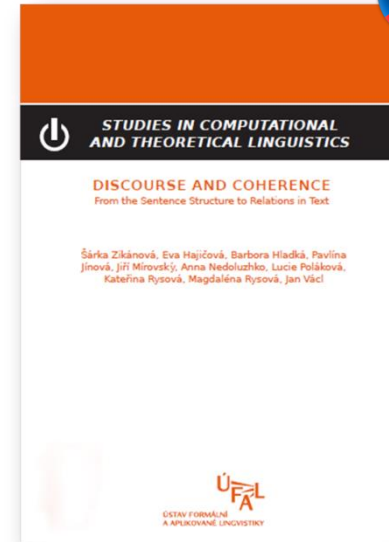


Sgall 1986, Functional grammar,
Zikánová et al., 2015
PDTB-like discourse

What we did:



Halliday and Hasan, 1976



Sgall 1986, Functional grammar,
Zikánová et al., 2015
PDTB-like discourse

MUTUAL RESULTING SCHEMA

What we have got:

MUTUAL RESULTING SCHEMA

	featID	Czech	German
IDENTITY	id1	coreference with pronouns	coreference with heads (no extended reference)
	id2	pronouns with arrows to segments and events	reference to verb phrases and longer segments
	id3	NP coreference	coreference with modifiers or def.articles
	id4	coreference with the word <i>same</i>	general comp.reference
	id5	coreference with local and temporal adverbs	coreference with local and temporal adverbs
NON-IDENTITY	nonid1	relations of MERONYMY	relations of MERONYMY
	nonid2	bridging CONTRAST	particular comparative reference and antonyms
DISCOURSE RELATIONS	temp	temporal	temporal
	cont	contingency	causal
	comp	comparison (contrast)	adversative
	expan	expansion	additive
ELLIPSIS	ellipsis	textual ellipsis	cohesive ellipsis

Lapshinova-Nedoluzhko-Kunz, 2015, NAACL, LAW, Denver

Data

- written discourse
- essays
- comparable topics (economic, political and social issues)

Data excerption

GERMAN

	Topics	sent	tokens
1	Germany and social market economy	121	2035
2	Optimistic remarks on globalisation	47	971
3	Politics and globalisation	103	1871
4	Globalisation and new challenges	27	478
5	The biggest currency changeover	85	1460
6	Globalisation and market economy	80	1782
7	Global market and technical progress	108	1851
8	Economic and technological changes	73	1795
	Total: ALL GERMAN	645	12243

CZECH

id	topics	sent	tokens
1-5	Germany, politics and history	170	687
6	Housing	83	1644
7	Technological changes	73	1795
8-9	Politics	121	1854
10-13	Economics	149	2568
14	Unemployment	112	2252
15	Television	55	969
	Total: ALL CZECH	628	11399

Questions:

- Are there any language contrasts with respect to coreference chain features between German and Czech?
- Are there any differences between German and Czech if abstract coreference is concerned?

Features to answer:

- longest chains (chains with the greatest number of elements in a text),
- number of coreference pairs (= number of anaphorically referring expressions),
- types of anaphors (e.g. pronouns, nominal modifiers, temporal and local adverbs and so on),
- number of anaphora referring to antecedents other than NP/pronoun,
- types of anaphora referring to antecedents other than NP/pronoun

Chain properties in German and Czech

	GERMAN	CZECH
number of coreference chains	225	550
length of coreference chains (mean value per text)	2.45	3.2
longest chain	11	27
number of coreference relations	327	1231

Chain properties in German and Czech

	GERMAN	CZECH
number of coreference chains	225	550
length of coreference chains (mean value per text)	2.45	3.2
longest chain	11	27
number of coreference relations	327	1231

still schemes differences

Chain properties in German and Czech

	GERMAN	CZECH
number of coreference chains	225	550
length of coreference chains (mean value per text)	2.45	3.2
longest chain	11	27
number of coreference relations	327	1231
	996	1027
	+ repetitions	- zeros

Chain properties in German and Czech

	GERMAN	CZECH
number of coreference chains	225	550
length of coreference chains (mean value per text)	2.45	3.2
longest chain	11	27
number of coreference relations	327	1231

Longest chains in German and Czech

GERMAN

mostly contain

- personal and demonstrative pronouns
- definite nominal phrases

CZECH

mostly contain

- hyperthematic named entities

*lexical cohesion
in GECCO*



(1) Úvahy padesát let od SNP

(3) Povstání a vůbec celou antifašistickou rezistenci považují za významnou událost, která podala nefalšovaný obraz Slovenska, řekl v srpnu 1994 v rozhovoru pro slovenský deník Pravda předseda Slovenského svazu protifašistických bojovníků generál Ján Husák.

(4) Také nejvýznamnější slovenští politici od prezidenta přes premiéra až po předsedu parlamentu považují Slovenské národní povstání za pozitivní jev ve slovenských dějinách.

(5) Dnes je už známé, že historie SNP byla překroucena ve prospěch komunistické ideologie.

(6) Po listopadu 1989 sice tento černobílý pohled zmizel, ale vakuum využili převážně exiloví historici a spojili obhajobu Slovenského státu s odsouzením SNP.

(7) Oprostit SNP od nánosu lží a mýtů si kladla za cíl první vědecká konference o SNP, která se uskutečnila letos v dubnu v Donovalech.

(8) S blížícím se 50. výročím SNP dochází na politické scéně i ve slovenském tisku k souboji dvou diametrálně odlišných koncepcí výkladu SNP: na straně jedné se povstání označuje za puč proti vlastní státnosti a na straně druhé je SNP chápáno jako vědomé odmítnutí totality a fašismu.

(9) Podle názoru místopředsedkyně Slovenské národní strany (SNS) Evy Slavkovské přineslo SNP slovenskému národu jen to, že povstal se zbraní v ruce proti fašismu a nepřijal ho trpně jako český národ.

(10) Ø Vzalo mu však mnohem víc: vlastní samostatnost.

(11) Podle komentáře slovenského levicového listu Pravda povstání bránilo hodnoty, na kterých Evropa stála a stojí.

(12) SNS a další národně orientované strany (např. Křesťanská sociální unie, Společnost dr.Jozefa Tisa, Slovenská národní jednota) se přiklání k názoru, že šlo o puč proti vlastní státnosti.

(13) Podle představitelů Slovenské národní jednoty Češi v SNP obrali Slováky o vlastní stát a na 40 let jim vnutili čechokomunistický režim.

(14) Strana demokratické levice po přehodnocení faktu, že vedoucí silou SNP byla komunistická strana, se jednoznačně staví za antifašistický a demokratický odkaz povstání.

(15) O postoji Hnutí za demokratické Slovensko (HZDS) k SNP se dá říci, že je přinejmenším nejasný.

(16) Na loňských oslavách 49. výročí SNP předseda HZDS a bývalý premiér Vladimír Mečiar sice zvůli.

(17) Ani postoj Ø Křesťanskodemokratického hnutí (KDH) nebyl vždy jasný.

(19) Představitelé KDH však neodmítli dialog a dnes považují povstání za významnou kapitolu slovenské historie.

(20) Slovenské národní povstání politicky organizovala Slovenská národní rada, která vznikla již v září 1943 a zastupovala komunistický a občanský odboj.

(22) Podle Vánoční dohody bylo cílem ozbrojeného boje obnovení Československé republiky.

(24) Během SNP působily na území Slovenska vojenské mise Spojenců, a to sovětská, britská a americká.

(27) Povstání zkomplikovalo plány nacistů na obranu Karpat, připoutalo nemalé nacistické jednotky na Slovensko, rozrušilo komunikační systém v týlu fronty a vytvořilo nebezpečné ohnisko nepokojů.

(28) Ø Posilovalo také odpor v sousedních zemích.

(29) Pro Slovensko Ø mělo význam také politický: Ø zlikvidovalo chápání Slovenska na mezinárodní scéně jako německého vazala.

(1) Úvahy padesát let od **SNP**

(3) **Povstání** a vůbec celou antifašistickou rezistenci považují za významnou událost, **která** podala nefalšovaný obraz Slovenska, řekl v srpnu 1994 v rozhovoru pro slovenský deník Pravda předseda Slovenského svazu protifašistických bojovníků generál Ján Husák.

(4) Také nejvýznamnější slovenští politici od prezidenta přes premiéra až po předsedu parlamentu považují **Slovenské národní povstání** za pozitivní jev ve slovenských dějinách.

(5) Dnes je už známé, že historie **SNP** byla překroucena ve prospěch komunistické ideologie.

(6) Po listopadu 1989 sice tento černobílý pohled zmizel, ale vakuum využili převážně exiloví historici a spojili obhajobu Slovenského státu s odsouzením **SNP**.

(7) Oprostit **SNP** od nánosu lži a mýtů si kladla za cíl první vědecká konference o **SNP**, **která** se uskutečnila letos v dubnu v Donovalech.

(8) S blížícím se 50. výročím **SNP** dochází na politické scéně i ve slovenském tisku k souboji dvou diametrálně odlišných koncepcí výkladu **SNP**: na straně jedné se **povstání** označuje za **puč** proti vlastní státnosti a na straně druhé je **SNP** chápáno jako vědomé odmítnutí totality a fašismu.

(9) Podle názoru místopředsedkyně Slovenské národní strany (SNS) Evy Slavkovské přineslo **SNP** slovenskému národu jen to, že povstal se zbraní v ruce proti fašismu a nepřijal ho trpně jako český národ.

(10) **Ø** Vzalo mu však mnohem víc: vlastní samostatnost.

(11) Podle komentáře slovenského levicového listu Pravda **povstání** bránilo hodnoty, na kterých Evropa stála a stojí.

(12) SNS a další národně orientované strany (např. Křesťanská sociální unie, Společnost dr.Jozefa Tisa, Slovenská národní jednota) se přiklání k názoru, že šlo o **puč** proti vlastní státnosti.

(13) Podle představitelů Slovenské národní jednoty Češi v **SNP** obrali Slováky o vlastní stát a na 40 let jim vnutili čechokomunistický režim.

(14) Strana demokratické levice po přehodnocení faktu, že vedoucí silou **SNP** byla komunistická strana, se jednoznačně staví za antifašistický a demokratický odkaz **povstání**.

(15) O postoji Hnutí za demokratické Slovensko (HZDS) k **SNP** se dá říci, že je přinejmenším nejasný.

(16) Na loňských oslavách 49. výročí **SNP** předseda HZDS a bývalý premiér Vladimír Mečiar sice zvůli.

(17) Ani postoj **Ø** Křesťanskodemokratického hnutí (KDH) nebyl vždy jasný.

(19) Představitelé KDH však neodmítli dialog a dnes považují **povstání** za významnou kapitolu slovenské historie.

(20) **Slovenské národní povstání** politicky organizovala Slovenská národní rada, která vznikla již v září 1943 a zastupovala komunistický a občanský odboj.

(22) Podle Vánoční dohody bylo cílem ozbrojeného boje obnovení Československé republiky.

(24) Během **SNP** působily na území Slovenska vojenské mise Spojenců, a to sovětská, britská a americká.

(27) **Povstání** zkomplikovalo plány nacistů na obranu Karpat, připoutalo nemalé nacistické jednotky na Slovensko, rozrušilo komunikační systém v týlu fronty a vytvořilo nebezpečné ohnisko nepokojů.

(28) **Ø** Posilovalo také odpor v sousedních zemích.

(29) Pro Slovensko **Ø** mělo význam také politický: **Ø** zlikvidovalo chápání Slovenska na mezinárodní scéně jako německého vazala.

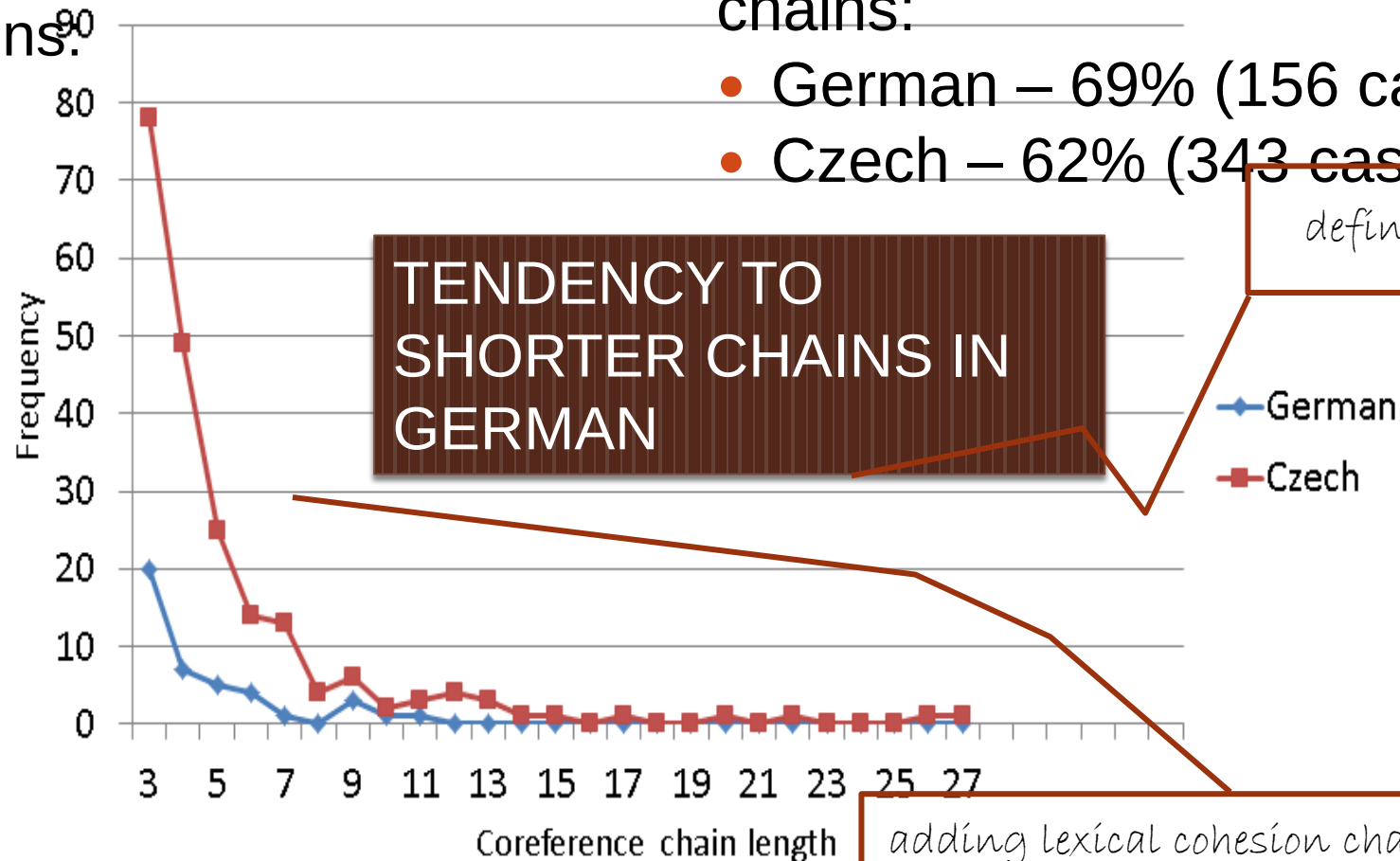
Das deutsche Modell **<der sozialen Marktwirtschaft>** ist eine ökonomische Erfolgsgeschichte : **<Sie>** verbindet wirtschaftlichen Wohlstand und soziale Gerechtigkeit auf eine Weise, die ihresgleichen sucht . Auch wenn einige **<ihrer Errungenschaften>** der Reform bedürfen - **<sie>** hat schon mehrfach bewiesen, dass **<sie>** die Kraft hat , Umbrüche zu bewältigen. Arbeiten wie die Polen , aber leben wie die Japaner - die Ökonomen der westlichen Welt lächelten milde über die Ideen der östlichen Nachbarn. Nur der französische Wirtschaftsphilosoph Michel Albert verteidigte sie 1991 gänzlich unironisch: "Ist eigentlich bekannt , dass Deutschland von dieser Vorstellung gar nicht so weit entfernt ist ? " Ob eine Wirtschaftsordnung nach deutschem Muster den Umbruch in Polen, Tschechien oder Russland besser gestaltet hätte , lässt sich nun nicht mehr herausfinden. In Deutschland allerdings hat **<die soziale Marktwirtschaft>** schon mehr als einen Umbruch bewältigt und dem Land die langjährige "Exportweltmeisterschaft" eingetragen . **<Sie>** hat wirtschaftlichen Wohlstand und soziale Gerechtigkeit miteinander verbunden, in einer Intensität , die weltweit ihresgleichen sucht. Auch wenn der deutsche Konjunkturmotor zur Zeit eher stottert als brummt: Die ökonomische Erfolgsgeschichte Deutschlands rechtfertigt die Suche nach ihren Erfolgsgeheimnissen . Dass es sich um reichlich deutsche Geheimnisse handelt , zeigt sich schon daran , dass die

Frequency of Coreference Chains of Various Length

more than 2-element chains:

Two-element coreference chains:

- German – 69% (156 cases)
- Czech – 62% (343 cases)

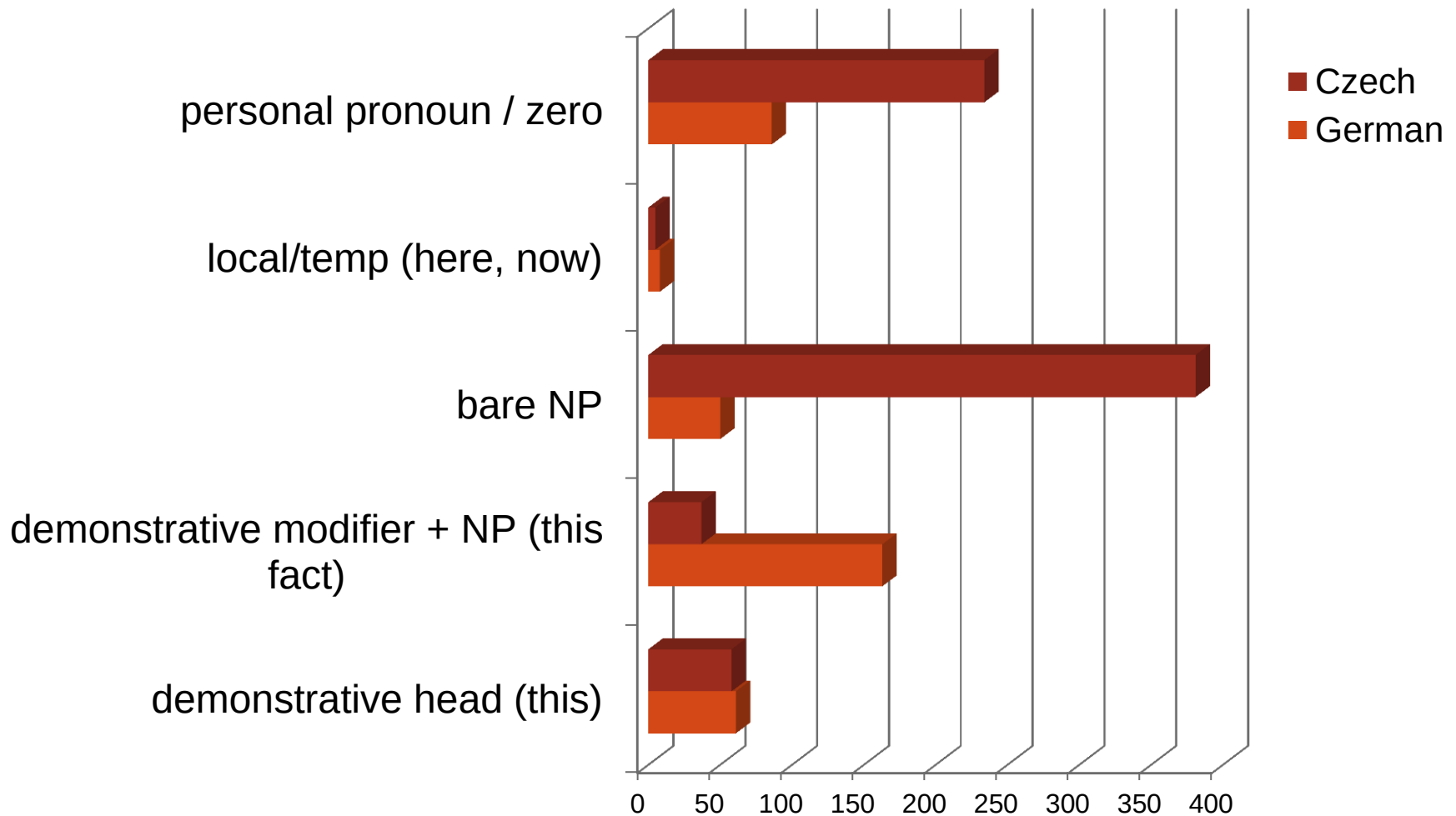


adding lexical cohesion changes the situation not very substantially

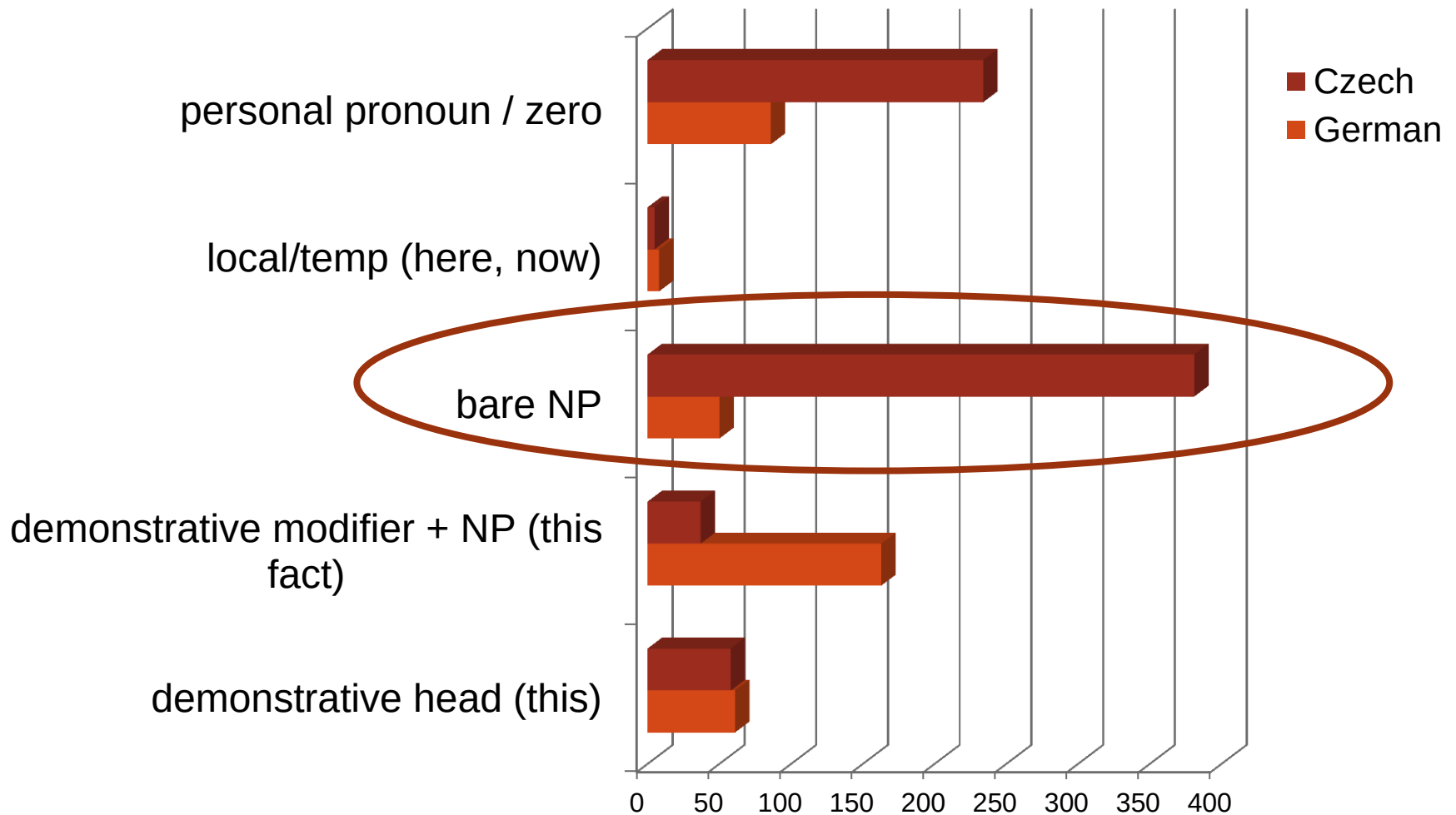
Results so far...

- comparable results as for the properties of coreference chains
- properties of chains - genre- or register-dependent???
- thematic similarities or differences of the selected texts

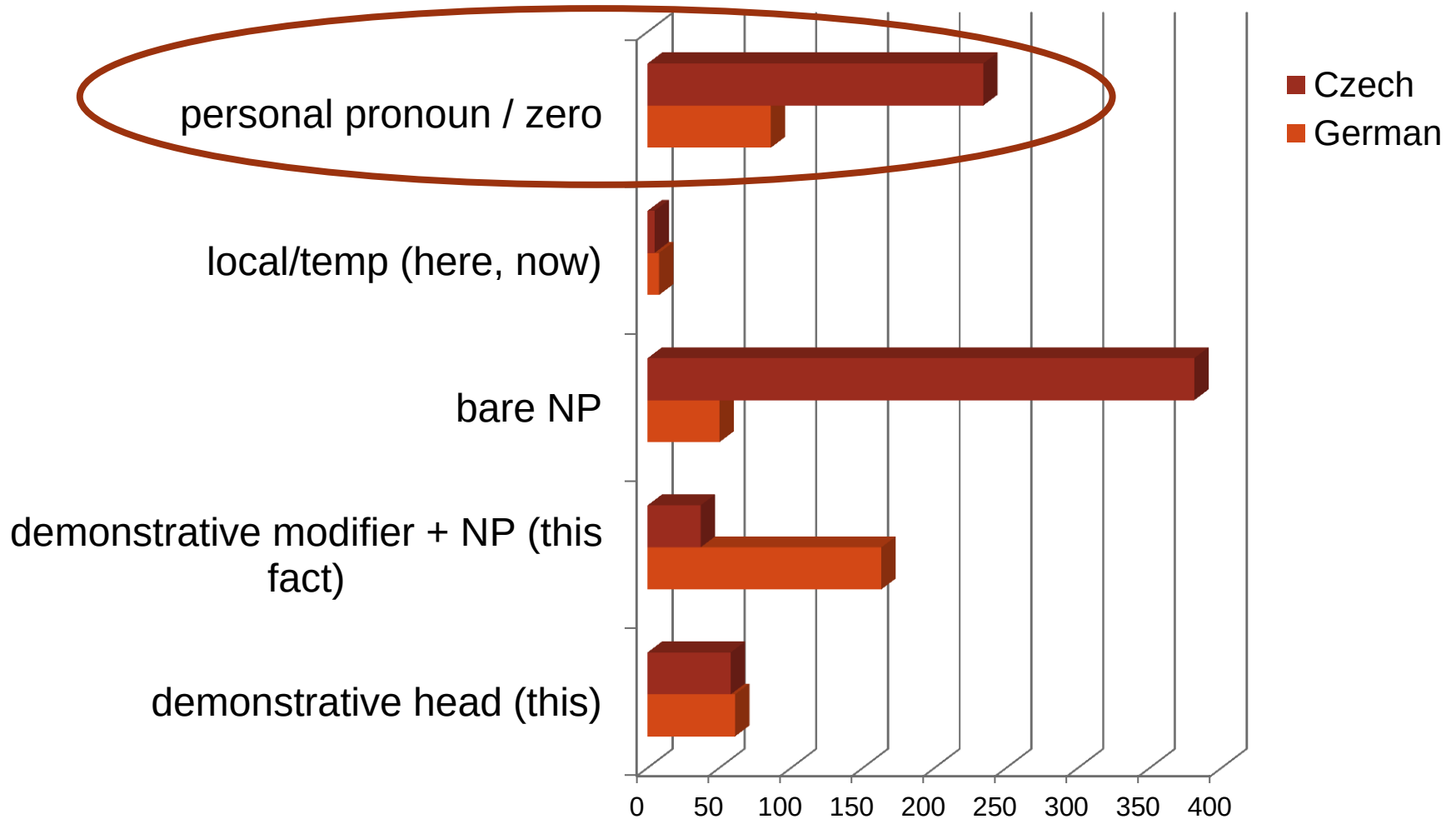
Expressive means of coreference



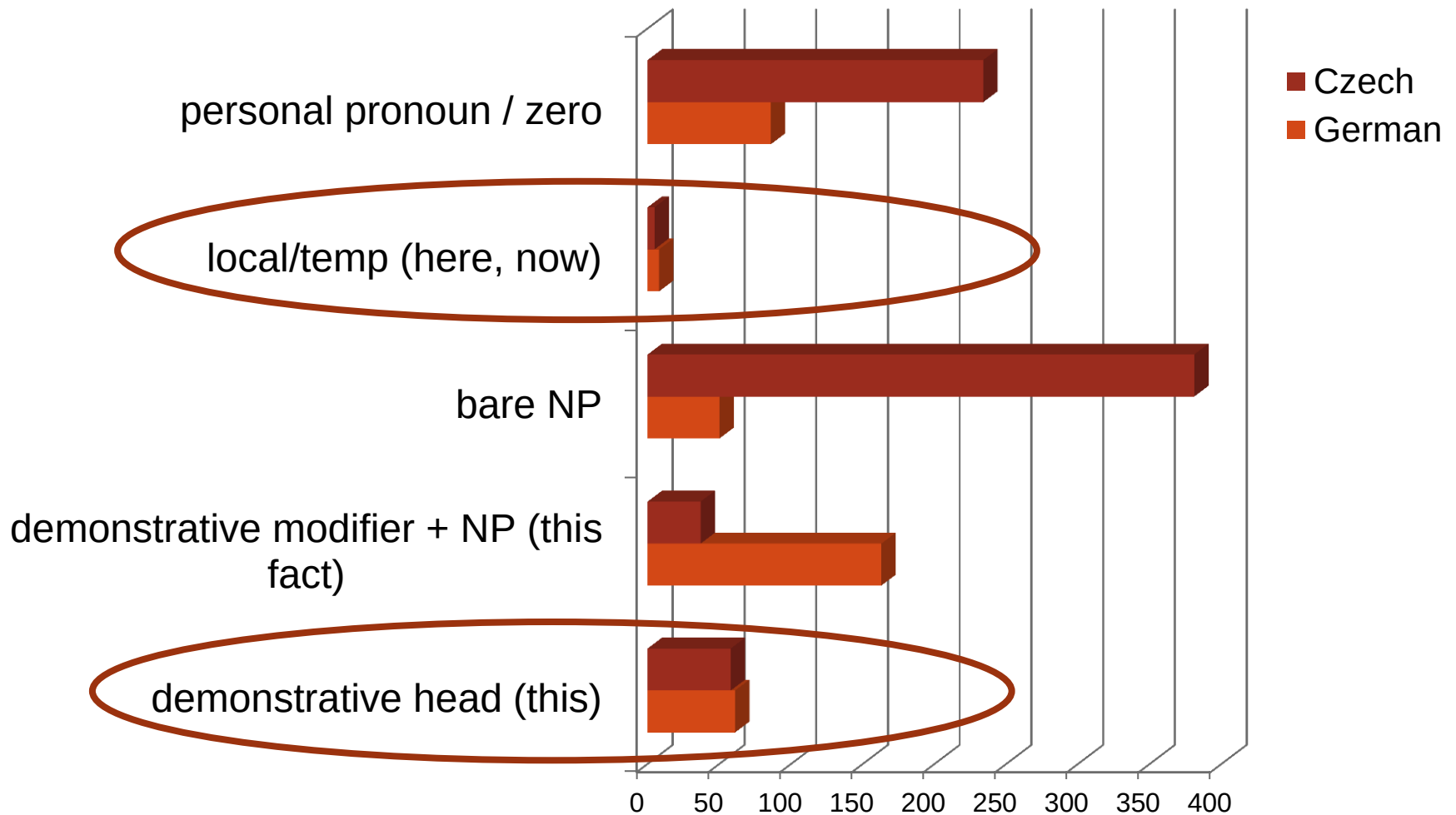
Expressive means of coreference



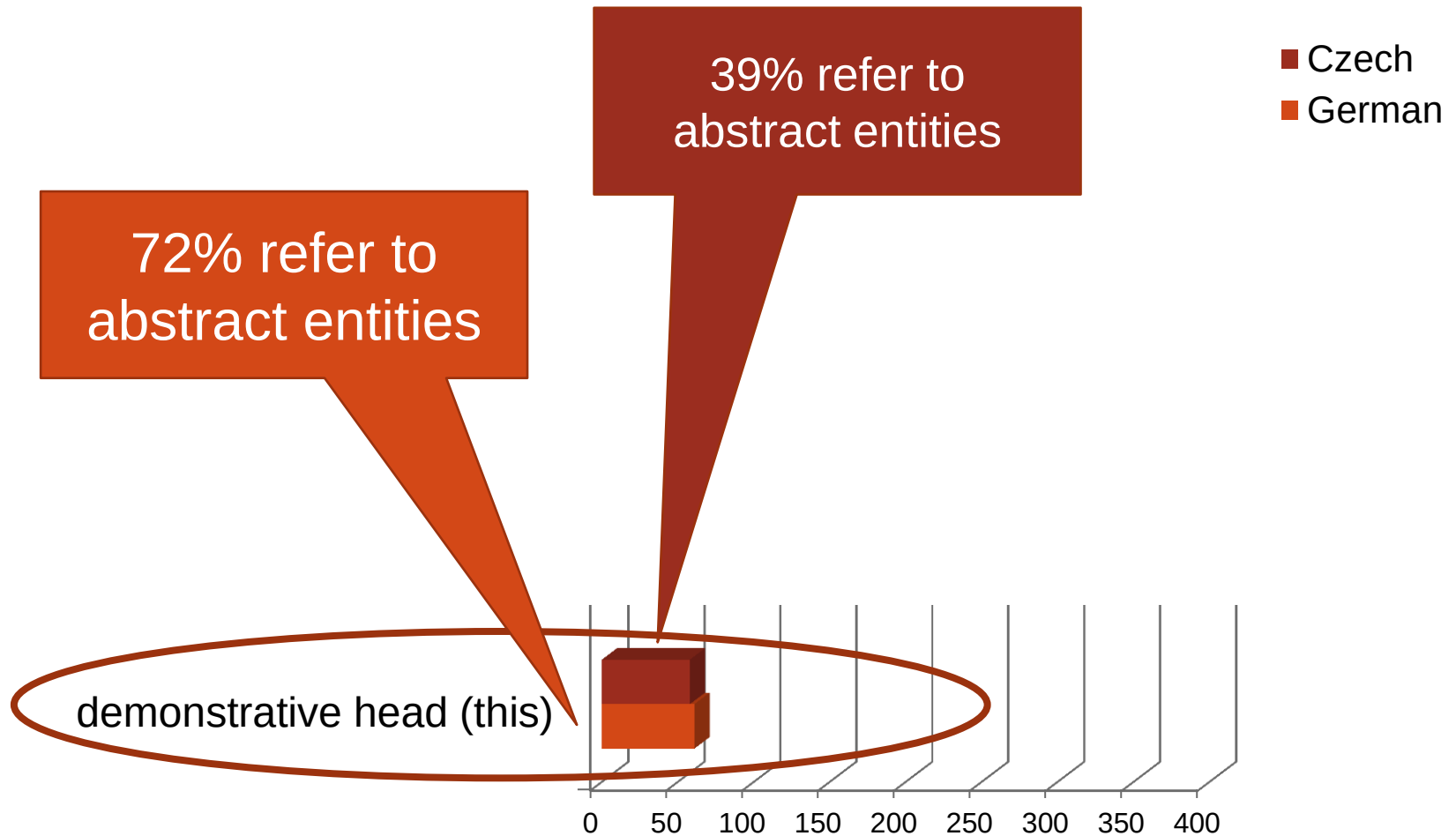
Expressive means of coreference



Expressive means of coreference



Antecedents of Demonstrative Heads



Abstract Anaphors

	GERMAN		CZECH	
	absolute	%	absolute	%
to clauses	5	7.35	13	20.63
to sentences	45	66.18	20	31.75
to bigger segments	18	26.47	30	47.62
TOTAL	68	100	63	100

Logically...

- demonstrative heads (this) – often to clauses

Cizinci podstatně přispěli k německému hospodářskému a kulturnímu vývoji, proč jejich počet naopak ve statistikách nezdůrazňovat a tím veřejně uznat jejich zásluhy o německou hospodářskou a politickou demokracii? [Foreigners have contributed significantly to the German economic and cultural development, so why not to emphasize their number in statistics, and to acknowledge their merit of the German economic and political democracy by this?]

- (dem.mod.+)NP (this/the decline) – often to sentences

Dnes se tento počet snížil na asi půl milionu, jenže důvodem <poklesu> je především skutečnost, že ten, kdo není zaměstnán déle než rok, již podporu nedostane. [Today, that number dropped to about half a million, but the reason for the decline is the fact that anyone who is not employed for more than a year, gets no support anymore.]

- dem.mod.+general noun – often to larger segments

Conclusion

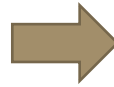


- technical issues
- differences of typological character (absence of definiteness or pro-drops) *also* have influence on the properties of coreference chains
- the main differences in the length and number of chains (reflecting the topic structuring) are rather genre- or even domain-dependent

Future work

- analysis of further genres and domains
- include more texts to have a better data representation

if the analysis of a greater number of texts (on different topics and from different genres) has a



the observed differences are not due to the properties of specific texts, but rather language- or

References

- BBN Technologies. Coreference Guidelines for English OntoNotes – Version 6.0. Linguistic Data Consortium. BBN Pronoun Coreference and Entity Type Corpus. 2006.
- Baker, M. (1995), Corpora in translation studies: An overview and some suggestions for future research. *Target* 7(2). 223–243.
- Botley, S. (2006), Indirect anaphora: Testing the limits of corpus-based linguistics, *International Journal of Corpus Linguistics*, 11(1), pp. 73–112.
- Caselli, T., Prodanof I. (2010), Annotating Event Anaphora: A Case Study, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.
- Dipper S., Zinsmeister H. (2009), Annotating discourse anaphora. *Proceedings of LAW III*, pp. 166–169.
- Francis G. (1994), Labelling discourse: an aspect of nominal group lexical cohesion. Coulthard M. (ed.), *Advances in Written Text Analysis*, London: Routledge, pp. 83–101.
- Fraurud K. (1992), Situation reference: What does 'it' refer to? GAP Working Paper No 24, Fachbereich Informatik, Universität Hamburg.
- Halliday M., Hasan R. (1976), *Cohesion in English*. Longman, London, New York.
- Hansen-Schirra S., Neumann, S., and Steiner, E. (2012), *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.
- Hedberg N., Jeanette K. G., and Zacharski, R. (2007), Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles, *Proceedings of DAARC-2007*, pp. 31–36.
- Kučová L., Hajičová, E. (2004), Coreferential Relations in the Prague Dependency Treebank. *Proceedings of DAARC-2004*, pp. 97–102.
- Lapshinova-Koltunski E., Nedoluzhko A., Kunz K. A. (2015), Across Languages and Genres: Creating a Universal Annotation Scheme for Textual Relations. *Proceedings of the The 9th Linguistic Annotation Workshop (LAW IX 2015) Association for Computational Linguistics*, Stroudsburg, PA, USA, pp. 168-177.
- Lapshinova-Koltunski, E., Kunz, K. (2014), Annotating cohesion for multilingual analysis. *Proceedings of the 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, Reykjavik, Iceland, May. LREC.
- Lapshinova-Koltunski, E., Kunz, K., and Amoia, M. (2012). Compiling a multilingual spoken corpus. In Mello H., Pettorino M., T. R. (ed.), *Proceedings of the VIIth GSCP International Conference: Speech and corpora*, Firenze, Firenze University Press, pp. 79–84.
- Müller Ch. (2008), Fully Automatic Resolution of 'it', 'this', and 'that' in Unrestricted Multi-Party Dialog. Ph.D. thesis, University of Tübingen.
- Navarretta C., Olsen S. (2009), The annotation of pronominal abstract anaphora in Danish texts and dialogues. DAD report 1. Centre for Language Technology, University of Copenhagen.
- Nedoluzhko A. (2011), Extended textual coreference and bridging anaphora [Rozšířená textová koreference a asociční anafora], Prague, Czech Republic, 268 pp., Dec 2011
- Nedoluzhko A., Toldova S., Novák M. (2015), Coreference chains in Czech, English and Russian: Preliminary findings. *Computational Linguistics and Intellectual Technologies*, Vol. 14, No. 21, RGGU, Moscow, Russia, ISSN 2221-7932, pp. 474-486.
- Nedoluzhko, A., Lapshinova-Koltunski, E. (2016), Abstract Coreference in a Multilingual Perspective: a View on Czech and German. In *Proceedings of the CORBON Workshop at NAACL2016*, San Diego, California.
- Pradhan S., Ramshaw L., Weischedel R., MacBride J., Micciulla L. (2007), Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the IEEE-ICSC*.
- Recasens M., Martí, A. (2010), AnCoraCO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources & Evaluation*.
- Recasens, M., Martí, M.-A., Taulé, A.M. (2007), Text as Scene: Discourse Deixis and Bridging Relations. *Procesamiento del Lenguaje Natural*, 39.
- Zinsmeister, H., Dipper, S., 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).

Thanks!