FROM MONOLINGUAL ANNOTATIONS TOWARDS CROSS-LINGUAL RESOURCES: AN INTEROPERABLE APPROACH TO THE ANALYSIS OF DISCOURSE

Ekaterina Lapshinova-Koltunski*, Anna Nedoluzhko, Kerstin Kunz*****

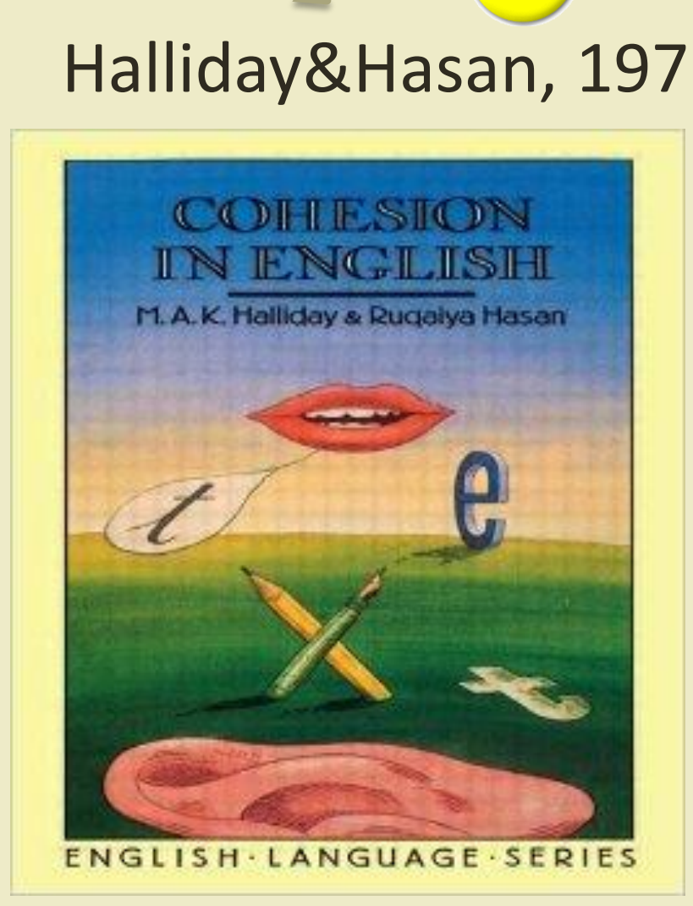
Saarland University*, Charles University in Prague**, University of Heidelberg***

e.lapshinova@mx.uni-saarland.de, nedoluzko@ufal.mff.cuni.cz, kerstin.kunz@iued.uni-heidelberg.de

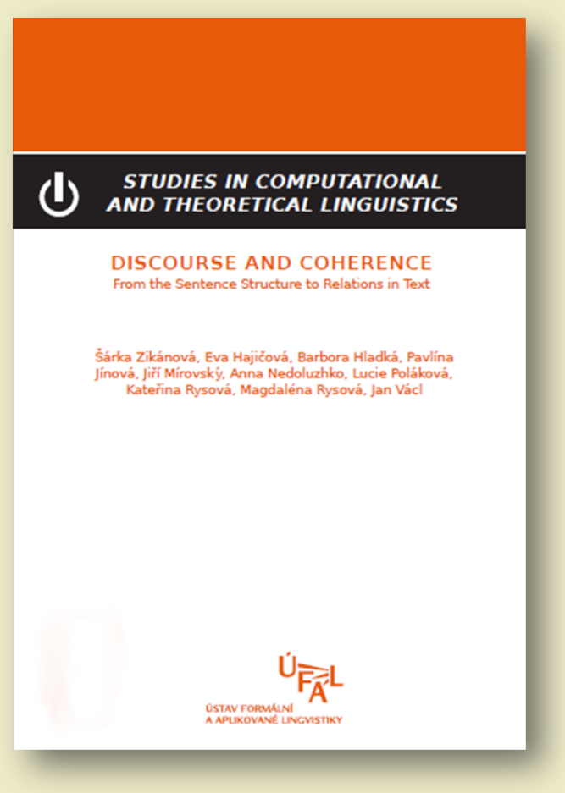
AIMS AND MOTIVATION

theories

Halliday&Hasan, 1976



Zikánová et al. 2015, PDTB-style discourse



resources

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

- ❖ 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 framework

GERMANIC AND SLAVIC LANGUAGES, METHODS AND DATA

German	Czech
❑ morphologically rich ❑ less isolating ❑ flexible word order ❑ no definite article	❑ reduced morphology ❑ more isolating ❑ less flexible word order ❑ definite article

RESULTING SCHEMA APPLIED

	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 *same*	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

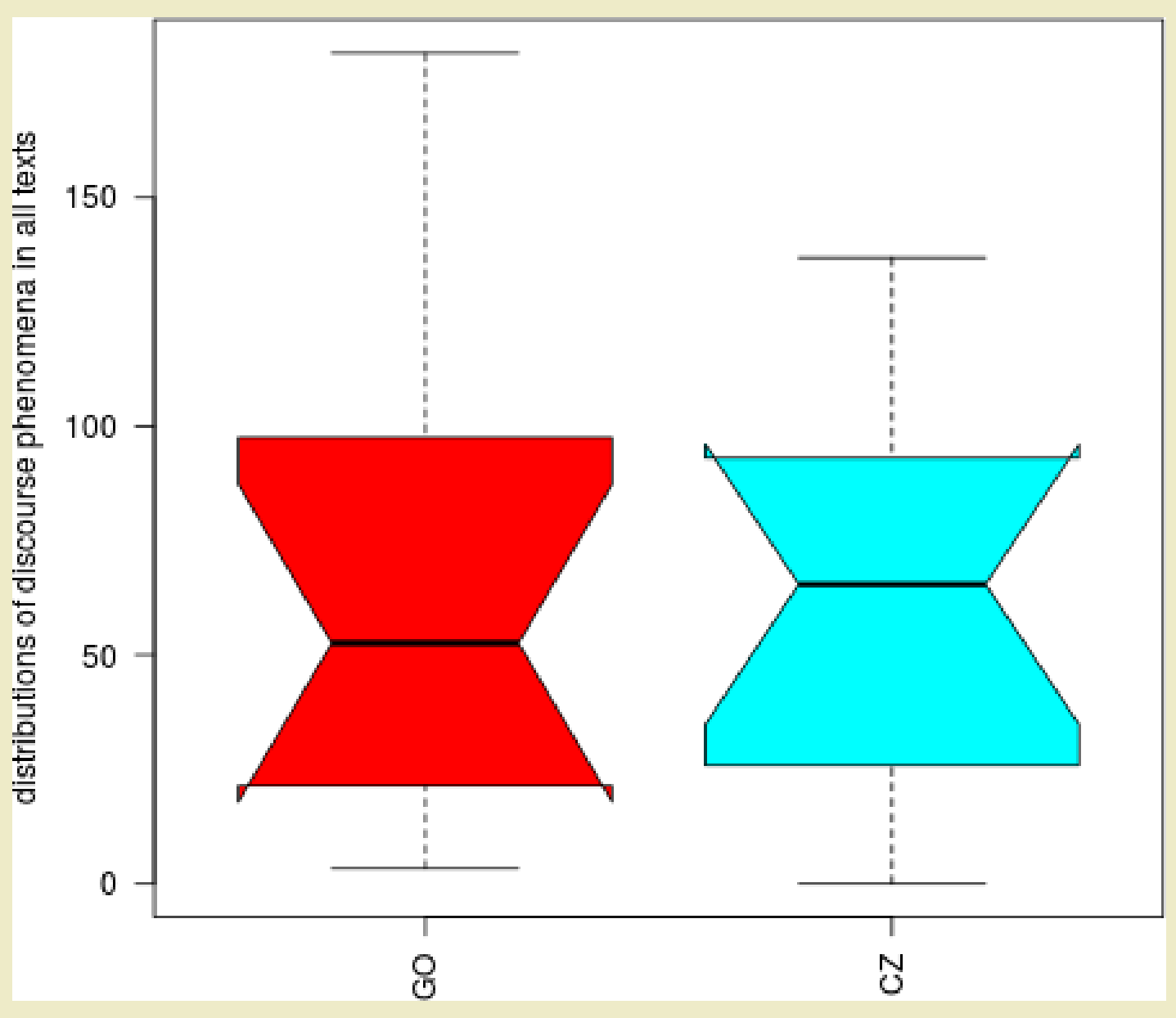
German data excerption

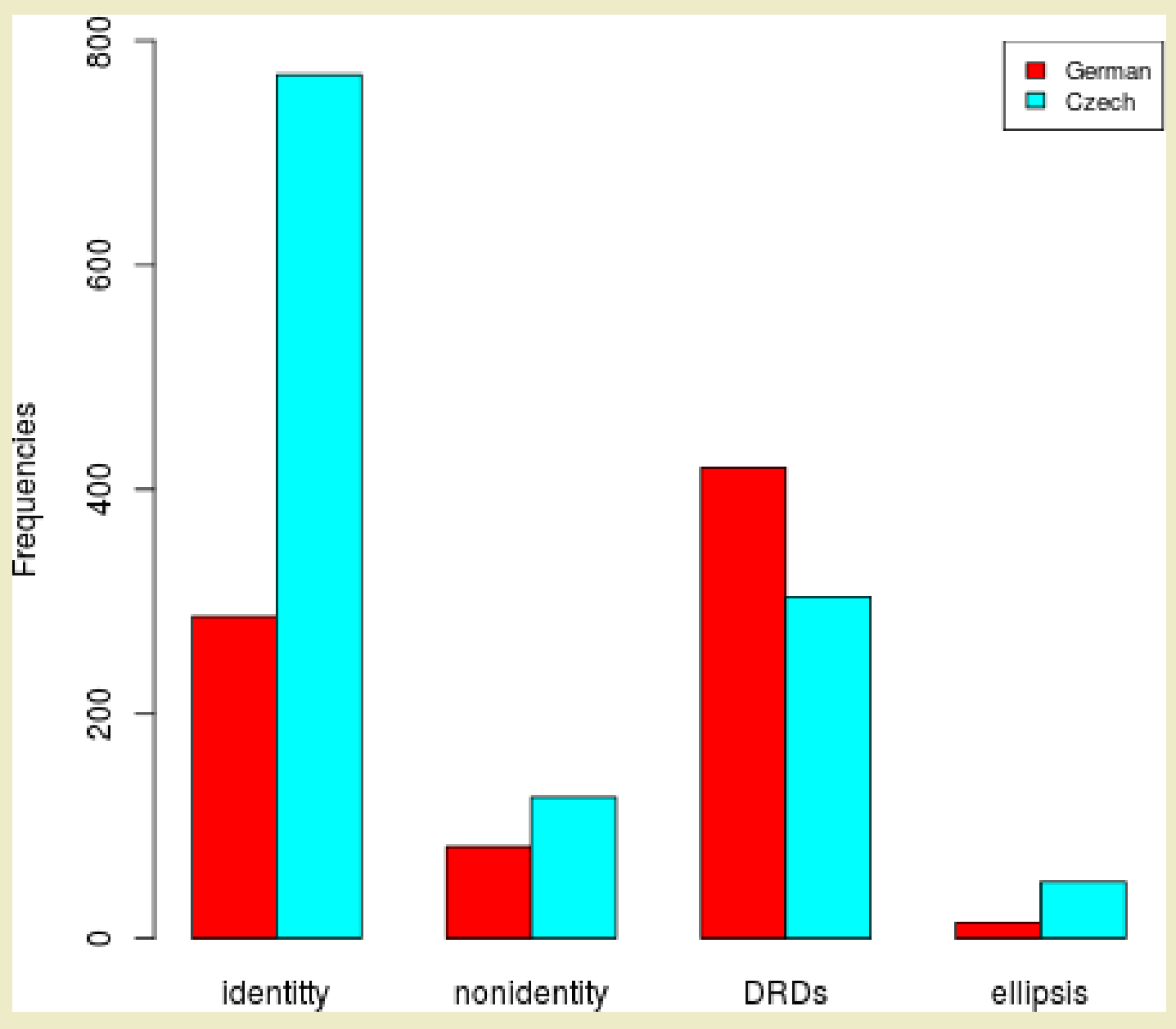
id	topics	sent	tokens
1	Germany and social market economy	121	2035
2	Optimistic remarks on globalisation	47	971
3	Politics and globalisation	103	1871
4	Globalisation and new challenges	27	478
5	The biggest currency changeover	85	1460
6	Globalisation and market economy	80	1782
7	Global market and technical progress	108	1851
8	Economic and technological changes	73	1795
9	Doctors and medical system	92	2687
	Total: **ALL GERMAN**	**736**	**14930**

Czech data excerption

id	topics	sent	tokens
1-5	Germany, politics and history	170	687
6	Housing	83	1644
7-8	Technological changes	73	1795
9-12	Politics	121	1854
13-14	Economics	149	2568
15-16	Unemployment	112	2252
17	Television	55	969
	Total: **ALL CZECH**	**763**	**11769**

ANALYSES AND RESULTS





SUBCATEGORIES *(normalised per 1000)*

featID	id1	id2	id3	id4	id5	nonid1	nonid2	temp	cont	comp	expan	ellipsis
German	88.4	38.2	144.7	3.3	12.0	52.9	28.8	106.5	52.2	79.0	181.5	14.1
Czech	97.7	64.6	597.3	0.0	10.2	88.4	37.4	14.4	66.3	86.7	136.8	50.1

RESULTS

- interoperable schemes permit multilingual analyses of resources from different approaches, saves time and effort as no additional annotation is required
- insights into differences between German and Czech
- limitations of the dataset: although comparable, still different sources, authors, size
- numbers show more differences in conceptualisations and annotation details, not in languages so far

EXAMPLES

id1 with pronouns referring to nominal antecedent

German: Als Superstar der sozialen Marktwirtschaft gilt aus gutem Grund Ludwig Erhard. Er hatte.. in den 50er Jahren... die produktiven Kräfte der Unternehmen entfesselt und daraus ein Wirtschaftswunder gezaubert... [Ludwig Erhard is regarded as the superstar of the social market economy, and for good reasons. ...in the nineteen-fifties..., he had unleashed the productive forces of business and in this way conjured up an economic miracle...]
Czech: Ta přijala strategii Bílého domu v domnění, že je to nejjistější cesta k vítězství. [She endorsed the White House strategy, believing it to be the surest way to victory.

nonid1 with meronymy relations

German: ...praktisch wird es dazu nicht kommen – dafür ist in Deutschland die Bereitschaft zur Solidarität, der Glaube an das “für alle” zu groß. Eine andere Gefahr ist da weit realer: daß die Deutschen... [In practice... it will not come to that– the readiness to practice solidarity, and people’s belief in the “for all” is too pronounced in Germany. Another dangers much more real, however: that the Germans....]
Czech: Jednotlivá studia v apartmánech jsou vybavena kuchyní, takže je možná individuální příprava stravy.[Studio apartments are equipped with kitchens, so everyone can prepare their own food.]

id2 with abstract anaphora

German: Gleichzeitig brauchen wir mindestens eine Verdoppelungdes Wohlstands. Wenn wir die Armutsgegendender Erde anschauen, weiß jeder sofort, dass dies das Mindeste an moralischer Herausforderung ist. [At the same time, we need to double the current level of prosperity. One look at the poor regions throughout the world is enough to make anyone realize that this is the most urgent moral challenge we face.]
Czech: 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?]

nominal ellipsis

German: All das ist eine kleine Revolution. Die grössere [] ist diese: … [But there is also a bigger [revolution], and it is this]
Czech: Klienti pojištoven, které ukončí svou činnost, se automaticky vrátí k Všeobecné []. [Clients of insurance companies which shut down will automatically return to the General [one].

nonid2 with contrast relations

German: Dazu gehören zum Beispiel die Halbierung der Energie-und Rohstoffintensität bis 2020 gegenüber 1990 und die Verdoppelung des Anteils erneuerbarer Energien am Energieverbrauch bis 2010. [For example, halving the amount of power and raw material consumption by 2020 compared to 1990 levels and doubling the percentage of renewable energy used as part of total energy consumption by 2010.]
Czech: Saldo běžného účtu platební bilance podle odhadu dosáhlo vloni cca 600 USD... I když letos a příští rok je nutné počítat se zpomalením růstu vývozu, prognózujeme, že saldo přesto zůstane kladné. [The balance of the current account deficit is estimated to reach $600 last year ... Although this and the next yeas we expect the slowdown in export growth, we forecast that the deficit will still remain positive.]

comp with discourse relations of comparison/contrast

German: Arbeiten wie die Polen, aber leben wie die Japaner...[Work like the Poles, but live like the Japanese...]
Czech: Poslední statistické sčítání dopravy proběhlo v roce 1990. Za poslední tři roky se však na českých silnicích zvýšil provoz. [The latest statistical traffic census took place in 1990. Over the past three years, however, traffic on Czech roads has increased.]

ÚFAL
UNIVERSITÄT DES SAARLANDES
GAČR GRANTOVÁ AGENTURA ČESKÉ REPUBLIKY
16-05394S