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

Coreference in Czech and cross-lingually – ideas and perspectives

Anna Nedoluzhko,

Warszawa, 30.4.2015

HUMAN CAPITAL
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

*Plan of the talk

* Corpora and coreference annotation
* Principles of coreference and bridging annotation
* Benefits and problems of annotation on t-trees
* Resolvers for coreference in Czech
* Inter-annotator agreement
* Analysis of agreement and disagreement

*PDT 3.0

* The Prague Dependency Treebank [Bejček et al. 2013]
* Czech newspaper texts,
* ca. 50000 sentences (3165 documents, 833195 tokens)
* The three PDT layers – capture grammatical information: morphological, surface shape (analytical) and underlying syntactic (tectogrammatical)
* t-layer includes
  * semantic labeling of content words and coordinating conjunctions
  * argument structure description based on a valency lexicon
  * ellipsis reconstruction
  * coreference annotation (pronominal, zero, NPs incl. differentiation to specific/generic)
  * bridging relations annotation
  * discourse structure annotation

*PCEDT 2.0+

* 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 (pronominal, zero, NPs – only specific)
  * ellipsis reconstruction

Corefence annotations in Prague corpora

	PDT 3.0	PCEDT	
		English	Czech
grammatical coreference	YES	YES	YES
pronominal textual coreference	YES	YES	YES
anaphoric zeros	YES	YES	YES
textual full-NP coreference - specific	YES	YES (PCEDT2.0+)	YES (PCEDT2.0+)
textual full-NP coreference - generic	YES	no	no
bridging relations	YES	no	no

*Coreference

- **grammatical coreference**
  - mostly possible to identify the antecedent on the basis of grammatical rules of the given languages
  - within one sentence
- **textual coreference**
  - not restricted to grammatical means alone, context
  - different means (pronominalisation, grammatical agreement, repetitions, synonyms, paraphrasing, hyponyms/hyperonyms, etc.)
  - often occurs between entities in different sentences

Grammatical coreference

- arguments in constructions with verbs of control
  - *John wants to [#Cor.ACT] kiss Mary.*
- reflexive pronouns
  - *John shaved himself.*
- relative pronouns
  - *John, who came late, apologized.*
- coreference with verbal modifications that have dual dependency
  - *John saw Mary [#Cor.ACT] stand on the windowsill and cry.*
- reciprocity
  - *John and Mary kissed [#Rcp.PAT].*

*Textual coreference

* personal and possessive pronouns (*Jonh left Mary. He wanted to see his mother*),
* demonstrative pronouns *ten, ta, to* (*It means that he doesn't really love Mary.*)
* with textual ellipsis (zeros) (*Více si Ø vážil své matky.*)
* nouns (*John asked his mother to advise him how he should behave with Mary, but mother ignored her son's wish.*)
* local adverbs (*John asked mother to come to Mary's place with him but she decided not to go there.*)
* some adjectives (*At last, Mary came to Prague herself and found the Prague atmosphere quite casual.*)
* reference to events (*Mary suggested Jonh to go to the theater, but Jonh ignored her wish*).
* If antecedent is a whole segment of (previous) text larger than one sentence (phrase) — special type of textual coreference segm(ent) without explicitly marked antecedent: (*The next day Mary suggested to visit his mother. Then she proposed to go swimming. Her last wish was just to look at the city center. Jonh denied all of it.*)

Textual coreference - types

* **SPEC(ific)** for coreference of NPs with specific reference (*Jonh$_a$ asked his$_a$ mother$_b$ to advise$_c$ him$_a$, but mother$_b$ ignored her$_b$ son$_a$'s wish$_c$*)
* **GEN(eric)** for coreference of NPs with generic reference, e.g.

  (Mary proposed Jonh to go to the Zoo to see animals. She believed that having looked at animals, Jonh will understand how wild he behaved to her.)

END OF STORY

*Bridging Relations

- **part – whole** (PART_WHOLE and WHOLE_PART)

  Germany – Bavaria - Munich

- **set — subset/element of set** (SET_SUB and SUB_SET)

  students – some students – a student

- **function - object** (P_FUNCT and FUNCT_P)

  prime-minister – government, trainer – football team

- **semantic contrast** (CONTRAST)

  *A přesvědčen jsem ještě o jednom - je třeba mít **vysoké cíle** a s **malými [cíli]** se nespokojit.* (= *And I am sure about one thing: it is necessary to have **lofty aims** and not to be satisfied with **small (ones)**.*)

- explicit **anaphora without coreference** (ANAF)

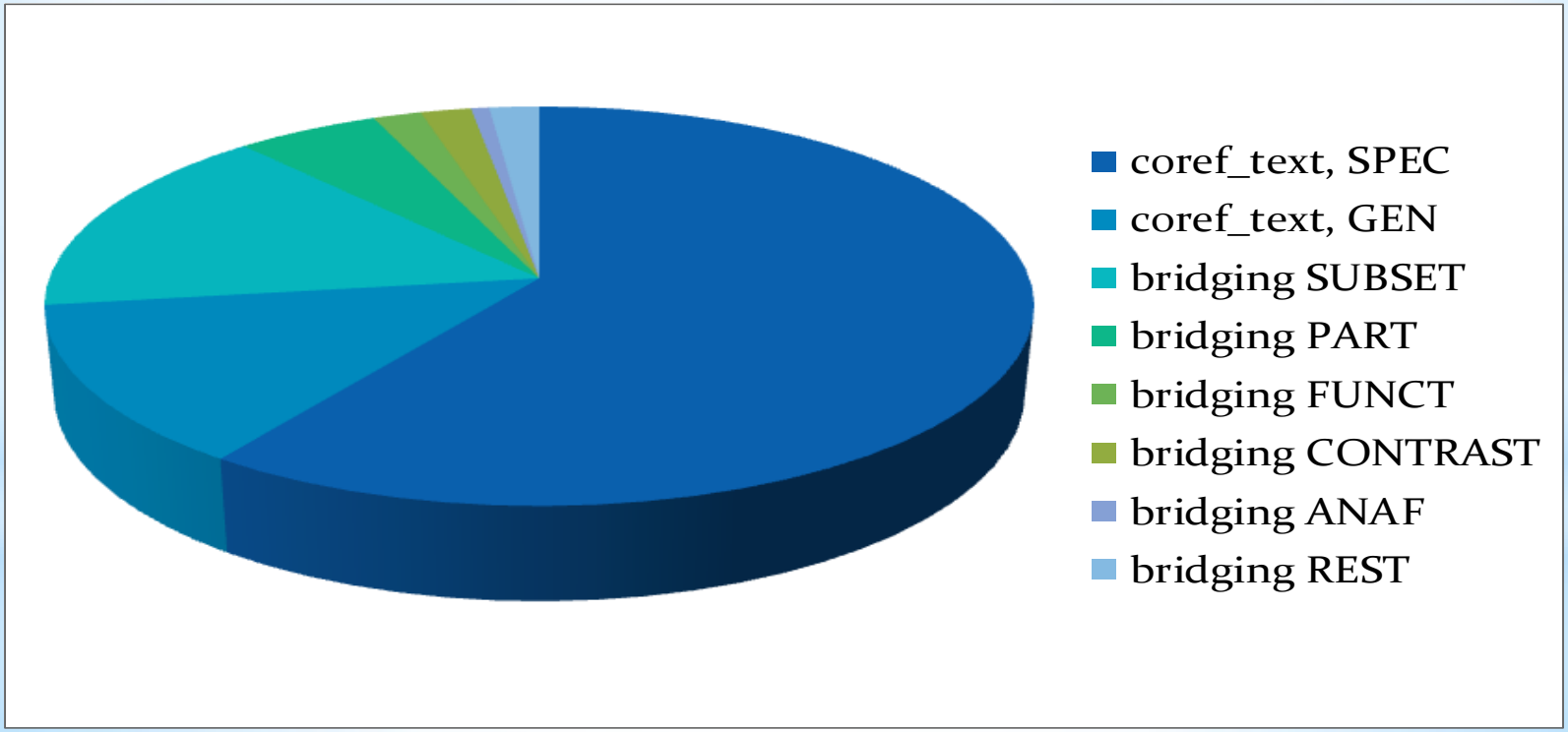
  *"**Duha**?" Kněz přiloží prst **k tomu slovu**, aby nezapomněl, kde skončil. - "A **rainbow**?" The priest pointed to **the word** ...*

  *Jak se Vám zamlouvala **Pragobanka Cup**? - **Takovou/podobnou/stejnou akci** bychom také uvítali - How did you like **the Pragobanka Cup**? We would welcome **a similar event** ...*

- **other** (REST)

  family *(grandfather - grandson)*, place – inhabitant, author – work, same denomination to support cohesion of the text ***(a chance** helped – **another chance** entered the game ...)* and event – participant of the event (***enterprise - entrepreneur)***

*Statistics – Types of Relations in PDT

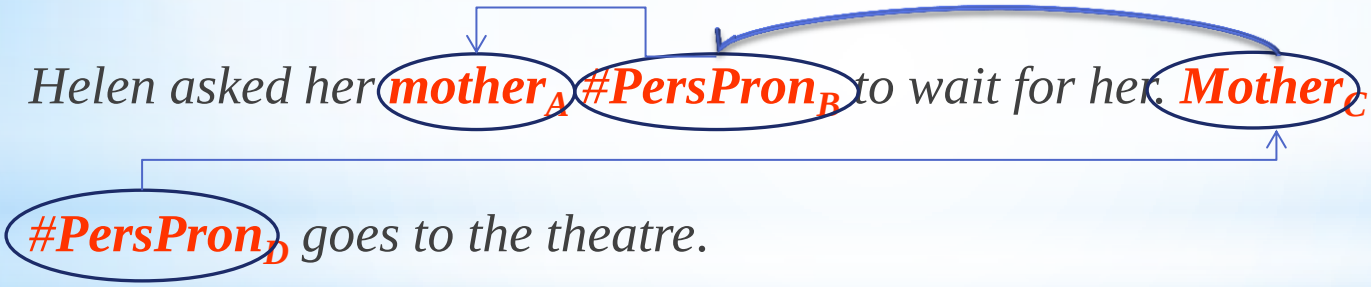


*Principles of Coreference Annotation-1

- chains – reference to the nearest antecedent
- maximal length of chains (incl. grammatical and textual coreference)

Example:

Helena poprosila svou ***maminku***$_A$*, aby* ***#PersPron***$_B$ *na ni počkala.*
Matka$_C$ *řekla, že* ***#PersPron***$_D$ *jde do divadla.*



the chain is established: A <= B <= C <= D

*Principles of Coreference Annotation-2

- maximal “scope” of the units: whole subtree
- “cooperation” with the TGTS’s: no special annotation of apposition, predicates etc.
- contribution to the coherence of the text
- preference of coreference over bridging anaphora: in case of multiple choice, we prefer textual coreference to bridging anaphora

  Mary – John – children in the class – Mary and John
- principle of preferring coreference to anaphora: coreference, not anaphora, is subject to annotation

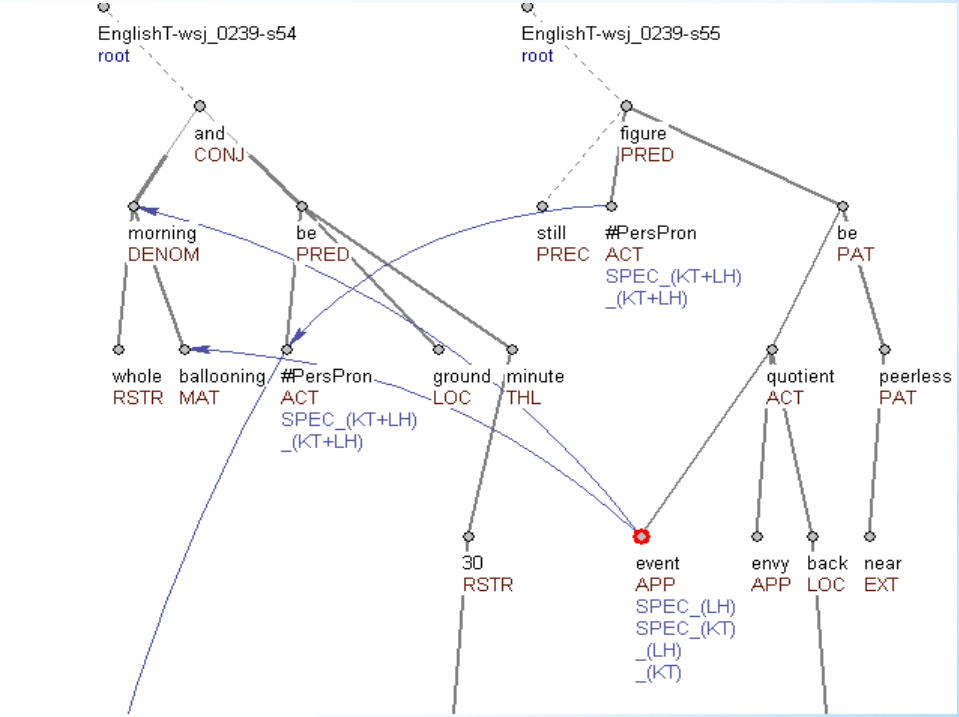
Benefits of dependency trees

* extraction of markables (MIN-IDs and maximal scope)
* reconstruction of syntactic zeros
  * #Perspron: personal or possessive pronouns
  * #Cor: controllee in control constructions
  * #Qcor: valency modification in constructions with quasi-control, e.g. *He offered Jan {#QCor} protection*.
  * #Rcp: participants that are left out as a result of reciprocation, e.g. *The lovers kissed {#Rcp.PAT}*.
  * a copy of the node representing the same lexical unit as the omitted element
* non-referring expressions:
  * appositions
  * coordinative constructions
  * verbal complements

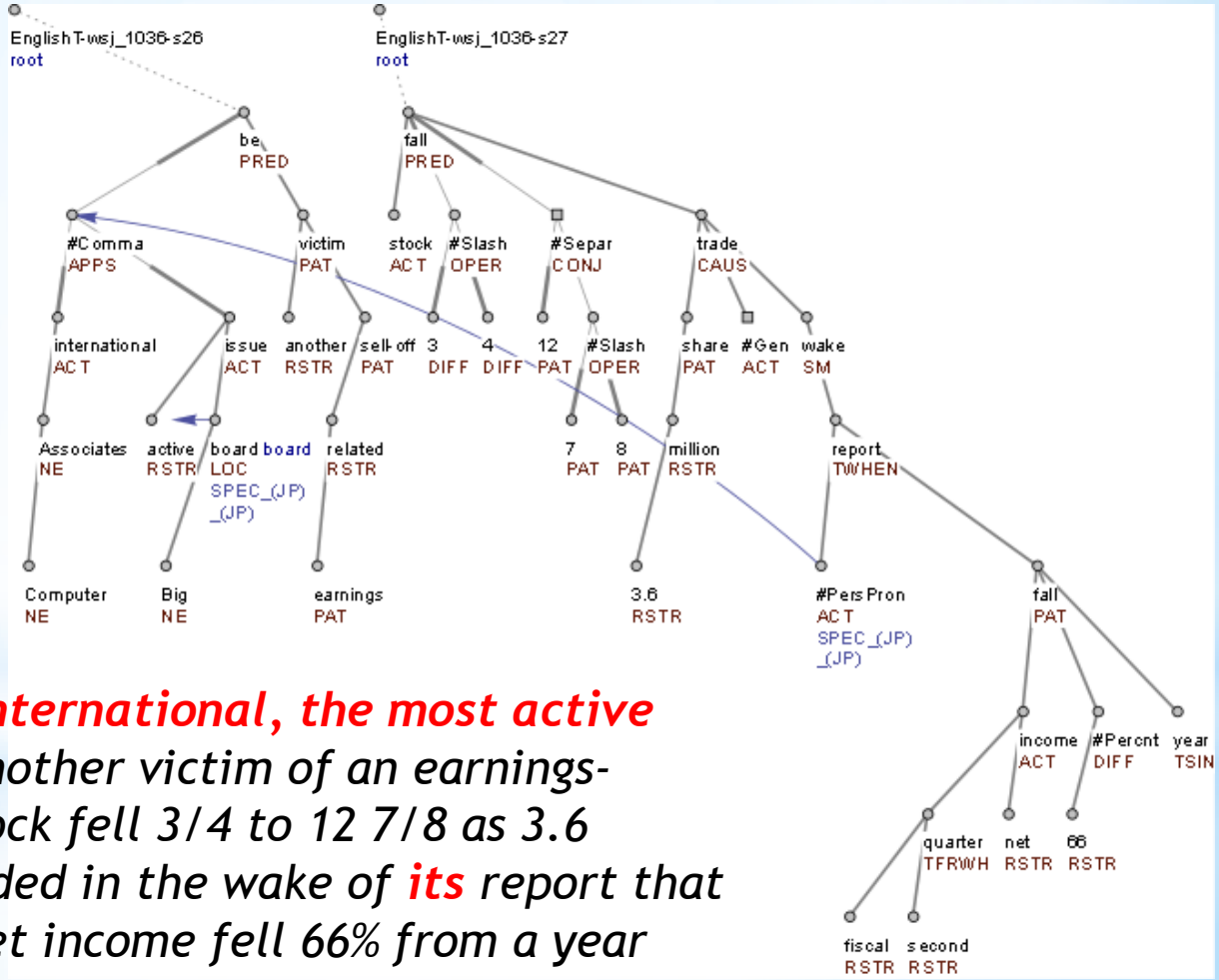
*Benefits of t-trees - markable identification

* convention: annotate larger antecedent positioned "higher" in TGS.

*{_A **A whole morning of** {_B **ballooning**}} and I had been off the ground barely 30 minutes. Still, I figured **the event's** envy-quotient back in the U.S.A. was near peerless.*



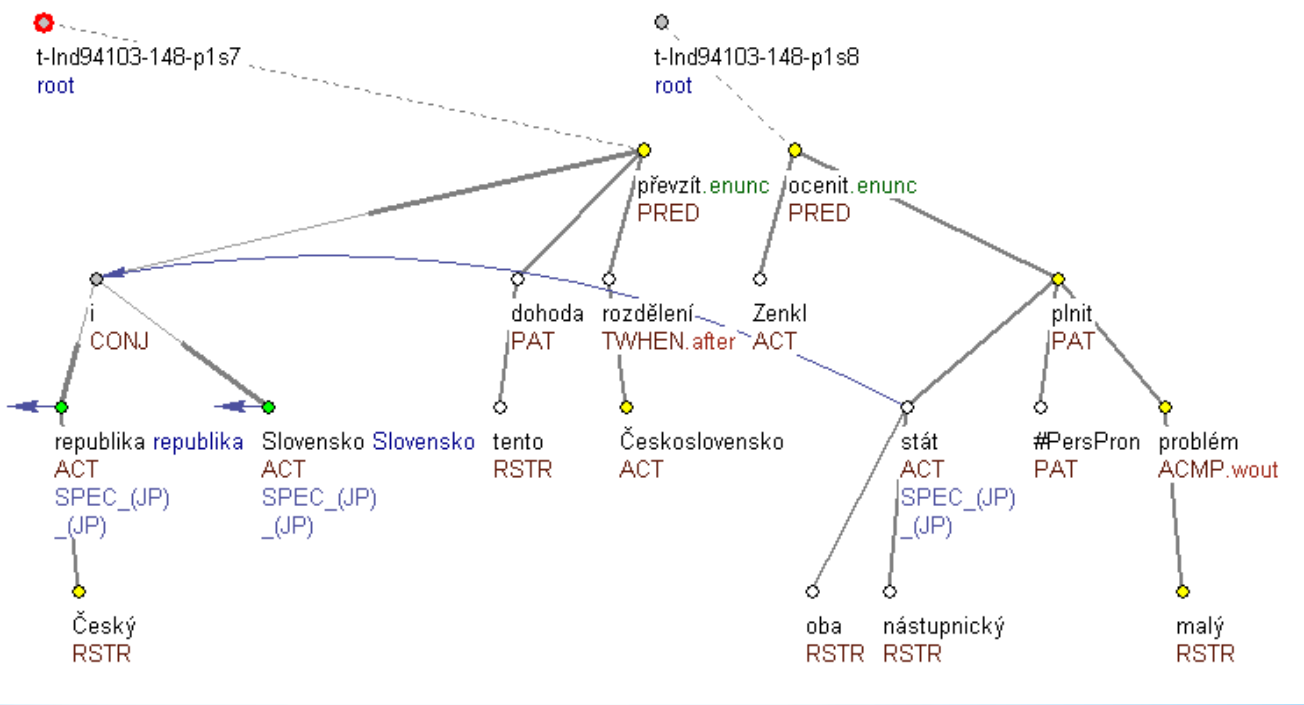
*Benefits of t-trees - Apposition



***Computer Associates International, the most active Big Board issue**, was another victim of an earnings-related sell-off. The stock fell 3/4 to 12 7/8 as 3.6 million shares were traded in the wake of **its** report that fiscal second-quarter net income fell 66% from a year ago.*

*Benefits of t-trees - Coordination

Česká republika i Slovensko tuto dohodu po rozdělení Československa převzaly. Zenkl ocenil, že ji oba nástupnické státy plní bez nejmenších problémů. - *Czech Republic and Slovakia took over this agreement after the split of Czechoslovakia. Zenkl appreciated that both successor states follow it without any problems.*

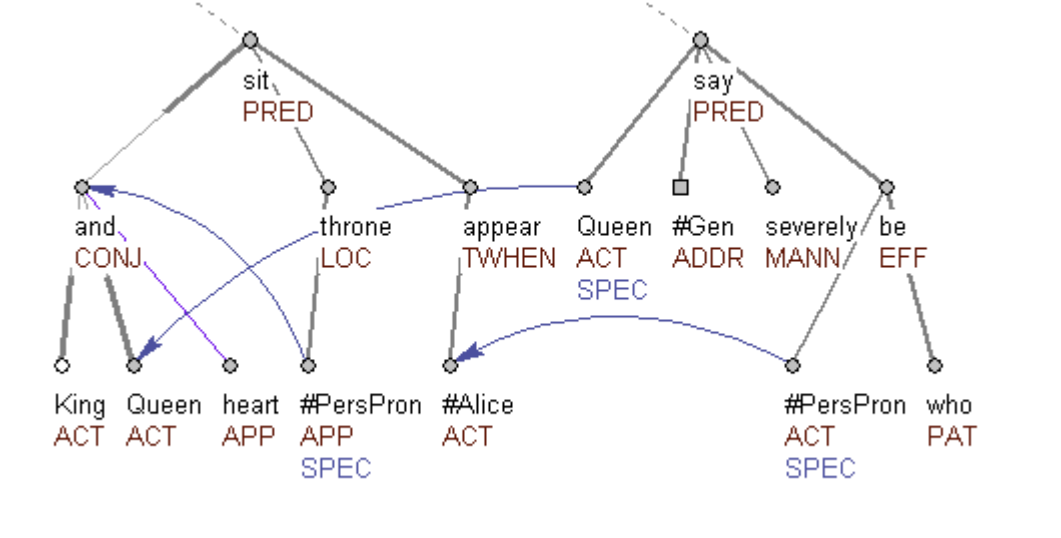


*Benefits of t-trees - Coordination

+ ellipsis of dependent element

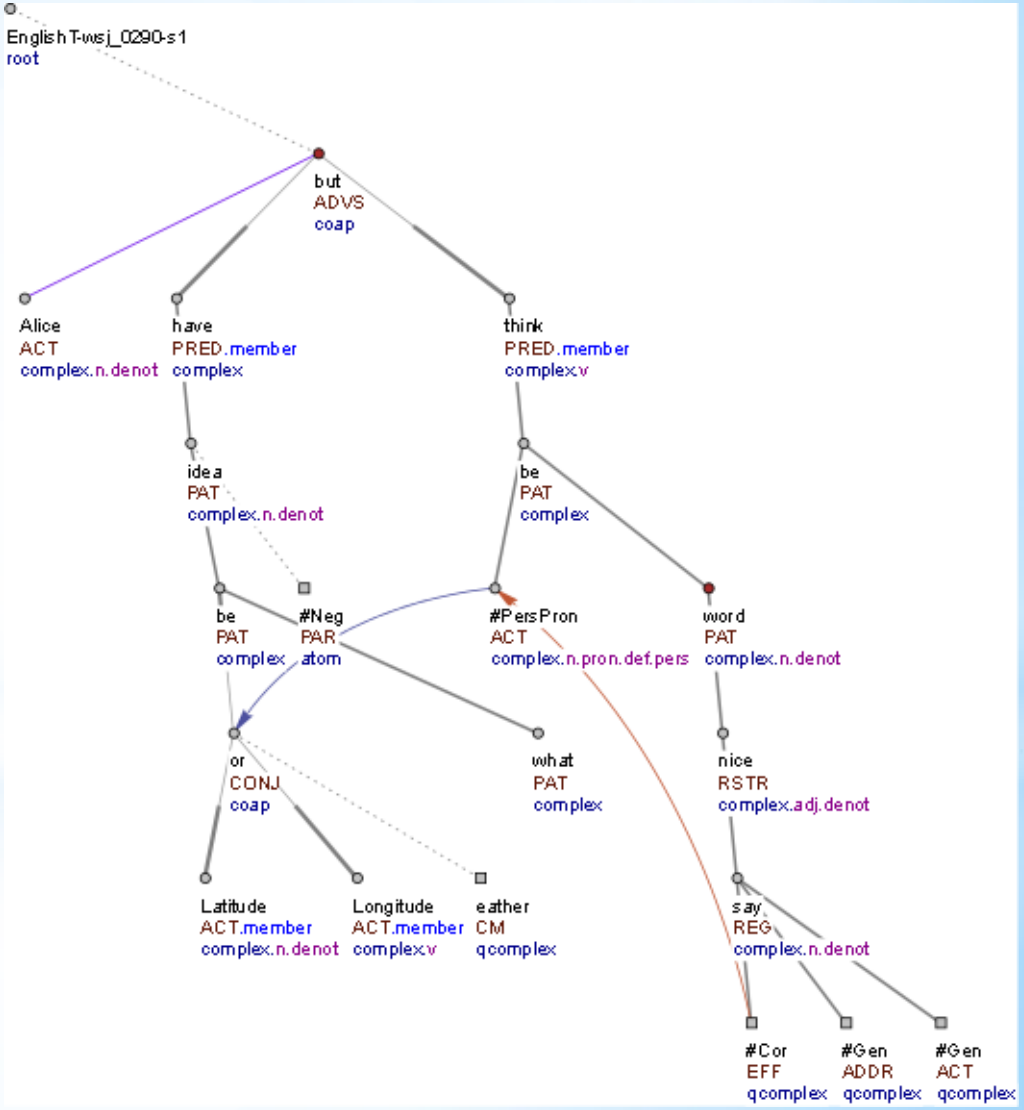
The King and Queen of Hearts *were sitting on* ***their*** *throne when Alice appeared.* ***The Queen*** *said severely "Who is she?"*

+ embedded dependent element



*Benefits of t-trees - Coordination

*Alice had no idea what **[Latitude] was**, or **[Longitude] either**, but thought **they** were nice grand words to say.*



Problematic issues: prepositional phrases

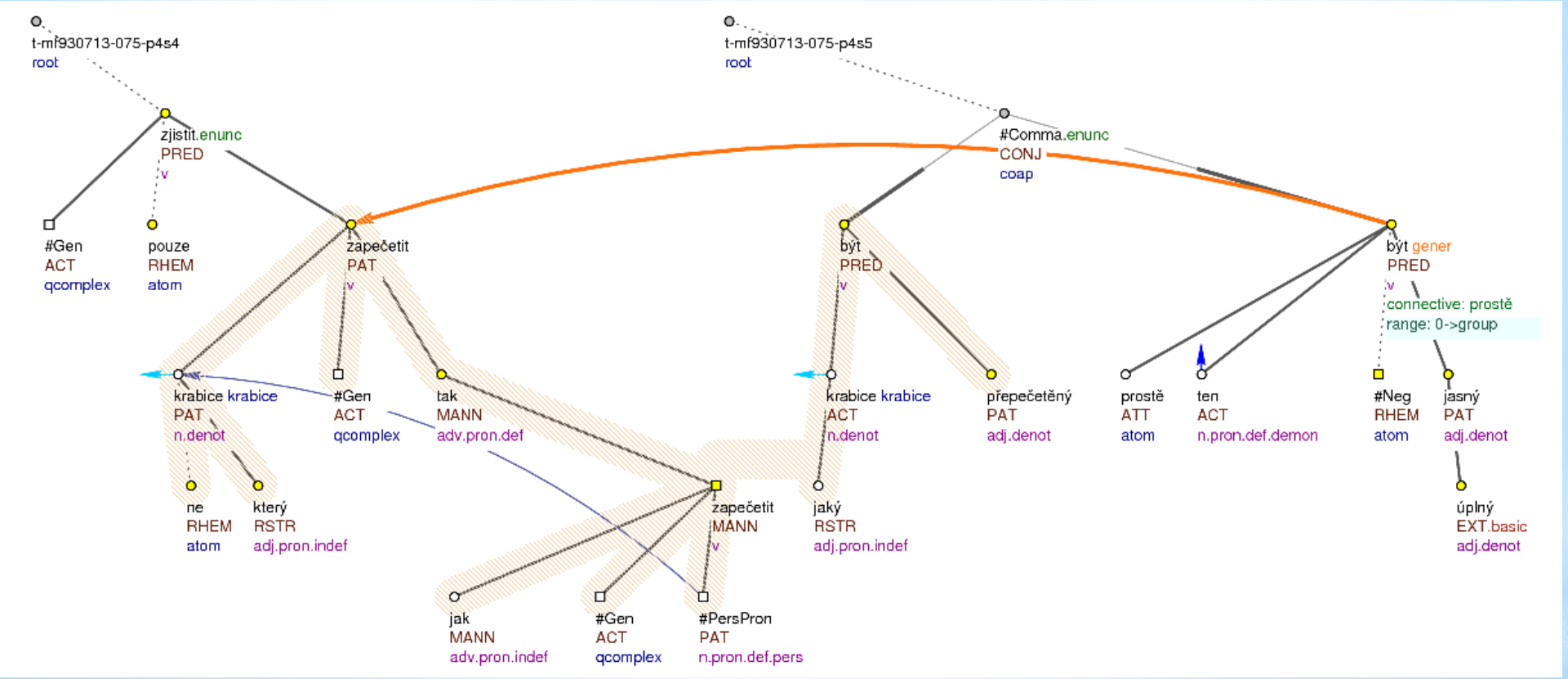
* in tectogrammatical structure, prepositions are embedded in tectogrammatical nodes
* PPs are annotated as NPs (*near Prague* = *Prague*, *before the war - during the war - after the war*)

not good but technically reasonable decision? (otherwise very low agreement for *in Prague - about Prague - for Prague* vs. *in Prague - around Prague - above Prague*) --- not always clear, see example:

Zatím se posunuje stále více za Prahu, čímž ztrácí na své účelnosti z hlediska dopravních spojení do jednotlivých částí města Na druhé straně by tu asi mohlo být víc pozemků vhodných k podnikání. Po dálnici bychom se měli svézt z Prahy až do Českých Budějovic, v roce 1997 pravděpodobně projedou první vozidla po dálnici Praha -Plzeň, dokončena by měla být i dálnice D8 z Prahy do Ústí nad Labem.

(= So far, people begin to move away from Prague, ... various parts of the city. On the other hand, there could be more lands suitable for business there. Highways could take us from Prague up to CeskéBudejovice … .)

Possibility to refer to bigger segments



Pouze se zjistilo, že ne všechny krabice jsou zapečetěny tak, jak by měly být. Nějaké krabice byly přepečetěné, nebylo to prostě úplně jasné. - However, there was found that not all the boxes are sealed as they should be. Some boxes were sealed too much, it wasn't just clear enough.

Performance of tools for coreference and bridging

type of the task	data	F_1
Grammatical coreference, verbs of control	PDT 2.0	91.5
Grammatical coreference, reflexive pronouns	PDT 2.0	97.1
Grammatical coreference, relative pronouns	PDT 2.0	99.6
Grammatical coreference, reciprocity	PDT 2.0	94.7
Pronominal coreference, rule-based	PDT 2.0	74.2
Pronominal coreference, perceptron ranking, gold features	PDT 2.0	79.4
Pronominal coreference, perceptron ranking, system features	PDT 2.0	50.3
NP-coreference, specific NPs	PDT 2.0	48.1 (P:59.7, R:40.3)
NP-coreference, generic NPs	PDT 2.0	1.8 (P:20, R:0.9)
bridging relations	PDT 2.0	0
Identification of an anaphoric unexpressed subject, rule-based	PCEDT 2.0	61.5
Identification of an anaphoric unexpressed subject, rule-based, exploiting English side	PCEDT 2.0	69.5

Performance of tools for coreference and bridging

type of the task	data	F_1
Grammatical coreference, verbs of control	PDT 2.0	91.5
Grammatical coreference, reflexive pronouns	PDT 2.0	97.1
Grammatical coreference, relative pronouns	PDT 2.0	99.6
Grammatical coreference, reciprocity	PDT 2.0	94.7
Pronominal coreference, rule-based	PDT 2.0	74.2
Pronominal coreference, perceptron ranking, gold features	PDT 2.0	79.4
Pronominal coreference, perceptron ranking, system features	PDT 2.0	50.3
NP-coreference, specific NPs	PDT 2.0	48.1 (P:59.7, R:40.3)
(NP-coreference, generic NPs)	PDT 2.0	1.8 (P:20, R:0.9)
(bridging relations) new features!	PDT 2.0	0
Identification of an anaphoric unexpressed subject, rule-based	PCEDT 2.0	61.5
Identification of an anaphoric unexpressed subject, rule-based, exploiting English side	PCEDT 2.0	69.5

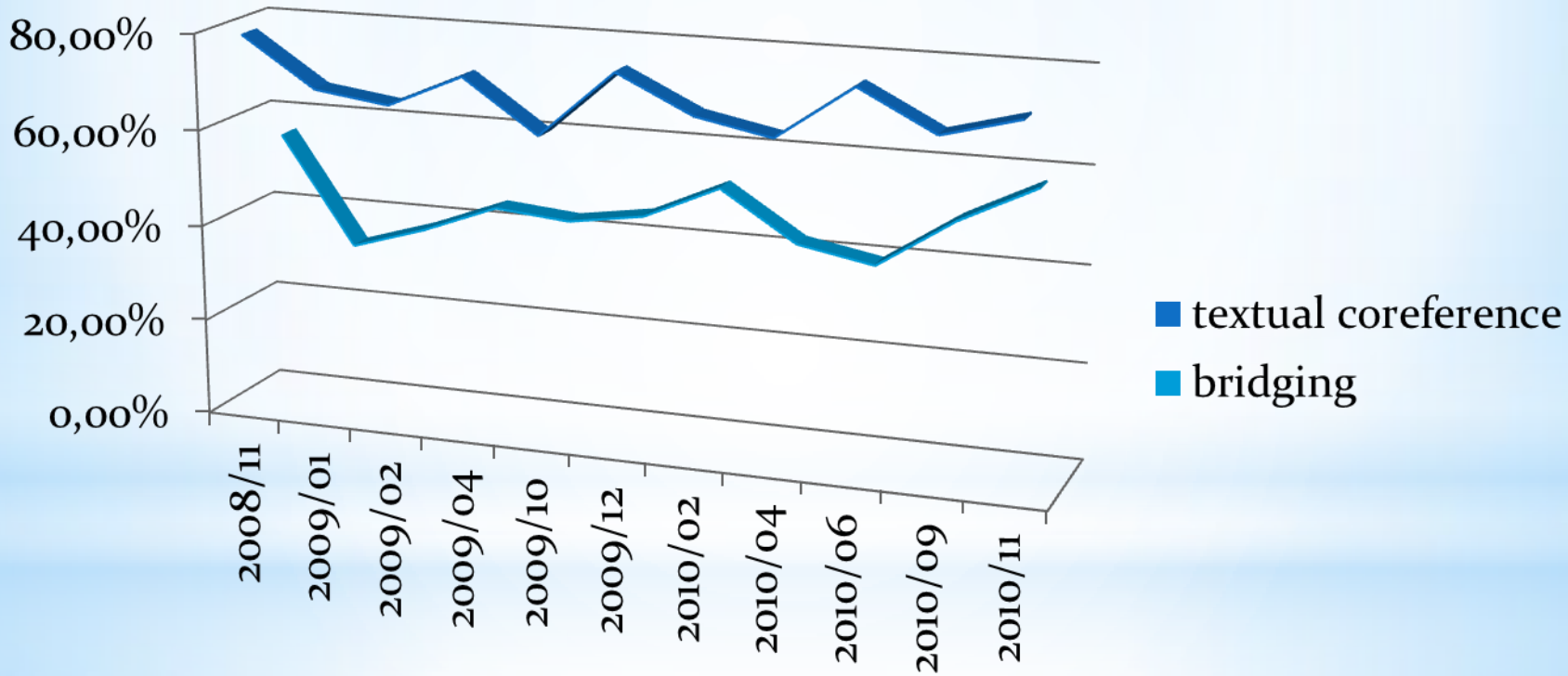
Performance of tools for coreference and bridging

type of the task	data	F_1
Grammatical coreference, verbs of control	PDT 2.0	91.5
Grammatical coreference, reflexive pronouns	PDT 2.0	97.1
Grammatical coreference, relative pronouns	PDT 2.0	99.6
Grammatical coreference, reciprocity	PDT 2.0	94.7
Pronominal coreference, rule-based	PDT 2.0	74.2
Pronominal coreference, perceptron ranking, gold features	PDT 2.0	79.4
Pronominal coreference, perceptron ranking, system features	PDT 2.0	50.3
NP-coreference, specific NPs	PDT 2.0	48.1 (P:59.7, R:40.3)
NP-coreference, generic NPs	PDT 2.0	1.8 (P:20, R:0.9)
bridging relations	PDT 2.0	0
Identification of an anaphoric unexpressed subject, rule-based	PCEDT 2.0	61.5
Identification of an anaphoric unexpressed subject, rule-based, exploiting English side	PCEDT 2.0	69.5

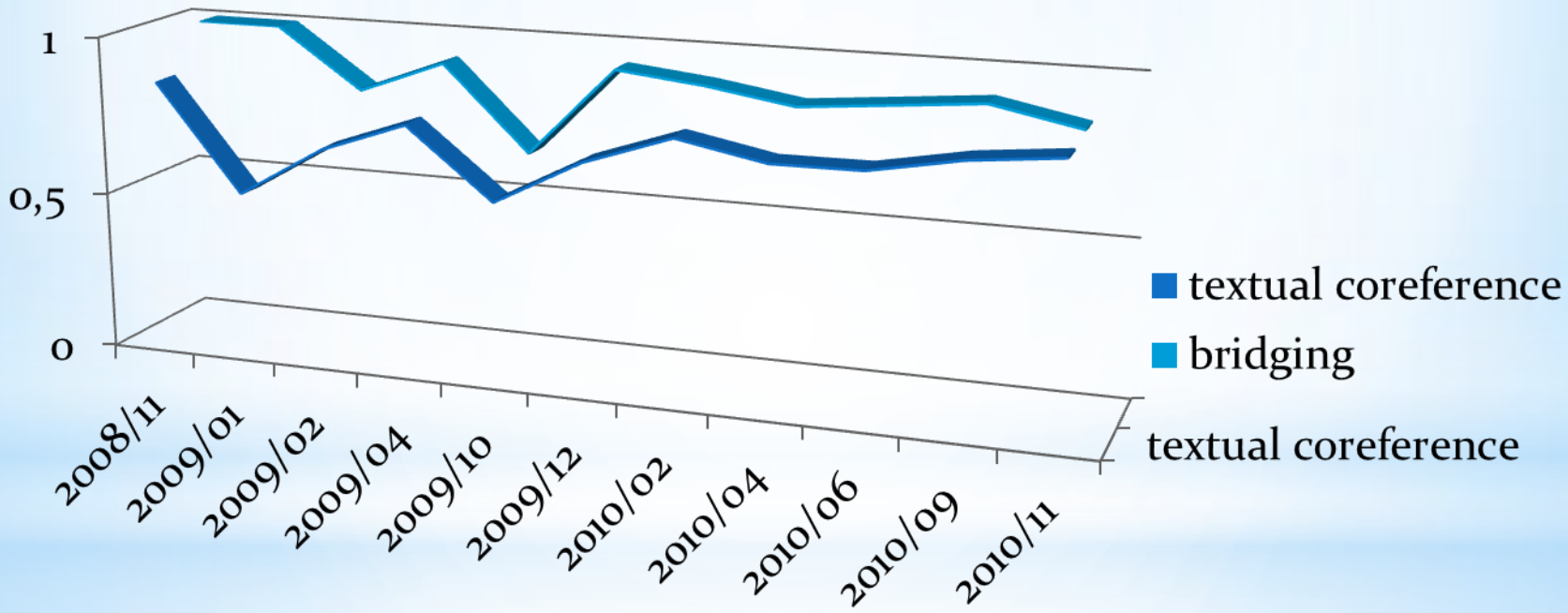
Inter-annotator Agreement for coreference and bridging in PDT

number of controlled documents	39
number of controlled sentences	1606 (3% PDT)
number of controlled tokens	26,520
F-1 on textual pronominal coreference (including zeros)	0,86
F-1 on textual coreference for specific NPs	0,705
F-1 on textual coreference for generic NPs	0,492
F-1 on bridging relations	0,455
textual NP_coref kappa of agreement on type	0,759
bridging kappa of agreement on type	0,889

*Inter-annotator Agreement



*Inter-annotator Agreement: *kappa* for Types



*Types of disagreement

* the relation should or should not be annotated for coreference/bridging
* what is the correct antecedent of a given noun phrase
* distinguishing between the bridging anaphora and the textual coreference
* selecting the type of the bridging anaphora or the textual coreference

Annotating / not annotating a relation

A když už byla knížka hotova, tak se zjistilo, že je praktická i pro rodiče. V této knize je poučení, jak snášejí děti rozvod a jak na něj reagují, a návod, jak se mají rodiče chovat, aby se utrpení dětí snížilo.

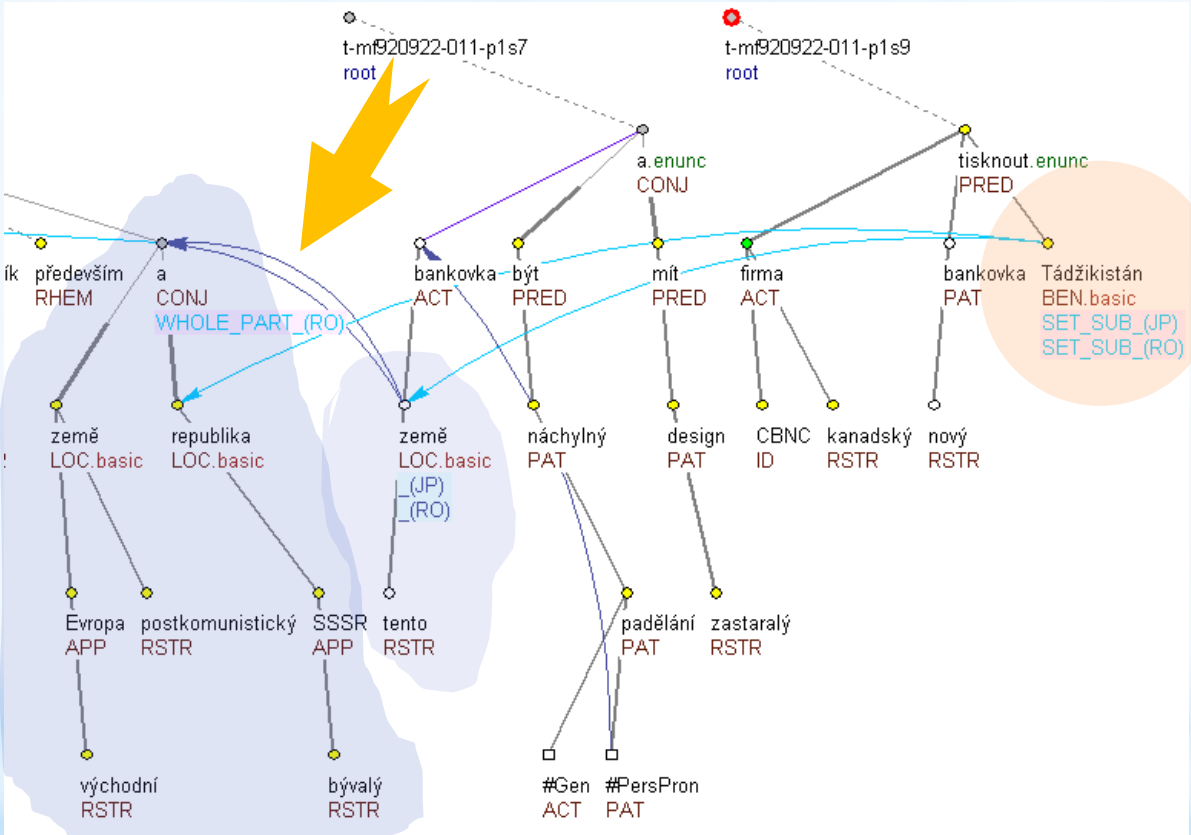
(=After the book had been already written, it was clear, that it is quite useful for parents too. The book contains explanations, how children go through divorce, how they react to it, and the instructions how parents should behave to minimize the suffering of their children..)

* Different selecting the antecedent/anaphoric element

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

(= *They have new clients, first of all in the post-soviet countries of East Europe and in the republics of the former USSR. Banknotes in these countries can be easily falsified. The CBNC Company will print banknotes for Tajikistan.*)

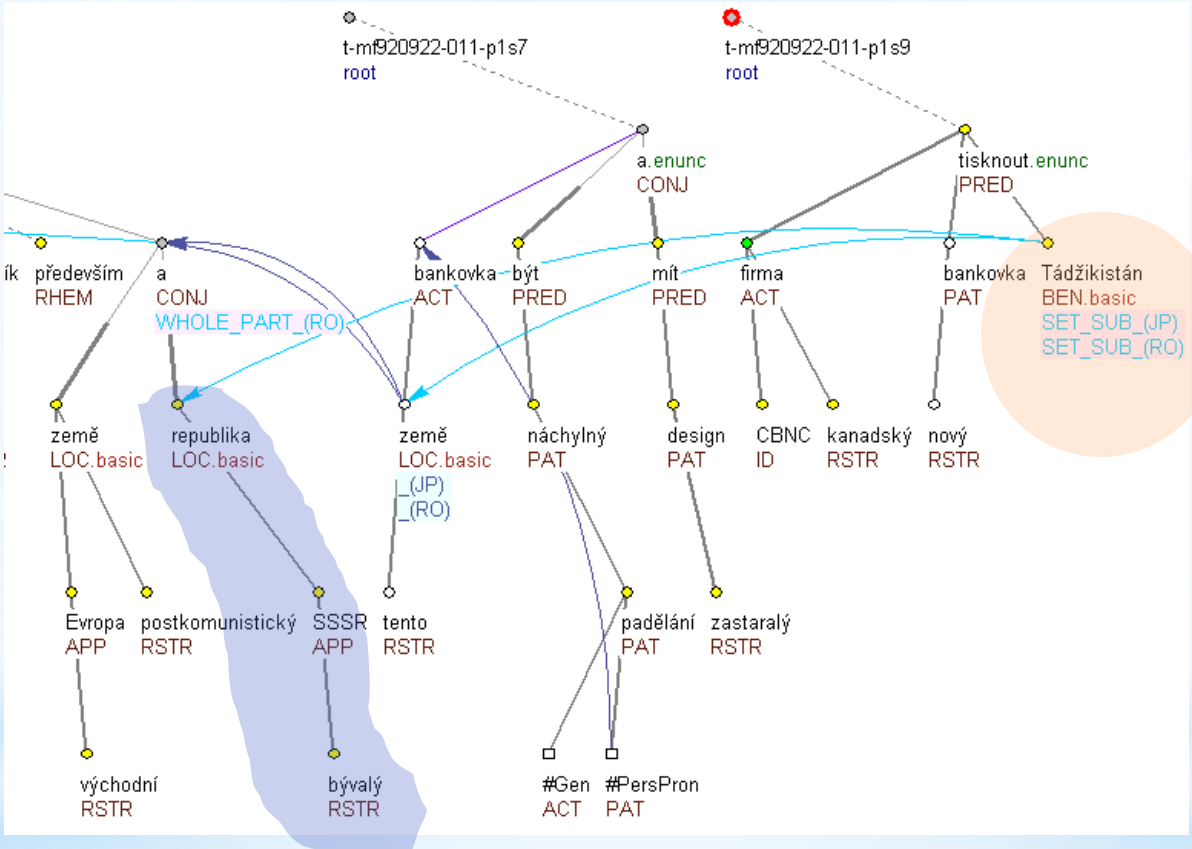
* Different selecting the antecedent/anaphoric element



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 .

(= They have new clients, first of all in the post-soviet countries of East Europe and in the republics of the former USSR. Banknotes in these countries can be easily falsified. The CBNC Company will print banknotes for Tajikistan.)

* Different selecting the antecedent/anaphoric element



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

(= *They have new clients, first of all in the post-soviet countries of East Europe and in the republics of the former USSR. Banknotes in these countries can be easily falsified. The CBNC Company will print banknotes for Tajikistan.*)

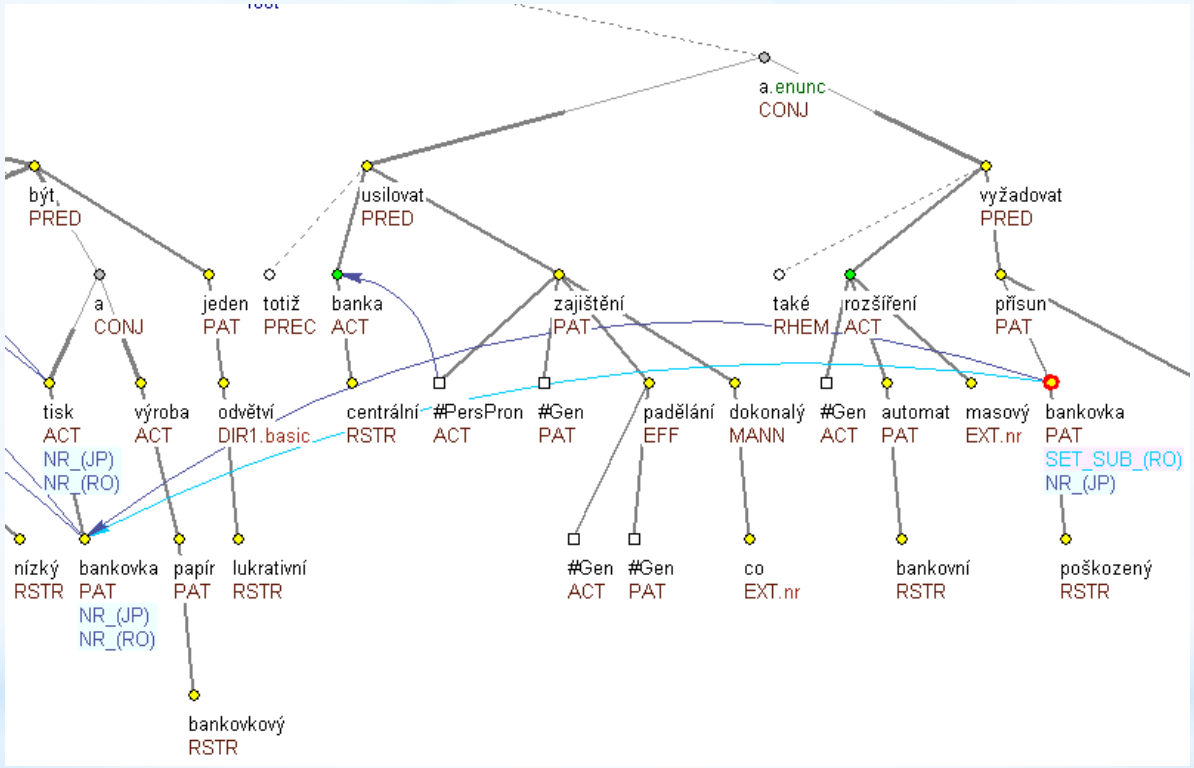
* Distinguishing between the bridging relations and the textual coreference

I přes klesající inflaci ve světě … je tisk bankovek a výroba bankovkového papíru jedním z nejlukrativnějších odvětví. […] *… Rozšíření bankovních automatů vyžaduje neustálý přísun nepoškozených bankovek.*

coreference (GEN) vs. bridging SUBSET

(= *Although inflation in the world rather decreases, … printing banknotes and production of banknote paper is still one of the most profitable areas. Mass expansion of ATMs calls for permanent increase of undamaged banknotes.*)

Distinguishing between the bridging relations and the textual coreference



I přes klesající inflaci ve světě … je tisk bankovek a výroba bankovkového papíru jedním z nejlukrativnějších odvětví. […] … Rozšíření bankovních automatů vyžaduje neustálý přísun nepoškozených bankovek.

(= Although inflation in the world rather decreases, … printing banknotes and production of banknote paper is still one of the most profitable areas. Mass expansion of ATMs calls for permanent increase of undamaged banknotes.)

* borderline cases between “specific” and “generic” coreference

U detergentu Toto jsme například řešili problém s udržením stálé kvality, protože jednotlivé partie byly nevyvážené. Investovali jsme dva miliony korun do nákupu pásových vah, zpřesnili dávkování a jakost pracího prášku stabilizovali.

engl. For example, for detergent Toto we thought about the problem of supporting the same quality …. We … made the dosage more exact and so we set the quality of washing powder.

In ambiguous cases between specific and generic co-reference, we choose specific co-reference.

Začal jsem provozováním hospody, která byla mnohokrát vykradena. [… 2 věty …] Hospoda byla jen startem, polem k podnikání s masem a masnými výrobky.

lit. engl. I began by carrying out a restaurant… […] A/the restaurant was just the beginning […]

*borderline cases between “specific” and “generic” coreference

K tématu pořadu TV NOVA TABU “Zrak za bílou hůl” byl přizván ke konzultaci Oldřich Čálek. Kateřina Hamrová, dramaturgyně pořadu, TV NOVA. (= *To consult the topic of the TV NOVA show TABU "Vision for a white cane", Oldřich Čálek was invited. Catherine Hamrová, the dramatist of the show, TV NOVA*)

Nic z toho se však nevyrovná míře neštěstí, které Romy postihlo v letech druhé světové války. Spolu se Židy byli označeni za méněcennou rasu a stali se objektem patologických fašistických opatření, jejichž cílem byla úplná genocida tohoto národa. (= *Nothing of this, however, compares to the misfortune that befell the Gipsies during the Second World War. Together with the Jews, they were called an inferior race and became the object of pathological fascist measures, their purpose being the complete genocide of the nation.*)

*Problem Cases - Reasons

different understanding of the content		
mostly don't have influence on understanding the text as a whole	"depth" of interpretation	guidelines "formalism"

Tak je knížka koncipována. V každé kapitole se mluví o určitém problému, uvádíme jak je rozsáhlý, kolik dětí je jím postiženo a co dělat. Je tam v podstatě konkrétní návod.
*This is the way **this book** is organised. **Every chapter** concerns a certain problem ... . There are actually specific instructions **there**.*

*I přes klesající inflaci ve světě … je tisk bankovek a výroba bankovkového papíru jedním z nejlukrativnějších odvětví. […] … Rozšíření bankovních automatů vyžaduje neustálý přísun nepoškozených bankovek. (= Although inflation in the world rather decreases, … printing **banknotes** and production of banknote paper is still one of the most profitable areas. Mass expansion of ATMs calls for permanent increase of **undamaged banknotes**.)*

Disagreement factors

- the text size
- degree of abstractedness of the text

Especially long texts with a large number of generic nouns, abstract and verbal nouns have the lowest inter-annotator agreement

- problematic are also
  - constructions with nouns of measure and time periods
  - generic noun phrases, abstract nouns and deverbatives
  - coreference between indefinite noun phrases

*Short text with 100% agreement

(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.

* Long text with low agreement

(11) Vaše kniha obsahuje ve třiadvaceti kapitolách různé problémy, od těžkých poškození dítěte až po lehčí disfunkci či vliv rozvodu na dítě.

(12) Tím ovšem jednu konkrétní rodinu může zajímat maximálně pět, přinejhorším deset kapitol.

(13) Zdeněk Matějček: Původně tato knížka byla určena pro zdravotnické pracovníky, a to především pro lékaře, kteří jsou ve styku s rodinou.

(14) Na druhé straně se ukázalo, že toto téma je stejně důležité pro pedagogy a vychovatele.

(15) Ti se přece setkávají i s postiženými nebo týranými dětmi.

(16) A když už byla knížka hotova, tak se zjistilo, že je praktická i pro rodiče.

(17) Samozřejmě ne každá kapitola ne pro každého rodiče.

(18) Zdeněk Dytrych: Kdyby se přímo dotýkalo některé rodiny deset kapitol, tak by to byla opravdu nešťastná rodina.

(19) Ale stačí jedna a většinou jich bude i víc.

(20) Vezměte si, kolik je rozvodů - třicet tisíc ročně v republice, to znamená, téměř třicet tisíc dětí je rozvodem nějakým způsobem postiženo.

(21) V této knize je poučení, jak snášejí děti rozvod a jak na něj reagují, a návod, jak se mají rodiče chovat, aby se utrpení dětí snížilo.

(22) Nebo například existuje lehká mozková disfunkce, kterou trpí podle našeho rozsáhlého výzkumu pět procent dětí.

(23) Toto postižení se velice špatně rozpoznává.

(24) Dítě je nemotorné, neklidné a není schopné se soustředit, ale přitom je většinou chytré.

(25) Rodiče ho považují za lajdáka a bývá trestáno třeba za špatný výkon ve škole, tím se zhoršuje vztah k učení atd.

(26) A tohle rodiče musí vědět.

(27) Samozřejmě i pedagogové a v této knížce je návod co s tím.

(28) Zdeněk Matějček: Předkládáme i problémy, na které se zapomíná.

(29) Tak například úmrtí dítěte nebo narození postiženého dítěte.

(30) Tady nejde jenom o rodiče, ale i o okolí, které musí vědět, jak se má chovat.

(31) Nebo úmrtí v rodině a jeho vliv na dítě a může to být třeba babička.

*Reasons for Disagreement – Abstract Nouns

one of very weak points in the PDT coreference annotaion

- attempted in PDT, also classified for specific and generic abstracts (e.g. according to the reference of valencies)
- actually my problem was that I couldn't reliably separate abstract nouns from concrete ones

*Preferuji širší předvedení s mnoha vnitřními souvislostmi, protože nám chybějí kritéria pro hodnocení **současné české výtvarné kultury**.* {… 11 sentences inbetween…} *Měli bychom se znovu pokusit … získávat **současné umění**, abychom jednou měli autentický soubor naší doby* (= *I prefer wider demonstration with many internal connections because we lack criteria for evaluation of contemporary Czech art*. {… 11 sentences inbetween …} *We should try ... to acquire the contemporary art again, in order to get an authentic set of our time.*)

antecedent is relatively far from the anaphoric NP

*Reasons for Disagreement – Abstract Nouns

*Tímto faktorem je podnikatel-inovátor, který se snaží o **zisk**, a proto logicky nemůže existovat ve stavu statiky, která nezná ani **zisk**, ani ztrátu.* (= *This factor is the enterpreneur-innovator, who is trying to gain profit, and hence, logically, cannot exist in a static state, where there is no profit or loss.*)

*Ve specifických podmínkách **české ekonomiky** růst nezaměstnanosti v letech 1991-1993 značně zaostal za poklesem HDP. […] Nejméně dvouprocentní růst **české ekonomiky** již letos.* (=*In the specific conditions of the Czech economy the growth of unemployment... This year at least a two percent growth of the Czech economy.*)

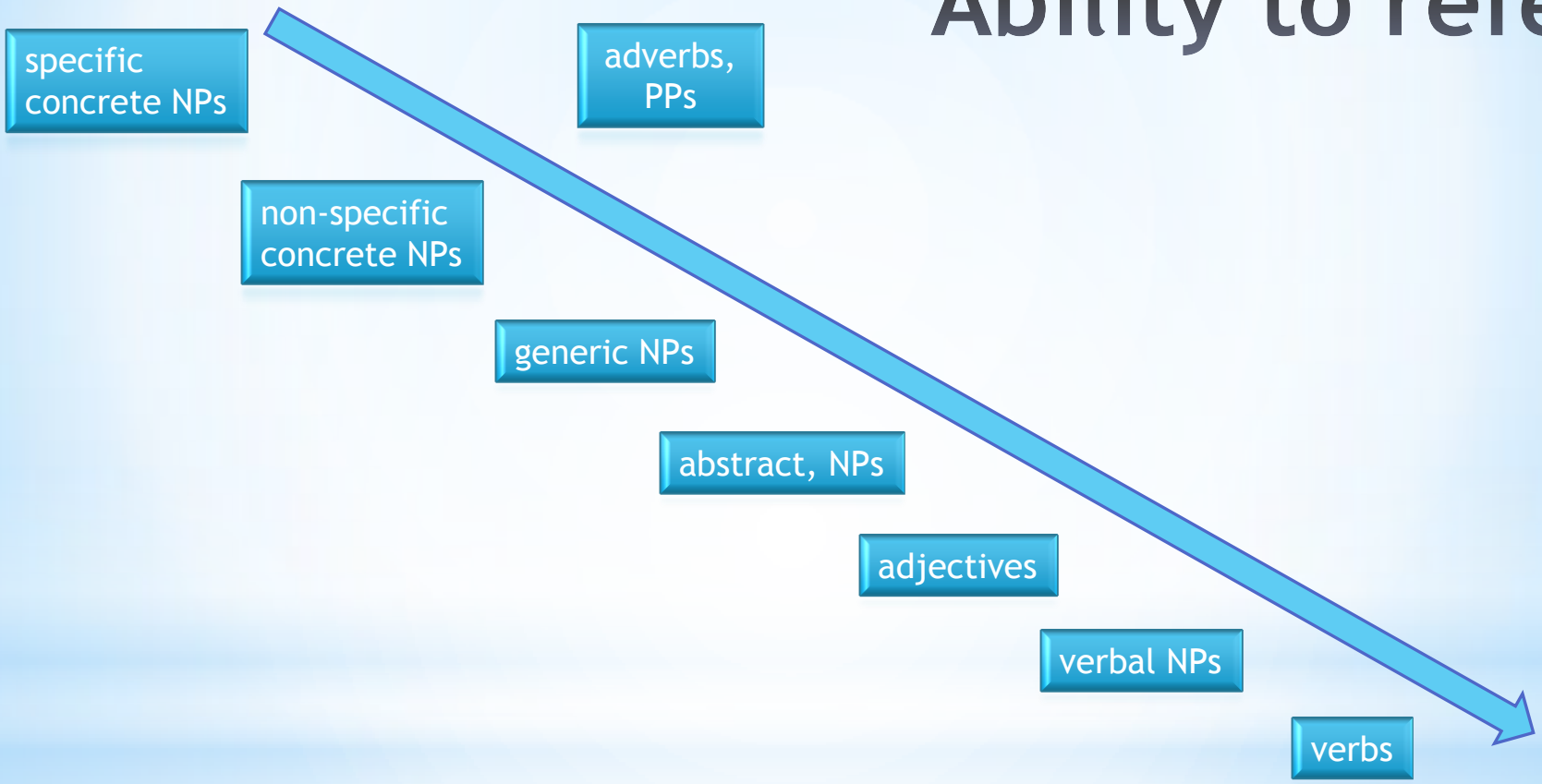
*In the Treasury market, investors paid scant attention to the day's economic reports, which for the most part provided a mixed view of **the economy**. ``Whether you thought **the economy** was growing weak or holding steady, yesterday's economic indicators didn't change your opinion," said Charles Lieberman, a managing director at Manufacturers Hanover Securities Corp.*

*Reasons for Disagreement – Verbal Nouns

*Vedení Pojišťovny Investiční a Poštovní banky nás upozornilo, že jejich pojišťovna nebyla zařazena mezi ty, které umožňují úrazové připojištění, ač tuto službu poskytují. Omlouváme se za **toto nedopatření**, dotyčná redaktorka byla pokutována.* (=*The Insurance Investment and the Post Bank management has notified us that their insurance company was not included among those that allow casualty insurance, although it provides this service. We apologize for this oversight, the editor who made the mistake was fined.*)

*Rychlé, avšak i bezpečné **vypořádání**. Rychlost **vypořádání** burzovních obchodů v čase odpovídá podle Jiřího Béra potřebám.* (= *Fast, yet safe transaction. According to Jiřího Bér's opinion, the speed of transaction corresponds to the needs.*)

*Ability to refer



* Reasons for Disagreement – measure NPs and other NPs with a 'container' meaning

skupina lidí (= a group of people)

počet akcií (= a number of stocks)

stádo krav (= a herd of cows)

dostatek financí (= abundance of finances)

milióny Židů (= millions of Jews)

sklenice piva (= a glass of beer)

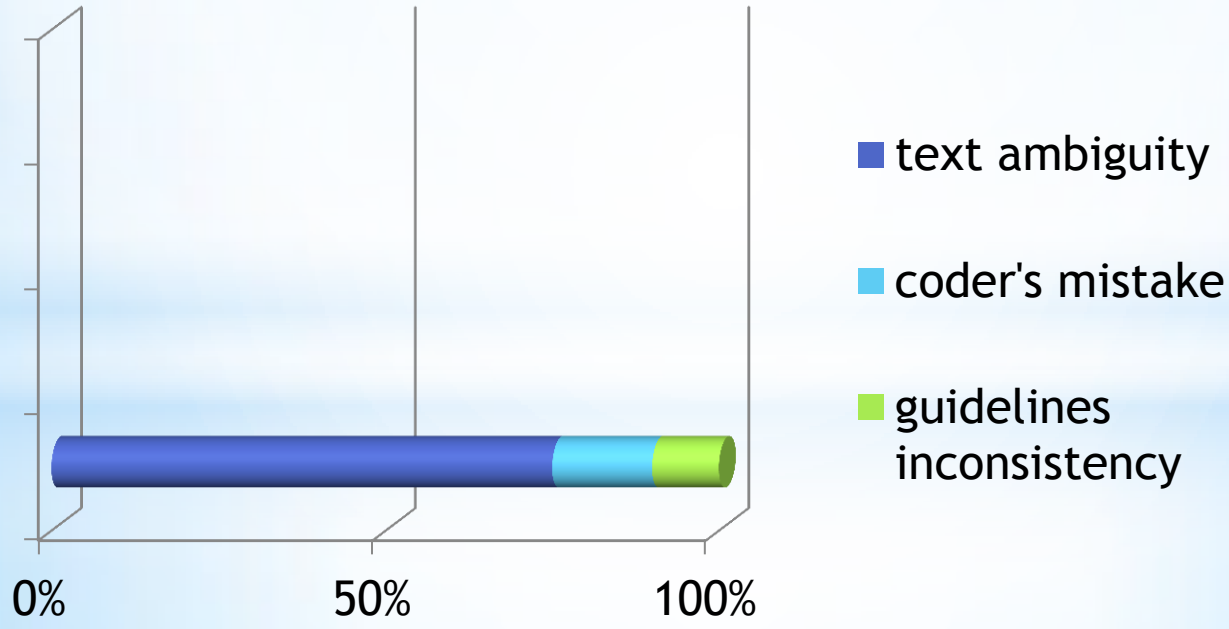
deset procent obyvatel (= ten percent of population)

* Reasons for Disagreement - Constructions with Time Periods

That compares with operating earnings of $132.9 million, or 49 cents a share, ***the year earlier****.*
The prior-year period *includes...*

* Reasons for Disagreement

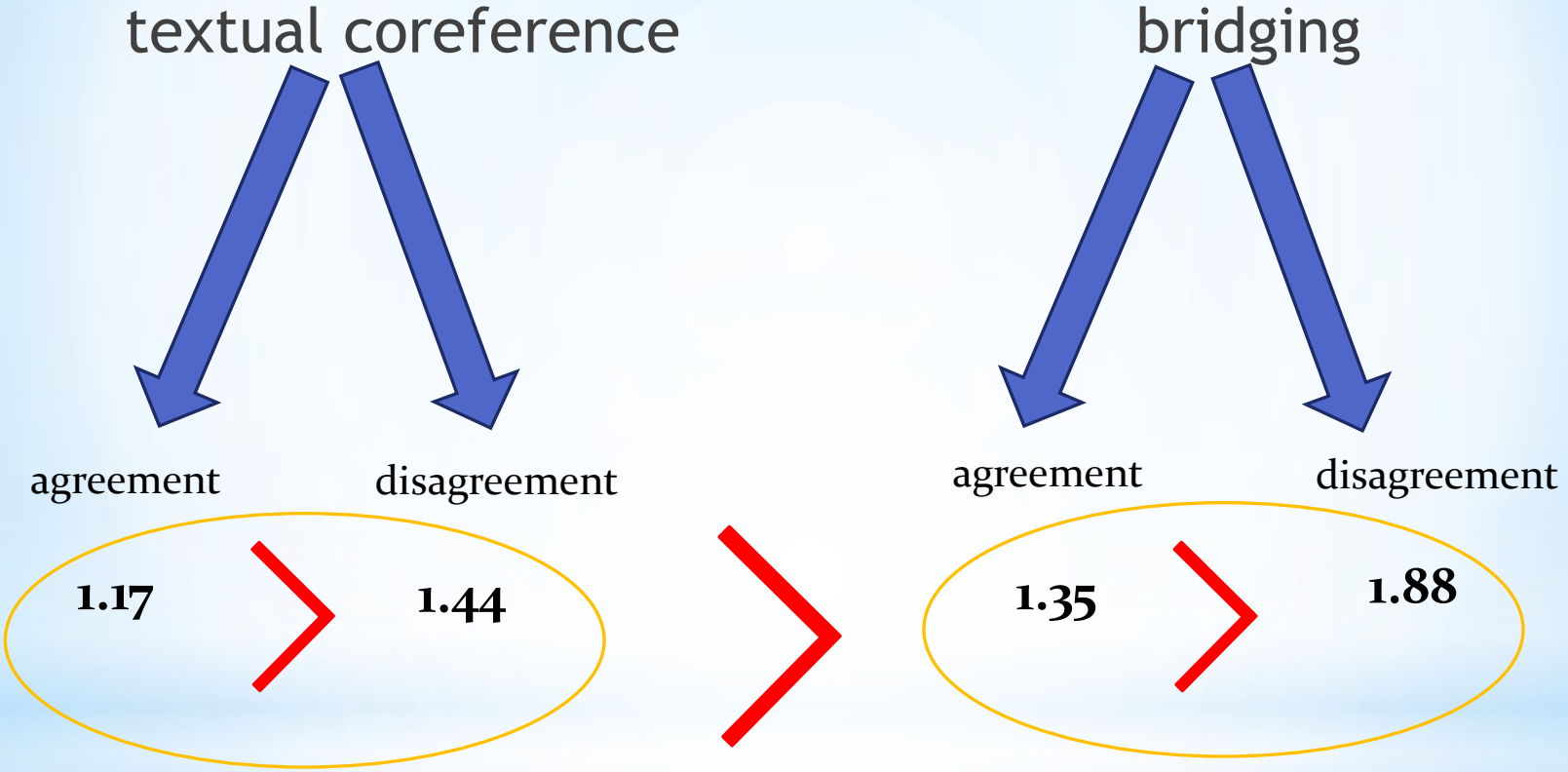
almost three fourth of the coders' disagreements come from the text ambiguity (empirically ambiguous or near-identical in the sense of Recasens (2010))



Experiment - Certainty of the manual annotations

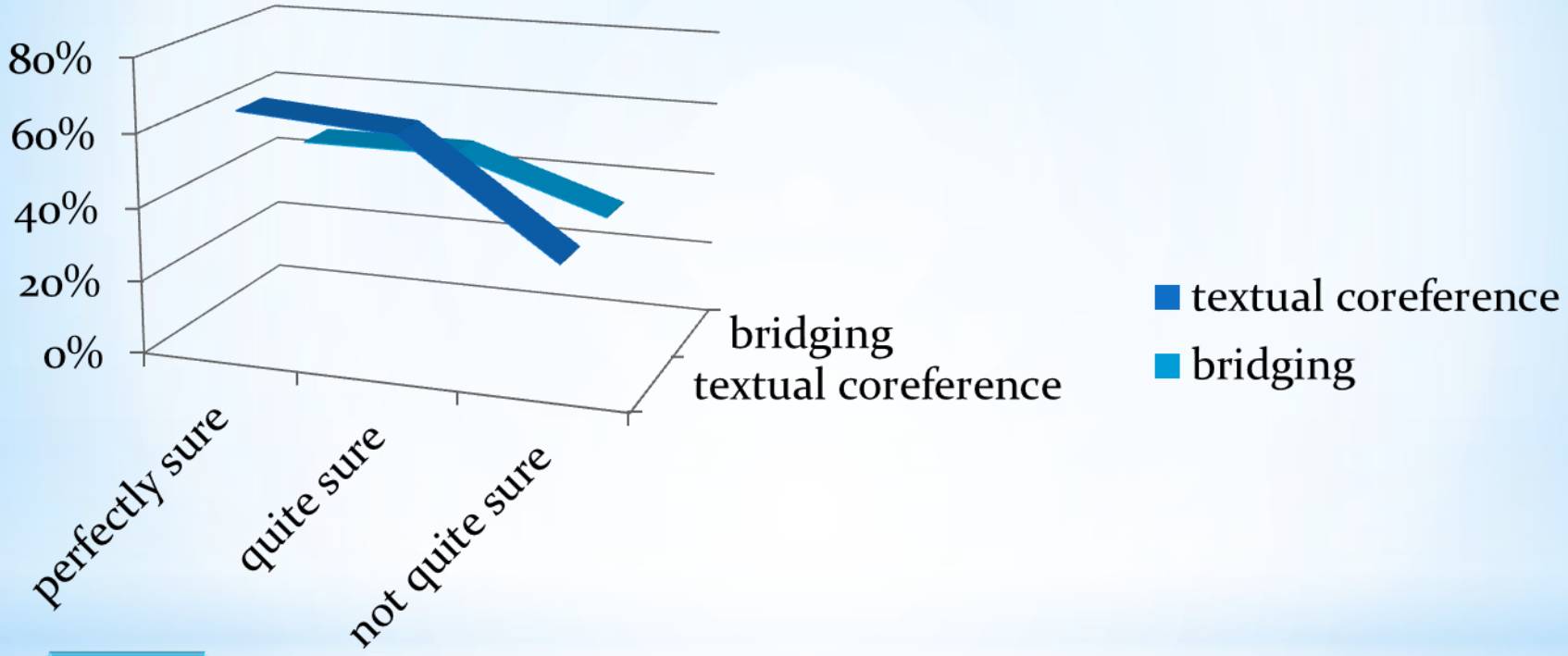
- annotators marked the certainty for their annotation decisions on the scale of 1 to 3
  - 1 : perfectly sure,
  - 2 : quite sure,
  - 3 : not quite sure
- certainty marked for
  - the presence of a relation,
  - selecting the antecedent,
  - distinguishing between the bridging relation and the textual coreference and
  - selecting the type of the bridging relation or the textual coreference

*Certainty in the Presence of a Relation



- naturally, the lower the agreement is, the less are the annotators sure
- the number of cases where the annotators didn't mark uncertainty but still disagreed exceeds all other cases (56 disagreements, only 26 were marked)

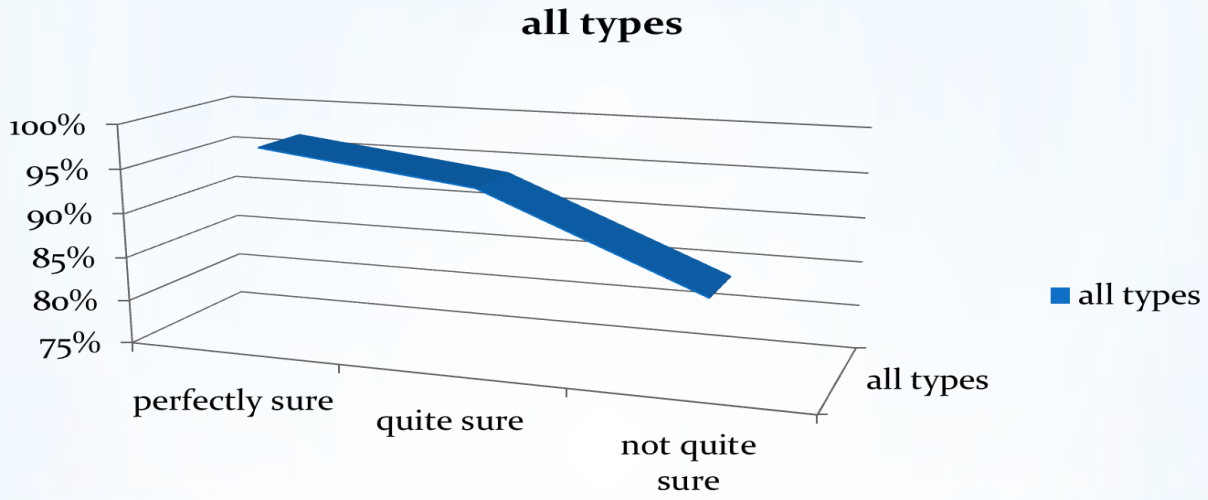
*Certainty in selecting the Antecedent



Again:

- the numbers show a lower agreement in cases where the annotators were not sure about the antecedent BUT
- from 27 disagreements in choosing the antecedent, only 16 were marked as uncertain by at least one annotator

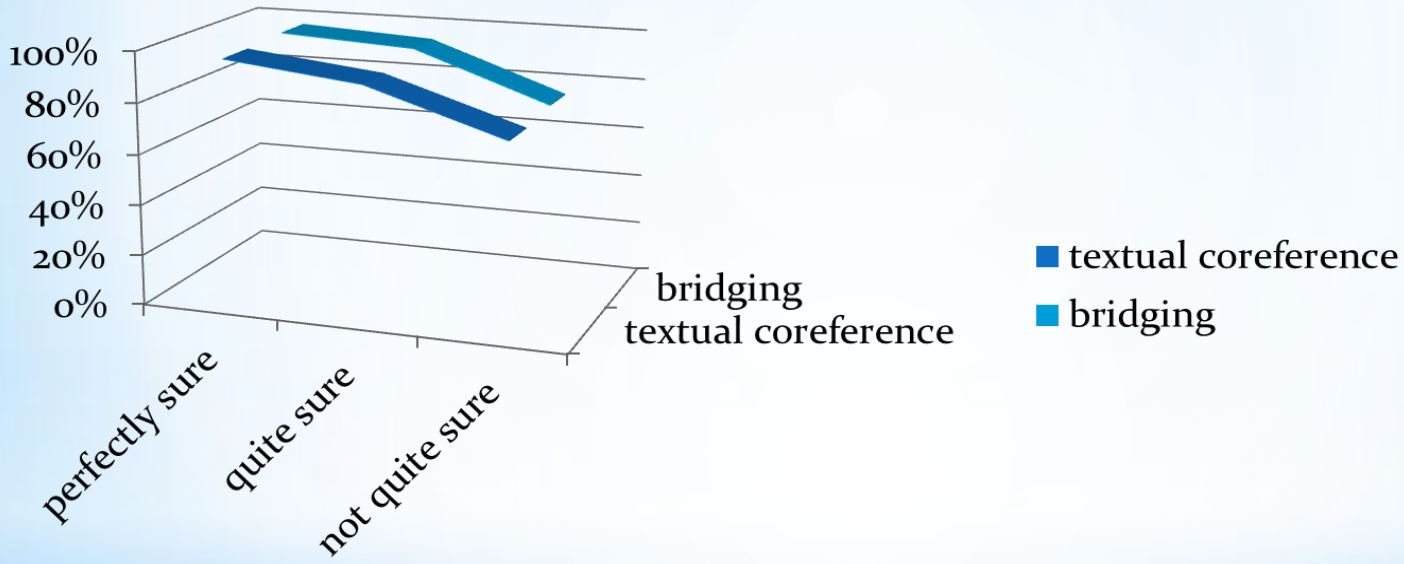
*Certainty in Distinguishing Between Coreference and Bridging



Numbers show:

- The difference in agreement between “certain” and “uncertain” relations in this case is not so relevant
- In most cases (21 of 32), the annotators marked ambiguity but still made the same decision

*Certainty in selecting the type of bridging or coreference relation



Again:

the numbers show a lower agreement in cases where the annotators were not sure about the type of the relation

* Results of the experiment and analysis

inter-annotator agreement + annotators' certainty reveal:

* empirical ambiguity is much more frequent on text level than on syntax level and lower
* the complexity of real corpus data which can never be reflected by any annotation guidelines
* ambiguity is frequently not detected by annotators
* in many cases world knowledge is needed
* annotators are more sure about relations between noun phrases in topic and contrastive topic than about those in focus

*Future plans?

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). 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).

Thank you for attention

http://ufal.mff.cuni.cz/discourse