CHARLES UNIVERSITY PRAGUE
faculty of mathematics and physics

LINDAT
CLARIN

Information Structure in an Annotated Parallel English-Czech Corpus: A Comparative Probe

Eva Hajičová, Jiří Mírovský, Kateřina Rysová and Magdaléna Rysová

Charles University, Prague

Starting hypothesis

the function of Information Structure:

- well comparable across languages (Prince 1981, Partee 1991, Steedman 2000, Krifka 2006)

=> in translations the IS of the source and of the target language sentences should be preserved

Prince (1981), Toward a taxonomy of given/new information; Partee (1991), Topic, Focus and Quantification; Steedman (2000), Information structure and the syntax-phonology interface; Krifka (2006), Basic Notions of Information Structure

Outline

1. Basic research questions
2. Theoretical background
3. Methodology and data
4. Analysis and results

1. Basic research questions

(1) How far does the assignment of Focus proper (= the last element of global Focus) agree in English and in Czech?

(2) If the assignment of Focus proper differs, is the Focus-proper element in English at least a member of the (global) Focus of the Czech sentence?

2. Theoretical background

the Praguian theory of Topic/Focus Articulation (TFA; Sgall 1967, Sgall et al. 1973, Hajičová et al. 1998)

(i) primary notion: contextual boundness

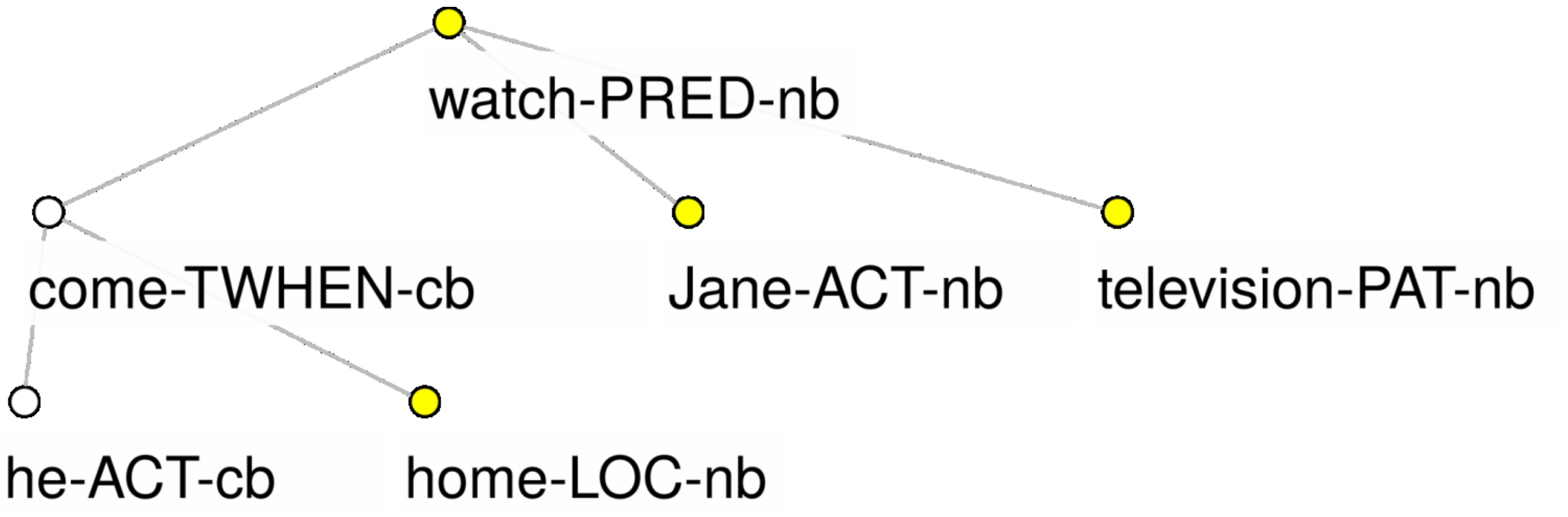
(ii) global bipartition: Topic – Focus

(iii) recursivity -> “local” topics and foci

(iv) contrastive topic

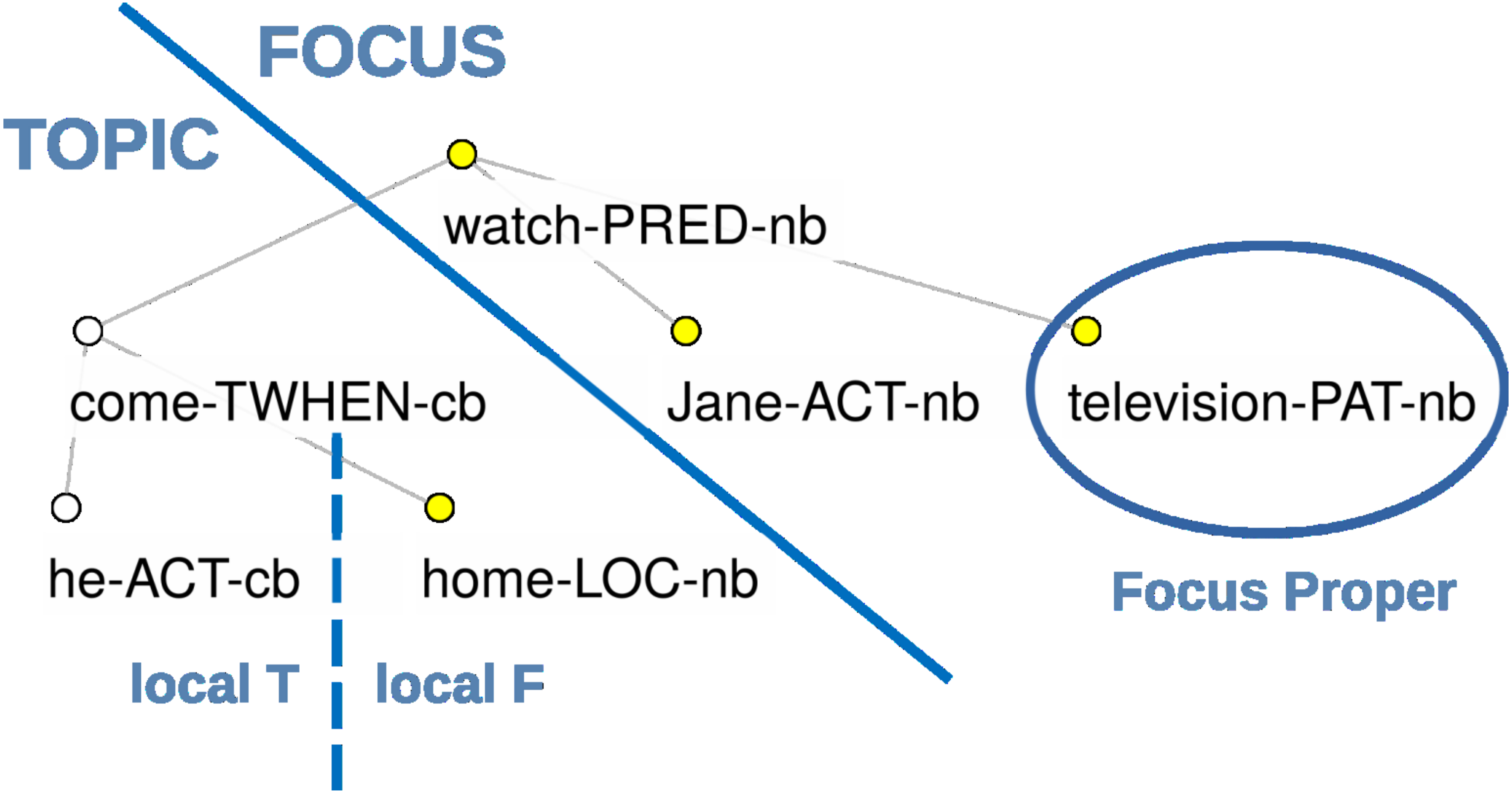
Semantic relevance, different means in different languages

Sgall (1967), *Functional Sentence Perspective in a Generative Description*; Sgall, Hajičová and Benešcvá (1973), *Topic, Focus and Generative Semantics;* Hajičová, Partee and Sgall (1998*), Topic-Focus Articulation, Tripartite Structures, and Semantic Content*



[Tom left his office after 6 o’clock.]

When he came home, Jane was watching television.



[Tom left his office after 6 o’clock.]

When he came home, Jane was watching television.

Contrastive studies

TFA: repeatedly tested on corpus material (Mírovský et al. 2013, Rysová et al. 2015, Hajičová and Mírovský 2018)

Contrastive studies: a parallel English(source)–Czech(target) corpus annotated both for deep syntactic structure and for TFA (PCEDT, Hajič et al. 2012)

Hajič et al. (2012): Announcing Prague Czech-English Dependency Treebank 2.0. (LREC 2012); Mírovský et al. (2013): (Pre-)Annotation of Topic-Focus Articulation in Prague Czech-English Dependency Treebank (IJCNLP 2013); Rysová et al. (2015): On an apparent freedom of Czech word order. A case study (TLT 2015); Hajičová, E., Mírovský, J. (2018): Discourse Coherence Through the Lens of an Annotated Text Corpus: A Case Study (LREC 2018).

Obstacles

(i) Information structure - a complex phenomenon, tricky to be annotated

(ii) In PCEDT: alignment on the basis of sentences, no alignment of individual words

➔

(i) Results of queries manually checked

(ii) Queries based on the search and comparison of (deep) syntactic values

3. Methodology and data

Both Czech (Mluvnice … 1987) and English (Quirk et al. 1985) representative grammars:

typical position of focus both for English and for Czech: end-focus

→comparison of English and corresponding Czech sentences as for the final position

→comparison of the syntactic value of the element in the final position of the dependency structure of the sentences

Corpus Data

The total number of automatically aligned sentences without coordination of the main predicates: 3857

out of which:

- 2514 cases (65,3%) with the same syntactic value at the last position of both source and target text (= Focus Proper)
- 1287 cases with a difference in this value

4. Research question (1)

- How far does the assignment of Focus proper (= the last element of global Focus) agree in English and in Czech?

(a) a manual inspection of the randomly selected 120 sentences (out of 1287) with a different syntactic label in the Focus Proper position

(b) filtering out cases where the syntactic label differed but the target lexical item corresponded to the source one => 24 examples of an actual difference in Focus Proper

Different syntactic label

Cataracts.ACT *refer to* a clouding.PAT of the eye’s natural lens

Šedým zákalem.PAT se *nazývá* ztmavnutí.ACT přirozených očních čoček.

The prospective buyers.ACT *included* investor.PAT Marvin Davis.

K potenciálním kupcům.PAT *patří* investor.ACT Marvin Davis..

An airline buy-out bill *was approved* by the House.ACT

Zákon o skupování aerolinek *prošel* Sněmovnou reprezentantů.DIR2

Analysis

(a) most frequently: the difference concerned the mutual position of the main predicate and its modification of time or place (in English, this modification frequently was in the Focus Proper)

(b) or the position of the predicate itself (in Czech, the predicate was in the Focus Proper)

(a) mutual position of the main PRED and its modification of TWHEN or LOC

Mr. Nixon / *met*.PRED Mr. Bush before coming to China on Saturday.TWHEN

V sobotu.TWHEN před odletem do Číny se Nixon / *setkal*.PRED s Bushem.

Both contracts / *have gained*.PRED a following since the 1987 global market crash.TWHEN

Od celostátního krachu trhů v roce 1987.TWHEN si obě smlouvy / *získaly*.PRED své stoupence.

Contrastive topic?

Accounting problems / *raise* more knotty issues.

Ještě.RHEM složitější problémy / jsou s účetnitcvím.

Many of the morning session winners / *turned out* to be losers by afternoon.

Mnoho vítězů z dopoledního obchodování se odpoledne / *změnilo* v poražené.

(b) the position of the predicate itself

In Czech: PRED may occupy the end position

New Zealand’s finance minister / lashed out.PRED at such suggestions.

Novozélandský ministr financí na takové názory / ostře útočí.PRED

As a private company, Random House / doesn’t report.PRED its earnings.

Random House jako soukromá společnost tyto zisky / neuvádí.PRED.

Other cases

E.: Grammatical rule requires the order Subj-Verb

If closing things could reduce volatility, stone tablets.ACT / should become the trade ticket of the future.PAT

Pokud by zpomalení procesu dokázalo zredukovat nestálost, pak by se budoucností obchodního světa.PAT / měly stát kamenné tabulky.ACT

Research question (2)

- Is the Focus Proper element in English at least a member of the (global) Focus of the Czech sentence?
- A manual filtering of the whole set of 171 sentences => a set of 30 sentences in which the element assigned Focus proper in English does not appear even in the global Focus part of the target Czech sentences

Member of global Focus

His longer analysis / *is to appear* in the Duke Law Journal.LOC later this year.TWHEN.

Jeho delší analýza / *by měla být publikována* v letošním roce.TWHEN v Duke Law Journal.LOC.

Chemical Bank / *spent* more than 58 million dollars to introduce its ChemPlus line according to Thomas Jacob in 1986.TWHEN

Chemical Bank / *utratila* v roce 1986.TWHEN podle Thomase Jacoba přes 58 milionů dolarů za zavedení své řady ChemPlus.

The company / *offered* two round-trip tickets.PAT to buyers of its Riviera luxury car.

Tato společnost / *nabízela* kupcům svého luxusního vozu Riviera dva lístky na okružní výlet.PAT

Analysis

(a) Most differences concerned temporal or local modifications which in the Czech sentence appeared in the Topic rather than in the global Focus.

(b) A second group of examples concerned the mutual position of the main predicate and its argument Patient: Patient was “topicalized” in Czech.

(a) Temporal or local modifications

- in the Czech sentence appeared in the Topic rather than in the global Focus: cf. examples above

Mr. Nixon *met*.PRED Mr. Bush before coming to China on Saturday.TWHEN

V sobotu.TWHEN před odletem do Číny se Nixon *setkal*.PRED s Bushem.

Both contracts *have gained*.PRED a following since the 1987 global market crash.TWHEN

Od celostátního krachu trhů v roce 1987.TWHEN si obě smlouvy *získaly*.PRED své stoupence.

(b) Main predicate and its argument

Patient is “topicalized” in Czech:

As a private company, Random House /doesn’t report.PRED its earnings.PAT

Random House jako soukromá společnost tyto zisky.PAT / neuvádí.PRED.

This / does not sit well.PRED with some clerics.PAT

To některým duchovním.PAT / nesedí.PRED

Some contexts offer both interpretations

[Context: A lot of people woud like to go back to 1979. Mr. Phalan said this week: I would like to go back to 1979.]

But we are not going back to 1979.

Jenže my se do roku 1979 nevrátíme.

(i) IC on *NOT* -> *to 1979* in TOPIC

(ii) IC on *1979* -> difference (*go back* -> phrase “vrátit se”)

Conclusions

(1) The differences in IS between the examined languages are rather rare

(2) Certain tendencies can be observed:

(a) the temporal and local modifications of the main predicate

(b) the position of the predicate itself

(c) some contexts actually offer both interpretations of the IS structure

Future work

- Analysis of the mutual order of temporal and local modifications of predicates to test:
  - the variability of the order of the given types of modifications in general
  - testing two hypotheses on their preferential order in particular, namely the SVOMPT hypothesis for English and the so-called systemic ordering hypothesis for both languages

CHARLES UNIVERSITY PRAGUE
faculty of mathematics and physics

LINDAT
CLARIN

THANK YOU FOR YOUR ATTENTION!

Questions?

Acknowledgements:

This work has been supported and using language resources and tools distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16 013/0001781). The authors also gratefully acknowledge support from the Grant Agency of the Czech Republic (project GA17-06123S).