Information Structure in written English
- a corpus study -

Oana Postolache
oana@coli.uni-saarland.de

IGK colloquium – 8 Dec 05

Information Structure (IS)

- Division of the sentence in two parts:
  1. Links the sentences to the discourse
  2. Advances the discourse (brings new information)

Rob needs to talk things out, and he certainly isn't going to do that with Dick or Barry. So, he talks to HIMSELF instead.

Topic Focus Topic

- Not the given/new distinction

Thesis Goal

- Develop computational methods to automatically detect IS for naturally occurring English sentences.

- Trial 1:
  - Use the PDT to develop a system that detects Topic & Focus for Czech.
  - Use a parallel corpus to transfer Topic & Focus to English, through word alignment (in order to create an English corpus).

- Trial 2: Investigation of English corpora.

Realization of IS in English

- Intonation
- Non-canonical word order
  - **Gregory Ward & Betty Birner studies**:
    - 1998 – **Information Status** and Non-canonical Word Order in English
    - 2001 – Discourse and **Information Structure**
    - 2004 – **Information Structure** and Non-canonical Syntax
  - Distinguish 5 types of non-canonical constructions which impose constraints on the IS of the sentence:
    - preposing, left-dislocation, postposing, right-dislocation and inversion
  - Their corpus consists in several thousands naturally occurring sentences collected over approx. 10 years.

What is this talk about?

- Consider 2 corpora:
  - WSJ – news (1,107,392 words)
  - “1984” – belletristic (104,136 words)

- Investigate:
  - How often these non-canonical constructions appear?
  - Do they comply with Ward & Birner constraints?
  - What is their Information Structure?

Outline

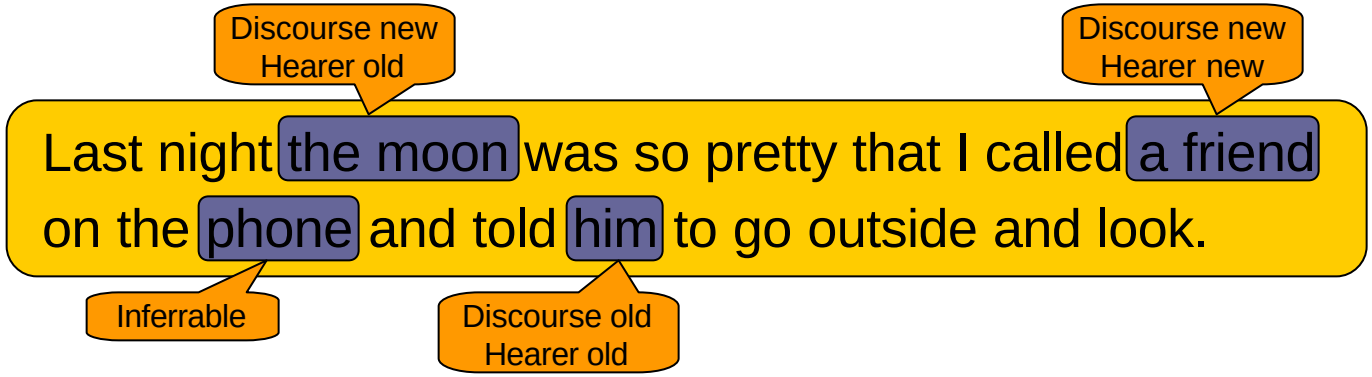
- Background
  - Information Status (vs Information Structure)
  - POSET relationship
  - Focus / Open-proposition theory
- 5 non-canonical constructions
  - Definition and exemplification
  - Ward & Birner constraints
  - Information Structure
  - Occurrence in corpora
- Summary

Outline

- Background
  - **Information Status (vs Information Structure)**
  - POSET relationship
  - Focus / Open-proposition theory
- 5 non-canonical constructions
  - Definition and exemplification
  - Ward & Birner constraints
  - Information Structure
  - Occurrence in corpora
- Summary

Information Status

- Regards the discourse familiarity or the hearer familiarity of an entity or event
  - Discourse-new / Discourse-old
  - Hearer-new / Hearer-old
  - Inferrable



Prince 1981, 1992

Outline

- Background
  - Information Status (vs Information Structure)
  - **POSET relationship**
  - Focus / Open-proposition theory
- 5 non-canonical constructions
  - Definition and exemplification
  - Ward & Birner constraints
  - Information Structure
  - Occurrence in corpora
- Summary

POSET relationship

- Linking relations: identity, type/subtype, entity/attribute, part/whole, etc.
- A POSET (Partially Ordered SET) is any set defined by a transitive partial ordering linking relation.

– Do you like this album?
– Yeah, **this song** I really like.

Relation = is-part-of, POSET = {album parts}

– Have you filled out the Summary Sheet?
– Yes, **both the Summary Sheet and the Recording Sheet** I've done.

Relation = is-a-member-of, POSET = {forms}

– Did you get any more answers for the crossword puzzle?
– No, **the cryptogram** I can do like that; the crossword puzzle is hard.

Relation = is-type-of, POSET = {newspaper puzzles}

Outline

- Background
  - Information Status (vs Information Structure)
  - POSET relationship
  - **Focus / Open-proposition theory**
- 5 non-canonical constructions
  - Definition and exemplification
  - Ward & Birner constraints
  - Information Structure
  - Occurrence in corpora
- Summary

Focus / Open-proposition theory

- Open-proposition (OP): the information in the sentence that is assumed by the writer to be shared by him and the reader.
- Focus: the complement of this presupposition.

I promised my father – on Christmas Eve it was – to kill a Frenchman at the first opportunity.

OP = It was X, where X ∈ {times}

X = on Christmas Eve

- What constitutes new information is the fact that a particular focus instantiates the variable in the open-proposition.

Prince 1981, 1984, 1986

Outline

- Background
  - Information Status (vs Information Structure)
  - POSET relationship
  - Focus / Open-proposition theory
- **5 non-canonical constructions**
  - Definition and exemplification
  - Ward & Birner constraints
  - Information Structure
  - Occurrence in corpora
- Summary

Preposing

1st non-canonical construction:

Preposing

- A canonically postverbal constituent appears in preverbal position.
- Restriction to lexically governed constituents.

In a basket, I put your clothes.

~~**In New York**, there's always something to do.~~

1st non-canonical construction:
Preposing – W&B Constraint

- The referent of the preposed constituent must be anaphorically linked to the previous discourse.
- The constituent is an element of a POSET which is salient or inferred.

  The POSET may contain only 1 element, the constituent, when it refers to a previous discourse entity.

1st non-canonical construction:

Preposing – Constraint Illustration

In principle, he is now capable of carrying out or determining the accuracy of any computation. ***Some computations*** *he may not be able to carry out in his head*. Paper and pencil are required.

POSET: {set of computations}

But keep in mind that no matter which type of equipment you choose, a weight-training regimen isn't likely to provide a cardiovascular workout as well. ***For that*** *you have to look elsewhere*.

POSET: {that = to provide a cardiovascular workout}

1st non-canonical construction:

Preposing – Information Structure

- Focus preposing

  Colonel Kadafy, you said you were planning on sending planes – ***M-16s*** *I believe they were* – to Sudan.

  OP: The planes were of type X, where X∈{types of military aircraft}
  Focus: X=M-16s

- Topicalization

  G: Do you like football?
  E: Yeah. ***Baseball*** *I like* *a lot better*.

  OP: I like to X degree {sports}, where X∈{degrees}
  Focus: X = a lot better

OP = Open proposition

1st non-canonical construction:
Preposing in Corpora

- In W&B corpus: 915 examples

	No. of examples	Is POSET salient / inferrable?		Information Structure		Is OP salient / inferrable?	
		yes	no	Focus	Topic	yes	no
WSJ 1.1 mil words	24	17	7	14	10	10	14
1984 0.1 mil words	68	39	29	29	39	28	40

Left-dislocation

2nd non-canonical construction:
Left-dislocation

- Preposing, but a referential pronoun is present in the canonical position of the preposed constituent.

***One of the guys I work with**, **he** said he bought over $100 in Powerball tickets.*

2nd non-canonical construction:

Left-dislocation - Constraints

- Simplifying left-dislocation

  The constituent is a discourse-new entity placed in a preposed position in order to simplify the discourse processing.

  I bet she had a nervous breakdown. That's not a good thing. ***Gallstones****, you have* ***them*** *out and they are out*. But a nervous breakdown, it's very bad.

- Left-dislocation triggering a POSET inference

  In her project, she'll use three groups of mice. One, she'll feed them mouse chow, just the regular stuff they make for mice. ***Another****, she'll feed* ***them*** *veggies*. And the third she'll feed junk food.

  POSET = {three groups of mice}

2nd non-canonical construction:

Left-dislocation – Inform. Structure

- The preposed constituent is Topic, the rest is Focus.
- In simplifying left-dislocation, we encounter examples of Topic that contains discourse new entities!

I bet she had a nervous breakdown. That's not a good thing. ***Gallstones***, *you have **them** out and they are out*. But a nervous breakdown, it's very bad.

In her project, she'll use three groups of mice. One, she'll feed them mouse chow, just the regular stuff they make for mice. ***Another***, *she'll feed **them** veggies*. And the third she'll feed junk food.

2nd non-canonical construction:
Left-dislocation in corpora

	No. of examples	Type	
		Simplifying	POSET triggering
WSJ 1.1 mil words	5	2	2
1984 0.1 mil words	11	8	3

Exception:

***A lifelong revolutionary with little education who fought both the French and the U.S.-backed Saigon regime**, **she** switched effortlessly to commerce after the war.*

Postposing

3rd non-canonical construction:

Postposing

- A canonically preverbal constituent (subject) is placed after the verb. Three types of postposing:
  - *Existential there*

    In Ireland's County Limerick, near the River Shannon, *there is* ***a quiet little suburb by the name of Garryowen***, which means "Garden of Owen".

  - *Presentational there*

    Daniel told me that shortly after Grumman arrived at Wideview Chalet *there arrived also* ***a man named Sleeman***.

  - *Extraposition*

    It was a shock to me ***that a bloodthirsty, cruel capitalist should be such a graceful fellow***.

3rd non-canonical construction:
Postposing - Constraints

- *Existential there*: the postverbal NP must be Hearer-new

  In Ireland's County Limerick, near the River Shannon, *there is* ***a quiet little suburb by the name of Garryowen***, which means "Garden of Owen".

- *Presentational there*: the postverbal NP must be Discourse-new

  Daniel told me that shortly after Grumman arrived at Wideview Chalet *there arrived also* ***his father***.

- *Extraposition*: the canonical variant is constrained - it is only possible when the embedded subject is Hearer-old; if it is new, extraposition is required.

  That a bloodthirsty, cruel capitalist should be such a graceful fellow *was a shock to me.*

3rd non-canonical construction:
Postposing – Information Structure

- *Existential & presentational there*: All Focus

  In Ireland's County Limerick, near the River Shannon,
  there is ***a quiet little suburb by the name of Garryowen***,
  which means "Garden of Owen".

- *Extraposition*:
  - Usually: All Focus

    Tom is not a very good student.
    It's a miracle ***that he turn in a term paper at all***.
  - Sometimes (theoretically): the extraposed part can be Topic

    Tom didn't turn in his term paper until a week after the deadline.
    It's a miracle ***that he turn in a term paper at all***.

- *Canonical variant of extraposition*: embedded subject - Topic

  That a bloodthirsty, cruel capitalist should be such a graceful fellow
  was a shock to me.

3rd non-canonical construction:
Postposing in corpora

	Existential there			Presentational there			Extraposition			Canonical variant of extraposition		
	All	Hearer new?		All	Discourse new?		All	IS		All	Hearer old	
		yes	no		yes	no		All Focus	with Topic		yes	no
WSJ 1.1 mil words	1,079	1,079	0	0	0	0	659	659	0	16	12	4
1984 0.1 mil words	446	431	15	19	19	0	311	311	0	16	14	2

Right-dislocation

4th non-canonical construction:
Right-dislocation

- A canonically preverbal constituent (subject) is placed in a postverbal position, while a referential pronoun is placed in the canonical position.

They *are really enormous*, ***those pipes***.

- *Constraint*: the postponed constituent must be Discourse-old

- *Information Structure*: the constituent – Topic, the rest – Focus.

4th non-canonical construction:
Right-dislocation in corpora

	No. of examples
WSJ 1.1 mil words	0
1984 0.1 mil words	1

***It**'s a beautiful thing, **the destruction of words**.*

? ***He**'s ever so good with hands, **Tom** is.*

Argument Reversal

5th non-canonical construction:
Argument Reversal

- Displacement of two arguments.
- Two types:

  - *By-passives*

    The mayor's present term of office expires Jan.1. ***He*** *will be succeeded* ***by Ivan Allen Jr****.*

  - *Inversion*

    George can you do me a favor? ***Up in my room, on the nightstand****, is* ***a pinkish-reddish envelope that has to go out immediately****.*

5th non-canonical construction:

Argument Reversal - Constraints

- *By-phrase*: the syntactic subject must not represent newer information within the discourse than does the NP in the by-phrase.

  The mayor's present term of office expires Jan.1.
  He *will be succeeded* ***by Ivan Allen Jr****.*

- *Inversion*: the preposed constituent must be more familiar than the postposed constituent.

  George can you do me a favor? ***Up in my room, on the nightstand****, is* ***a pinkish-reddish envelope that has to go out immediately****.*

5th non-canonical construction:
Argument Reversal - IS

- Usually:
  - The preposed constituent is Topic and the rest is Focus.
- It can also be:
  - All Focus

The mayor's present term of office expires Jan.1.
He *will be succeeded* ***by Ivan Allen Jr****.*

George can you do me a favor? ***Up in my room, on the nightstand,*** *is* ***a pinkish-reddish envelope that has to go out immediately****.*

5th non-canonical construction:
Argument Reversal in corpora

	By-phrase			Inversion		
	All	Is A more familiar than B		All	Is A more familiar than B	
		yes	no		yes	no
WSJ 1.1 mil words	3,138	3,131	7	280	274	6
1984 0.1 mil words	90	86	4	31	23	8

A = preposed constituent
B = postposed constituent

Summary

- The aim of this study had 3 goals:
  1. How many of these constructions are in corpora?

	Preposing	Left Dislocation	Postposing	Right Dislocation	Argument Reversal
WSJ	24	5	1,754	0	3,418
1984	68	11	776	1	133

  2. Do they comply with Ward & Birner constraints?
     Do not comply:

WSJ	7	0	4	0	13
1984	29	0	17	0	12

  3. What is their Information Structure?
     - Does the syntactic construction triggers a certain IS?

       YES: Left-Dislocation, Postposing, Right-Dislocation

       NO: Preposing, Argument Reversal

Thank you!

Preposing: the POSET is not inferred

- Manville, having rid itself of asbestos, now sells fiberglass, forest products, minerals and industrial goods. ***Heady stuff*** *it's not*.

- He has put some of his aesthetic ideas into practice with his design of the four-star Quilted Giraffe restaurant -- ``***architecturally impeccable***,'' *Progressive Architecture magazine called it* -- and his remodeling of Paul Stuart, the Madison Avenue clothing store.

- But major packaged-goods players of the world -- such as Procter & Gamble, Colgate-Palmolive and Unilever -- have steadfastly eluded the agency. *"**Three of our favorite names**," Mr. Della Femina calls that roster* …

- The instrument (***the telescreen***, *I was called*) could be dimmed …

- He was the commander of a vast shadowy army, an underground network of conspirators dedicated to the overthrow of the State. ***The Brotherhood***, *its name was supposed to be*.

Existential there: not Hearer new

- One had the impression that there were dust in the creases of her face. […] In the better light of the living-room he noticed with interest that *there actually was* ***dust in the creases of her face***.

- Suddenly there sprang into his mind, ready made as it were, the image of a certain Comrade Ogilvy. [..] It was true that *there was* ***no such person as Comrade Ogilvy***.

- There was also something called the jus primae noctis, which would probably not be mentioned in a textbook for children. […] For all he knew *there might never have been any* ***such law as the jus primae noctis*** …

Canonical version of extraposition: not Hearer old

- It never ceases to amaze me how the business world continues to trivialize the world's environmental problems ("Is Science, or Private Gain, Driving Ozone Policy?" by George Melloan, Business World, Oct. 24). ***To suggest that a 10% drop in ozone by the middle of the next century would be negligible*** *is irresponsible and shortsighted*.

- In the long run, a hierarchical society was only possible on a basis of poverty and ignorance. ***To return to the agricultural past , as some thinkers about the beginning of the twentieth century dreamed of doing,*** *was not a practicable solution.*

By-Phrase

The preposed is NOT more familiar than the postposed.

- The telescreen received and transmitted simultaneously. ***Every sound that Winston made, above the level of a very low whisper****, would be picked up* ***by it*** *…*

- The heirs of the French, English, and American revolutions had partly believed in their own phrases about the rights of man, freedom of speech, equality before the law, and the like, and have even allowed ***their conduct*** *to be influenced* ***by them*** to some extent.

- ***The word well****, for example, was replaced* ***by goodwise***.

- Xtra , a transportation leasing company , said in a statement 0 it would have no comment on Mr. Gintel 's plans until `` further information has been disclosed * by him . "

Inversion

The preposed is NOT more familiar than the postposed.

- ***Following*** *is* ***a weekly listing of unedited net asset*** *values of publicly traded investment fund shares , reported by the companies as of Friday 's close.*

- ***Conspicuous by its absence*** *is* ***California****.*

- ***Out of the mouths of revolutionaries*** *are coming* ***words of moderation****.*

- ***Never again*** *will* ***you*** *be able of ordinary human feeling.*