Bridging Corpus for Russian

Denis Khachko,
Institute Of Mathematical Problems In Biology,
Russia
Mordol@lpm.org.ru

Anna Roitberg,
School Of Linguistics In High State Of Economics,
Institute Of Mathematical Problems In Biology
Russia
Cvi@yandex.ru

March,7 2015
Valladolid

Some Bridging Corpora

Corpus	Language	Link
VENEX corpus	Italian	http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf
Copenhagen treebank	Danish	http://www.linguistics.ruhr-uni-bochum.de/bla/beyondsem2011/korzen_final.pdf
Prague Dependency Treebank 2.0 (PDT 2.0)	Czech	http://ufal.mff.cuni.cz/~nedoluzko/koref_anot/DAARC_Nedoluzhko.pdf
SemDok corpus of German scientific articles	German	http://media.dwds.de/jlcl/2008_Heft2/Baerenfaenger,Goecke,Hilbert,Luengen,Stuehrenberg.pdf

Anaphora. Basic Terms

- Anaphor –word or phrase refers back to an earlier word or phrase

- Antecedent - An earliery entity to which another word refers back.

The men came. **He** was very tall.

The men = antecedent

He = anaphor

Direct Anaphora vs. Indirect Anaphora

- **Direct anaphora**. Anaphor and Antecedent are coreferential
  - A man came. **He** was tall.
    He is coreferential to *A man*
- **Indirect anaphora** Anaphor and Antecedent aren't coreferential
  - I came to a room. **The walls** were white.
    The walls isn't coreferential to *a room*. But there are some anaphoric relations.

Bridging Anaphora

Bridging

(indirect or associative anaphora)

An anaphoric relations between two non-coreferential elements

Bridging Basic Terms

Bridging	Anaphora
Bridging element	Anaphor
Anchor	Antecedent, Anchor
Bridging link	Anaphoric link

Genitive construction in Russian

N_1 + (Pron.Gen) + (Adj.Gen) + N_2 Gen

(1) [Plat'e [sestry]$_{Gen}$] - 'sister's dress'

'dress'.Nom + 'sister'.Gen

(2) [Plat'e [moej starshej sestry]$_{Gen}$]

'dress'.Nom + 'my'.F.Gen + 'eldest'.F.Gen + 'sister'.Gen

'My eldest sister's dress'

Genitive construction in Russian

Not only possessive. For example:

- Litr moloka - ‘a liter of milk’

  ‘Liter’ ‘milk’.Gen

- Ministr pravitelstva - ‘a minister of government’

  ‘minister’ ‘government’.Gen

- Vozrogdenie gorodov - ‘urban renewal’

  ‘renewal’ ‘urban’.Pl.Gen

Genitive Construction and Bridging

NB. A genitive dependent can be missed out if in the previous text its coreferential item had appeared.

V avtobuse nachalsya pojar. **Voditel'** (~~avtobusa~~) potushil ogon'.

'In' 'bus' 'start' 'fire' 'driver' (~~'bus'.Gen~~) 'put out' 'fire'

'The fire began in a bus. The driver (~~of the bus~~) putted out the fire '

Bridging in Genitive Constructions

We consider just on bridging in genitive construction.

➡ "Genitive Bridging"

V avtobuse nachalsya pojar. Voditel' (~~avtobusa~~) potushil ogon'.

'In' 'bus' 'start' 'fire' 'driver' (~~'bus'.Gen~~) 'put out' 'fire'

'The fire began in a bus. The driver (~~of the bus~~) putted out the fire '

What Do We Annotate?

- Text?
  - Short news: 80 – 560 words, most texts are 150 -200 words
- Source?

  polit.ru – one of the biggest news resource in Russia.
- Topic?
  - Different topic, but there are more political news than others.
- When these news were written?
  - From 2013 to 2015

How Do We Annotate?

1. Automatic morphological annotation:

   FreeLing http://nlp.lsi.upc.edu/freeling/
   - Part of speech
   - Morphological categories

2. Manual annotation of bridging relations

   brat http://brat.nlplab.org for manual annotation and visualization

Automatic Bridging Anaphora

The goal of these corpus is a creation of automatic bridging anaphora resolution system.

So we don't have semantic annotation:

- There are no full enough and free semantic thesauruses or ontologies for Russian
- It's no much point to train our system on semantic relations – we can't use this knowledge in automatic bridging resolution

Manual Annotation Marks

Marks for relations:

1. Bridge = Genitive bridging
2. Coref = Coreference (necessary for bridging)
3. Coref-bridge
4. Non-gen = bridging close to genitive

Type of Links: bridge

Bridging in genitive construction:

V avtobuse nachalsya pojar. **Voditel'** (~~avtobusa~~) potushil ogon'.

'In' 'bus' 'start' 'fire' 'driver' (~~'bus'.Gen~~) 'put out' 'fire'

'The fire began in a bus. The driver (~~of the bus~~) putted out the fire '

Types of Links: coref; coref-bridge

Ob etom soobshil [advocat [uchastnikov$_{Gen}$ [akcii]]]$_i$ <…>
'lawyer' 'of the participants.Gen'

Po slovam[zashitnika]$_i$ (~~uchastnikov~~) akcii shestero britanzev
'defender' 'participants.Gen'

pokinuli Rossiju

'The lawyer of the participants of the action said<…> According to the defender six britishes left Russia'

- 'Defender' —coref-bridge→ 'the participants'
- 'Defender' —coref→ 'Lawyer of the participants'

Types of Bridge: non - gen

Moskovskij taxist otobral u pasagira (~~taxi~~ Gen) 1.5 milliona rublej

'taxi driver' passenger (~~of the taxi~~)

'Moscow taxi-driver took of 1.5 billions rubles from the passenger'

- 'Passenger' —bridge→ 'taxi'

But we haven't "taxi" in previous text

We have "Taxi driver" and it's one word in Russian (taxist)

→ 'Passenger' —non-gen→ 'taxi driver'

Manual Annotation Overview

It’s difficult to be concentrate on bridging pairs.

Annotators miss

- near to 40% of bridging cases in the first time.
- near to 15% after rechecking

It’s twice as much as in case of annotating direct anaphora or coreference

Manual Annotation Overview

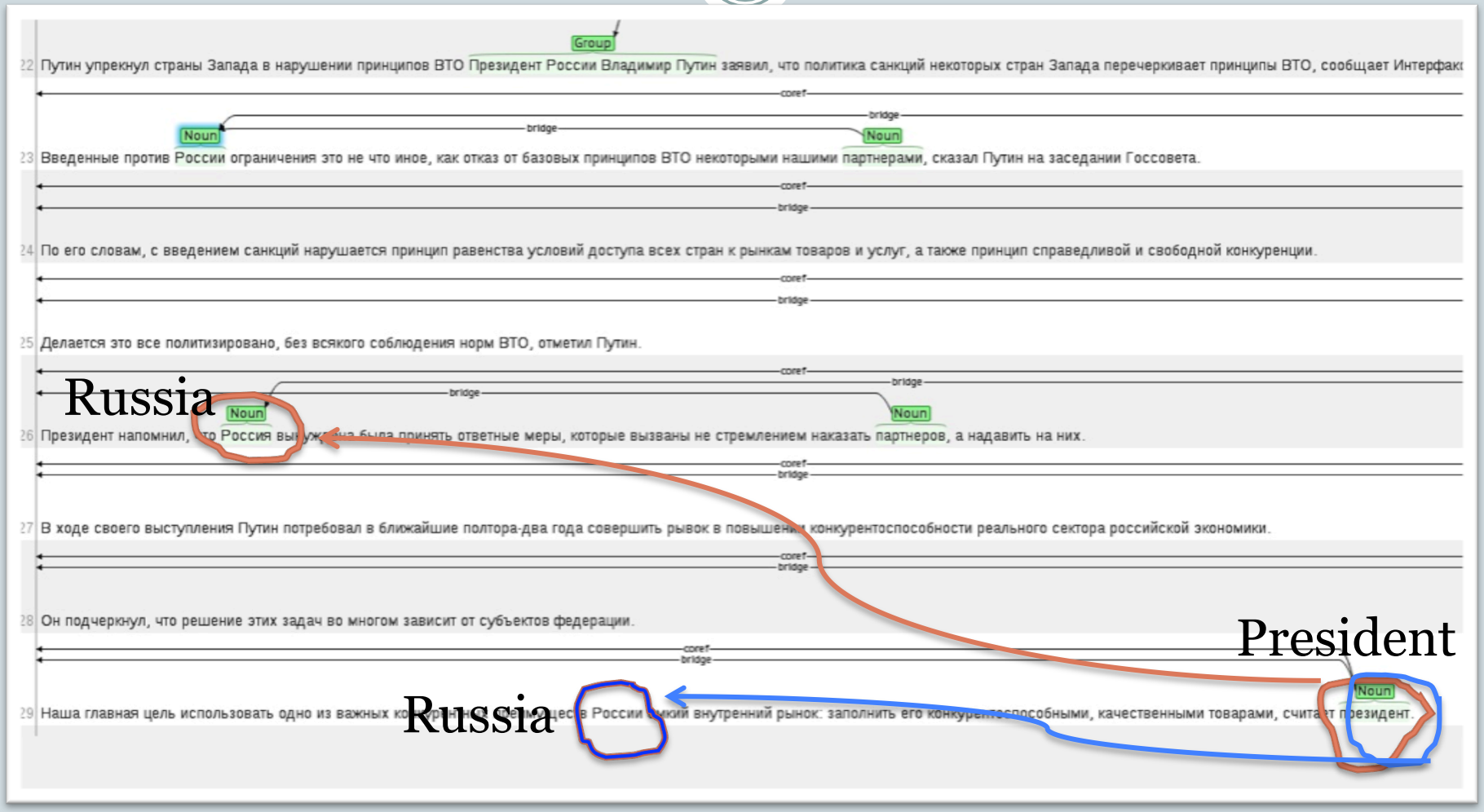
There are a very few situations where annotators are really disagree.

Similar bridging elements are usually linked to similar anchors with similar link.

Except where anchors are different but coreferent

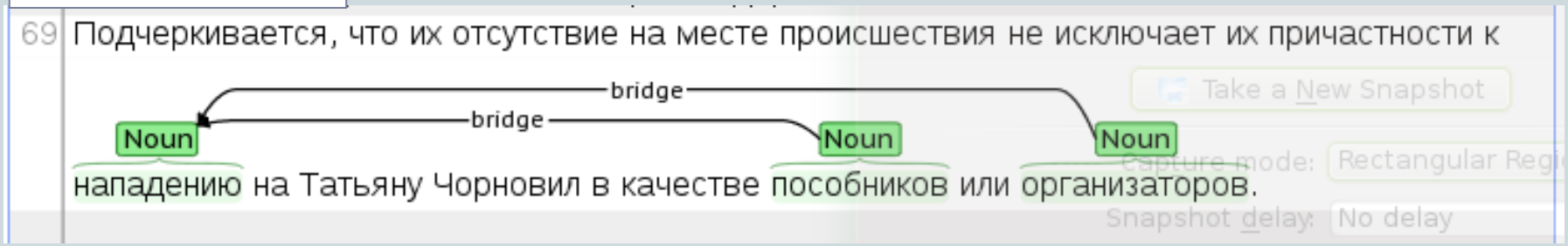
-> one of annotators had missed the closest anchor

Manual Annotation. Mistakes.



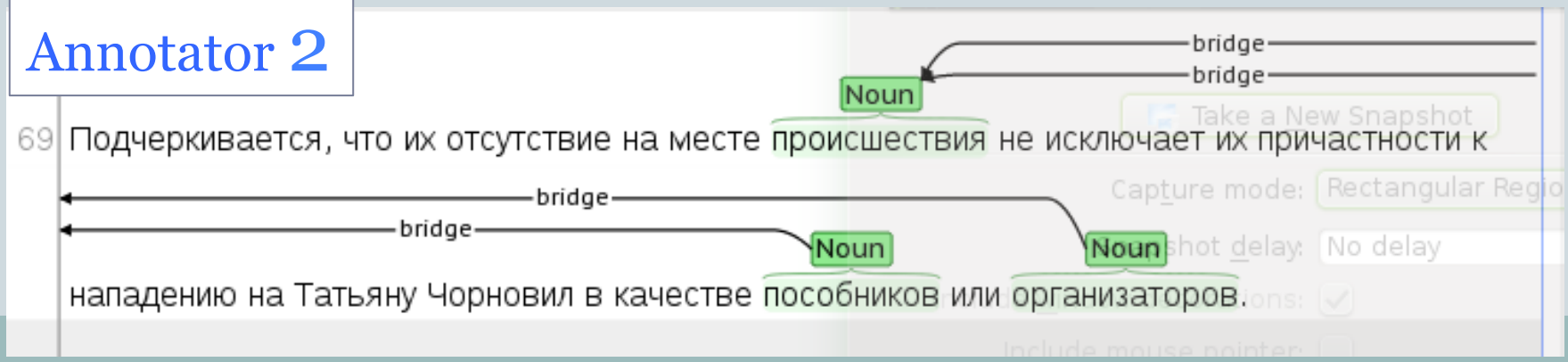
Manual Annotation. Mistakes.

Annotator 1



Importantly, their absence at the $\underline{\text{event}}_{i2}$ does not preclude their implication in $\underline{\text{the attack}}_{i1}$ on Tatyana Chernovil as either $\textit{accomplices}_{1;2}$ or $\textit{organizers}_{1;2}$.

Annotator 2



Typical Mistakes. “Local”

“Local”

Pravitelstvo Moskvy predlogilo <…>
‘government’ ‘Moscow’ ‘offered’

Mestnye giteli ~~(Moskvy.~~Gen) <…>
‘local’ ‘citizens’

“Local” is in anaphoric relation with actualized geographic object.

So “local citizens” is looks like “citizens of Moscow ”, and annotator forgets to draw an arrow.

Typical Mistakes. "National"

"National"

Bank Rossii provodit politiku podderganiya kursa nacional'noj valyuty (~~Rossii~~.Gen).

Bank of Russia maintain the policy of supporting the national currency (~~of Russia~~) rate.

"National" is in anaphoric relation with actualized geographic object.

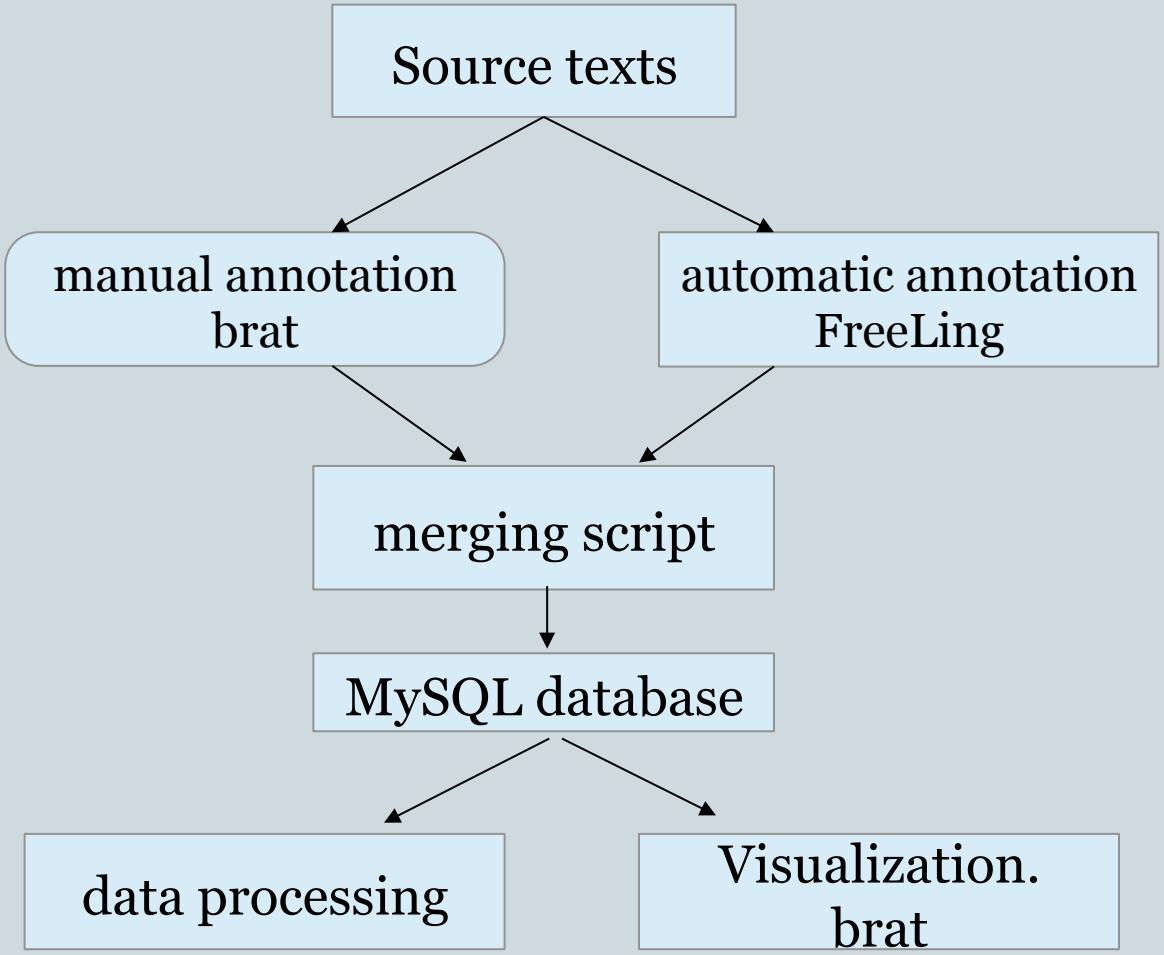
So "national currency" is looks like "Russian currency", and annotator forgets to draw an arrow.

Russian Bridging Corpus. Statistics

Texts	161
Bridging arows	249
Nests*	169
Max Nest-size	8
Min Nest-size	1
Max Arrow Length	516
Min Arrow Length	1

* Nest – is a set of bridging elements which refer to the same anchor

Russian Bridging Corpus. Technical Overview



Future Work on Corpus

- Annotate more texts.
  - We want to have 500 texts annotated with two annotators and checked by supervisor
- Add automatic syntactic annotation

The Last Slide

Спасибо!

¡Gracia!

Thank you!

1. Find Possible Bridging Elements

- Possible bridging elements?
- All the words with determinatives etc.

- But in Russian:
  - No articles
  - Usage of possessive pronouns and determinative elements is not obligatory (as e.g. in English)

4 Necessary Steps for Automatic Bridging Resolution

1. Find possible bridging elements
2. Choose the most probable bridging element
3. Find possible anchors
4. Choose the most probable anchor

Find Possible Bridging Elements

So we need more complicated rules to find potential bridging elements

-> the set of likely-bridging criterion

Corpus can give us all statistics on necessary criterion

Find Possible Anchors

-> the set of likely-anchor criterion

Corpus can give us all statistics on necessary criterion