FACTORS OF REFERENTIAL CHOICE: COMPUTATIONAL MODELING

N. Loukachevitch [2], G. Dobrov [3],
A. Kibrik [1], M. Khudyakova [4], A. Linnik [4]

1) Институт русского языка им. В.В. Виноградова РАН
2) НИВЦ МГУ им. М.В. Ломоносова
3) Факультет ВМиК МГУ им. М.В. Ломоносова
4) Филологический факультет МГУ им. М.В.Ломоносова

Грант РФФИ № 09-06-00390

Plan of presentation

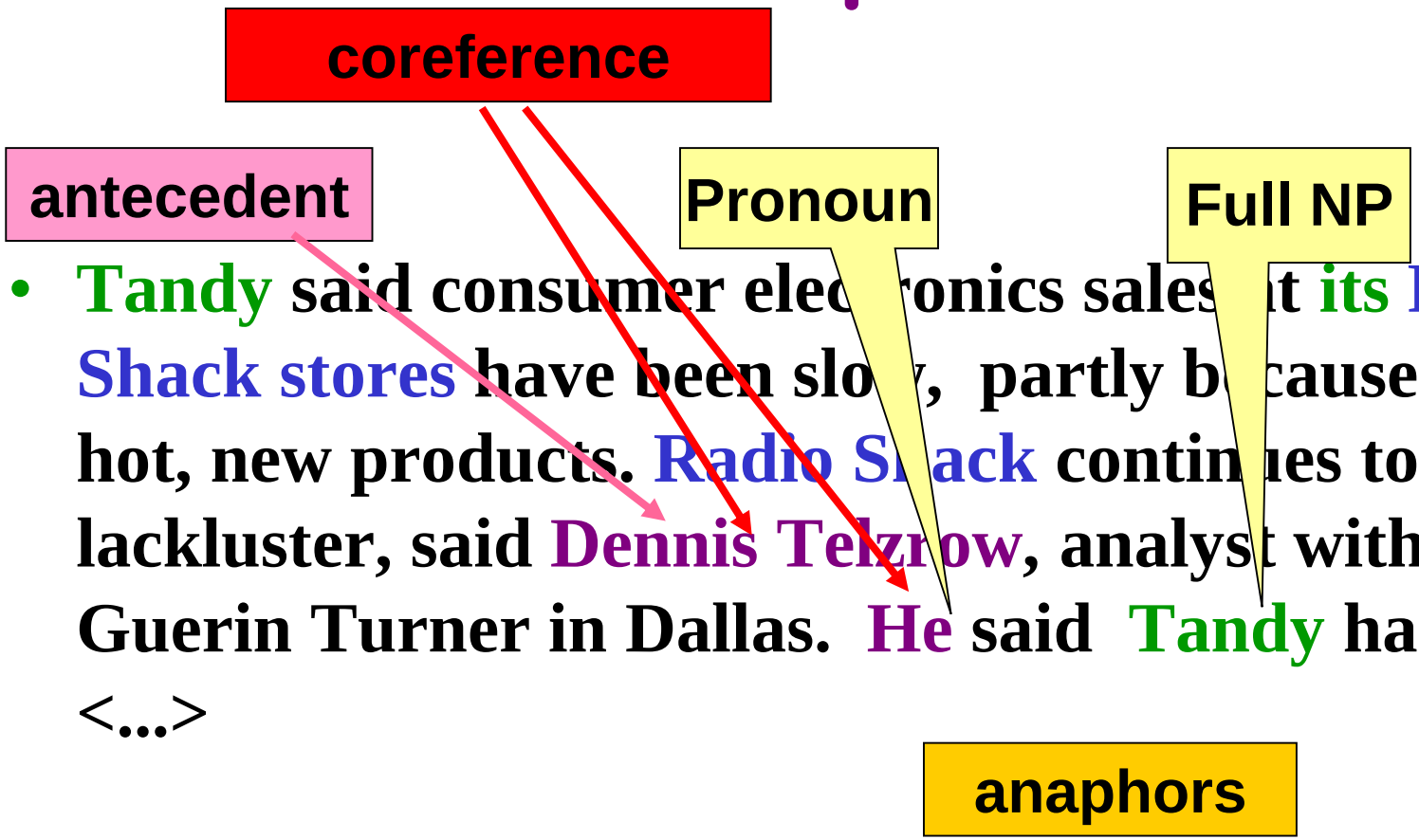
- **Preliminary work was presented at Dialog-2010**
- **This year**
  - **New factors**
  - **Machine learning: compositions of algorithms**

- **Plan**
  - **Referential choice**
  - **Refrhet corpus**
  - **Factor of referential choice**
  - **Machine learning algorighms**
  - **Results**

Referential choice in discourse

- **When a speaker needs to mention (or refer to) a specific, definite referent, s/he chooses between several options, including:**
  - **Full noun phrase (NP)**
    - **Proper name (e.g. Pushkin)**
    - **Common noun (with modifiers) = definite description (e.g. the poet)**
  - **Reduced NP, particularly a third person pronoun (e.g. he)**

Example



- **Tandy said consumer electronics sales at its Radio Shack stores have been slow, partly because a lack of hot, new products. Radio Shack continues to be lackluster, said Dennis Telzrow, analyst with Eppler, Guerin Turner in Dallas. He said Tandy has done <...>**

➢ How is this choice made?

Why is this important?

- **Reference is among the most basic cognitive operations performed by language users**
- **It is the linguistic representation of what is known as attention and working memory in psychology**
- **Reference constitutes a lion's share of all information in natural communication**
- **Reference is an important property of coherent text -> natural language processing**

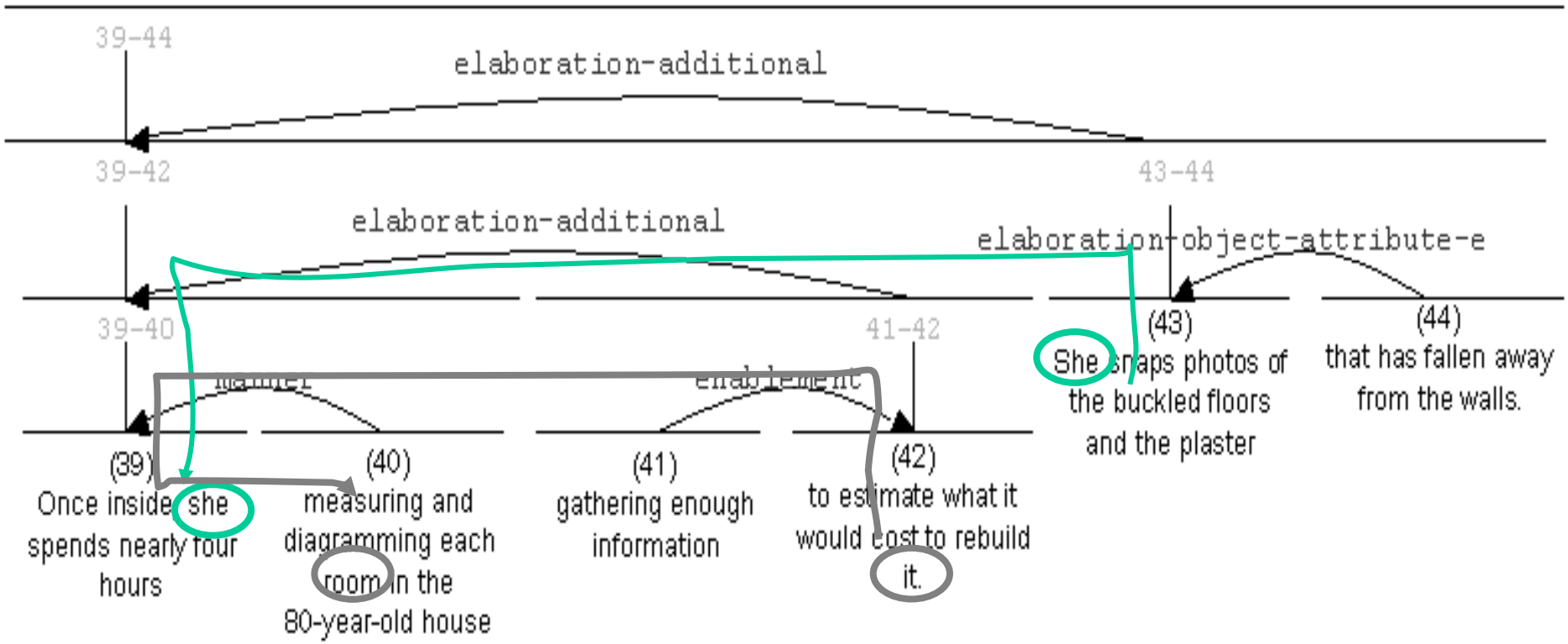
The RefRhet corpus

- **English**
- **Business prose**
- **Initial material – the RST Discourse Treebank**
  - **Annotated for hierarchical discourse structure**
  - **385 articles from Wall Street Journal**
- **The added component – referential annotation**

➢ **The RefRhet corpus**

- **Over 30 000 referential expressions**
- **157 texts are annotated twice**
- **193 texts are annotated once**

Example of a hierarchical graph



Scheme of referential annotation

- **The MMAX2 program**
- **Krasavina and Chiarcos 2007**
- **All markables are annotated, including:**
  - **Referential expressions**
  - **Their antecedents**
- **Coreference relations are annotated**
- **Features of referents and context are annotated that can potentially be factors of referential choice**

Current state of the RefRhet referential annotation

- **2/3 completed**
- **Further results are based on the following data:**
  - **247 texts**
  - **110 thousand words**
  - **26 024 markables**
    - **7097 proper names**
    - **8560 definite descriptions**
    - **1797 third person pronouns**
  - **4291 reliable pairs «anaphor – antecedent»**
    - **Full noun phrases — 2854**
    - **Pronouns — 1437**

Factors of referential choice: referent

- **Animacy: animate (human) or inanimate (non-human)**
- **Gender and number (agreement): masculine, feminine, neuter, plural**
- **Protagonism, that is a referent's centrality in discourse**
  - **the ratio of its referential chain length to the maximal length of a referential chain,**
  - **ratio of its referential chain to the gross number of referential expressions in the text.**

Experiment: Protagonism Modeling

- **Participants (English native speakers) were required to read the text and to identify the central entity**
- **30 texts**
- **15 texts: participants were unanimous**
- **11 texts: agreement between two (out of three) participants**
- **26 texts: reliable information about a protagonist**
- **human assessment and the computer's assessment coincide in 24 instances – 92%**

Factors of referential choice: antecendent

- **Affiliation in direct speech (dir_speech); particularly important are the situations in which anaphor and atecedent are located across a direct speech boundary**
- **Type of phrase (phrase_type):**
  noun phrase, prepositional phrase
- **Grammatical role (gramm_role):**
  subject, direct object, indirect object
- **Referential form (np_form, def_np_form):**
  definite NP, with further indication of subtype, vs. proper name vs. indefinite NPs
- ***Antecedent length, in words***
- ***Number of markables from the anaphor back to the nearest full NP antecedent***

Factors of referential choice: anaphor

- **Introductory vs. repeated mention (referentiality)**
- ***Number of referent mention in the referential chain***
- **Affiliation in direct speech (dir_speech)**
- **Type of phrase (phrase_type): noun phrase, prepositional phrase**
- **Grammatical role (gramm_role): subject, direct object, indirect object**

Factors of referential choice: distances

- **Distance in words**
- **Distance in markables - referential competition in a discourse context - referential conflict**
- **Linear distance in elementary discourse units, as found in the rhetorical representation**
- **Rhetorical distance in elementary discourse units, as found in the rhetorical representation**
- ***Distance in sentences***
- ***Distance in paragraphs*.**

Factors 2011

- **Gender and number (agreement): masculine, feminine, neuter, plural**
- **Antecedent length, in words**
- **Number of markables from the anaphor back to the nearest full NP antecedent**
- **Number of referent mention in the referential chain**
- **Distance in sentences**
- **Distance in paragraphs**

Goals for machine learning study

- **Dependent variable:**
  - **Form of referential expression (np_form)**
- **Binary prediction:**
  - **Full NP vs. pronoun**
- **Three-way prediction:**
  - **Definite description vs. proper name vs. pronoun**
- **Accuracy maximization:**
  - **Ratio of correct predictions to the overall number of instances**

Machine learning methods (Weka, a data mining system)

- **Easily interpretable methods:**
  - **Logical algorithms**
    - **Decision trees (C4.5)**
    - **Decision rules (JRip)**
- **Higher quality:**
  - **Logistic regression**
- **Quality control – the cross-validation method**
- **This year – composition of classifiers**
  - **Bagging**
  - **Boosting**

Examples of decision rules generated by the JRip algorithm

- (Antecedent's grammatical role = subject) &
  (Hierarchical distance ≤ 1.5) &
  (Distance in words ≤ 7)
  => pronoun

- (Animate) &
  (Distance in markables ≥ 2) &
  (Distance in words ≤ 11)
  => pronoun

Composition of classifiers: boosting algorithm

- **Base algorithm (C4.5 Decision trees)**
- **Iterative process**
- **Each additional classifier applies to the objects that were not properly classified by the already constructed composition**
- **At each iteration the weights of each wrongly classified object increase, so that the new classifier focuses on such objects**

Composition of classifiers: bagging

- **Base algorithm (C4.5 Decision trees)**
- **Bagging randomly selects a subset of the training samples to train the base algorithm**
- **Set of algorithms built on different, potentially intersecting, training subsamples**
- **A decision on classification is done through a voting procedure in which all the constructed classifiers take part.**

Two-way task: full noun phrase vs. pronoun.

Algorithm	Accuracy 2010	Accuracy 2011
Logistic regression	85.6%	87.0%
Decision tree algorithm	84.3%	86.3%
Deciding rules algorithm	84.5%	86.2%
Boosting	88.2%	**89.9%**
Bagging	86.6%	87.6%

Three-way task: descriptive np, proper names, pronouns

Algorithm	Accuracy 2010	Accuracy 2011
Logistic regression	76.0%	77.4%
Decision tree algorithm	74.3%	76.7%
Deciding rules algorithm	72.5%	75.4%
Boosting	79.3%	**80.7% (50 it)** **80.9% (100 it)**
Bagging	78.0%	79.5% (50 it) 79.6% (100 it)

Significance of factors in the three-way task of referential choice-1

Factors	Accuracy
All factors, including the newly added ones	80.7%
without protagonism	80.0%
without animacy	80.68%
without the anaphor's grammatical role	79.3%
without the antecedent's grammatical role	80.2%
without grammatical role	79.2%
without the antecedent's referential form	77.0%

Significance of factors in the three-way task of referential choice - 2

Factors – distances (6)	Accuracy
All factors, including the newly added ones	80.7%
without all distances	73.5%
- except for rhetorical distance only	74.9%
- except for the distance in words only	79.0%
- except for the distances in words and paragraphs	79.0%
- except for the distances in words and sentences	79.5%
- except for rhetorical distance and the distances in words and sentences	79.7%
- except for the distances in words, markables, and paragraphs	<u>80.47%</u>

Conclusion

- **We have presented the recent results of our modeling study in referential choice, based on the RefRhet corpus**
- **The account of additional factors and the employment of compositions of machine learning techniques have led to an improvement of referential choice prediction**
- **The majority of factors taken into account are significant for modeling referential choice**