Identifying Multidocument Relations

Erick G. Maziero
Maria Lucía R. Castro Jorge
Thiago A. S. Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

FAPESP & CNPq

Multidocument scenario

- Huge amount of information
  - IDC: 800 exabytes of new information only in 2009
- Several information sources with a variety of multidocument phenomena
  - Redundant, complementary and contradictory information
  - Events that evolve in time
  - Different perspectives and positions
  - Diverse writing styles

Multidocument processing

- Google and GoogleNews are not enough

- Text Summarization
  - For example, *NewsBlaster* (McKeown et al., 2001) and *MEAD* (Radev et al., 2000)

- Question answering
  - For example, *Wolfram Alpha* and *Ask.com*

Multidocument processing

- Room for a lot of improvements in the available systems
  - Appropriately dealing with multidocument phenomena

- Possible solution
  - Better understanding and representation of the multidocument phenomena
    - How text parts relate to one another
  - Multidocument parsing

Multidocument parsing

- Questions to answer
  - Which multidocument phenomena happen in news texts?
    - Which ones are more frequent?
  - Are we able to grasp them?
    - How good we are?
  - Is it possible to automate this task?

This work

- **Our experience**
  - Method, tools and results for corpus annotation
  - Experiment on automatic multidocument parsing

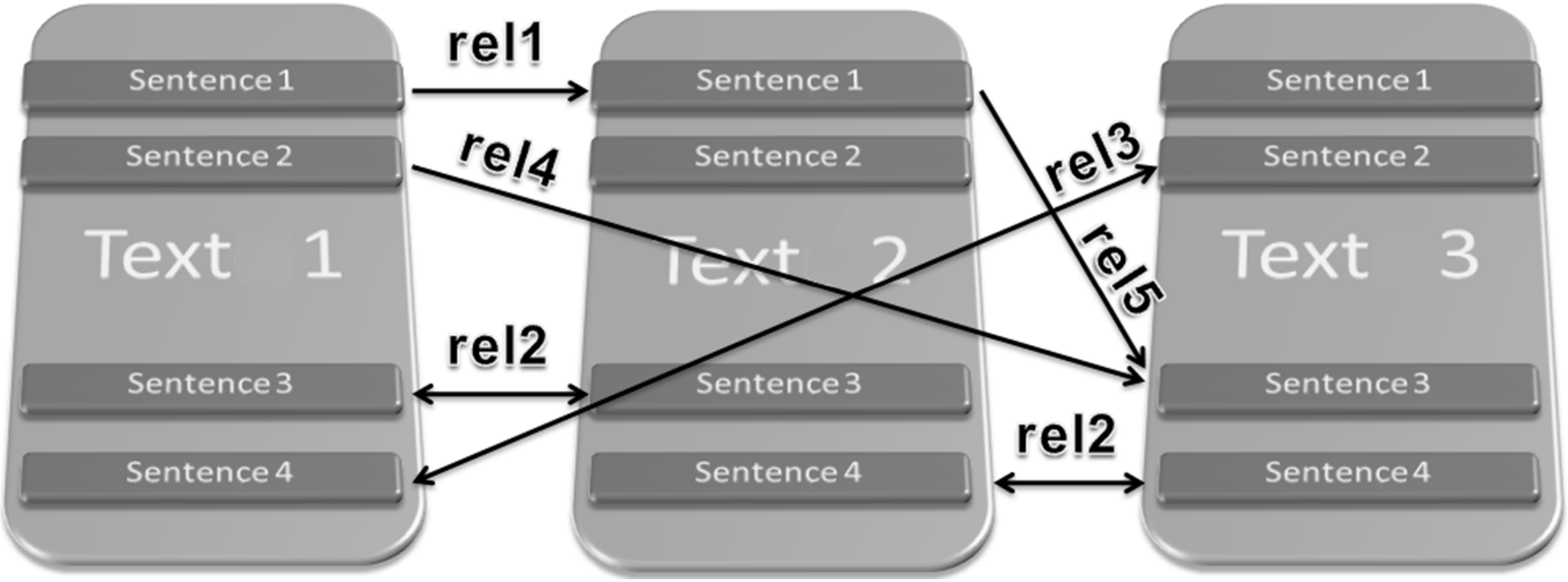
- Language: *Brazilian Portuguese*

Previous work

- Trigg et al. (1983, 1986) and the TextNet system for scientific papers
- RST (Rhetorical Structure Theory) (Mann and Thompson, 1987): single document relations
- Radev and Mckeown (1995): SUMMONS and its operators
- Allan (1996): typology of links for relating documents
- Radev et al. (2000, 2001, 2002): CST (Cross-document Structure Theory) and initiatives of automatic parsing
- Afantenos et al. (2004, 2007): problems with CST and new proposal

CST (Radev, 2000)

- Model of multidocument relationship for related texts
  - Any level of analysis is possible



CST (Radev, 2000)

- Originally 24 relations

Identity	Modality	Judgment
Equivalence	Attribution	Fulfillment
Translation	Summary	Description
Subsumption	Follow-up	Reader profile
Contradiction	Elaboration	Contrast
Historical background	Indirect speech	Parallel
Cross-reference	Refinement	Generalization
Citation	Agreement	Change of perspective

CST refinement (Zhang et al., 2003)

- 18 relations

Identity	Modality	Change of perspective
Equivalence	Attribution	Fulfillment
Translation	Summary	Description
Subsumption	Follow-up	Reader profile
Contradiction	Elaboration	Citation
Historical background	Indirect speech	Generalization

CST: example

- Contradiction, overlap, historical background (←)

D1: A plane crash in the town of Bukavu in Congo killed 13 people on Thursday afternoon, said on Friday a spokesman from the United Nations.

D2: At least 17 people died with the crash of a plane in Congo. According to a spokesman from the UN, the plane was trying to land in the airport of Bukavu during a storm. Congo has a history of more than 30 aircraft accidents.

CST parsing

- CSTBank (Radev et al., 2004): unique corpus for English
  - Clusters of related news texts
  - For a sample of 88 segment pairs
    - 58% of total or partial annotation agreement
    - No kappa values reported
  - Some relations are difficult to understand (Afantenos et al., 2004)

CST parsing

- Zhang et al., 2003, 2004: only known attempt for English
  - 2 steps
    - Determining which segments may present relations
    - Finding the relations
  - Machine learning
    - Simple features: number of words, POS tags, semantic similarity of words (using Wordnet), etc.
    - Subset of relations: equivalence, subsumption, follow-up, elaboration and overlap
    - Best results: 0.29 average f-measure

Our experiments: corpus

- **CSTNews** (Aleixo and Pardo, 2008)
  - 50 clusters of related news texts from several on-line sources
    - Each cluster has 2-4 texts
    - Each text has ½-1 page

Our experiments: annotation tool

- **CSTTool** (Aleixo and Pardo, 2008)
  - Automatic sentence segmentation
  - Suggestion of segment pairs to relate
    - Also based on Zhang and Radev (2004), word overlap measure
      - Otherwise, too many segment pairs to consider
      - Zhang et al. (2003): *CST relations are unlikely to exist between segments that are lexically very dissimilar to each other*
  - XML output in CSTBank format

Our experiments: 1st annotation

- The problem was harder than we thought
  - 2 computational linguistics with some study and training in CST
  - Very low agreement: 0.26 in the traditional kappa measure
    - Very naïve approach!
      - Not enough training
      - No suggestions from CSTTool

Our experiments: 2nd annotation

- Consistent training step with 4 computational linguists
  - 1-2 months

- CST refinement
  - Refined relation set
  - Better relations definitions
  - Relations typology
  - Constraints

Our experiments: 2nd annotation

- **New relation set**: 14 relations
  - Some confusing relations were joined
  - Some relations that were never observed were not considered

Example of definition

<u>Relation name</u>: subsumption

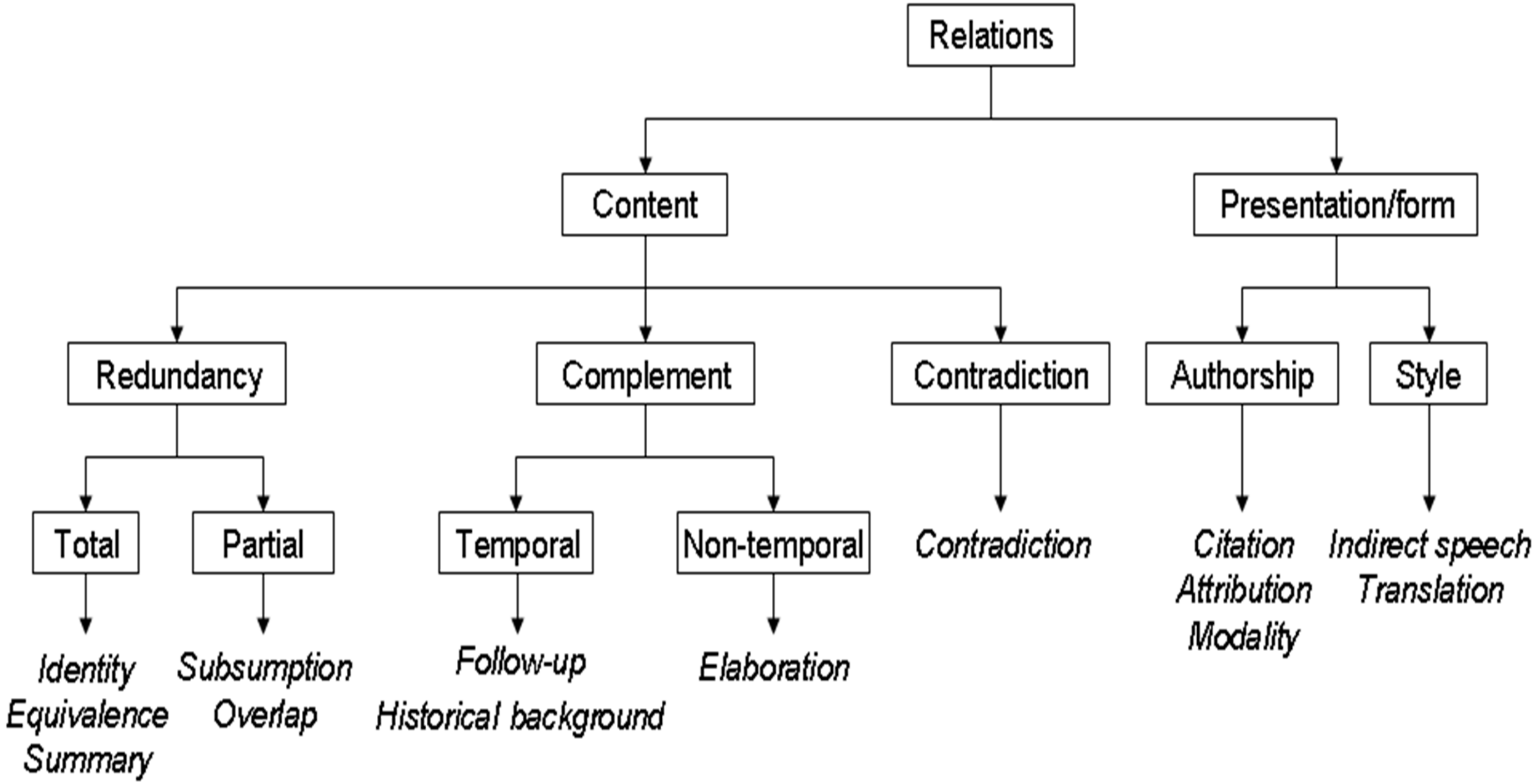
<u>Directionality</u>: S1→S2

<u>Restrictions</u>: S1 presents the information of S2 and as well as additional information

<u>Comments</u>: S1 presents contents X and Y, S2 presents only X

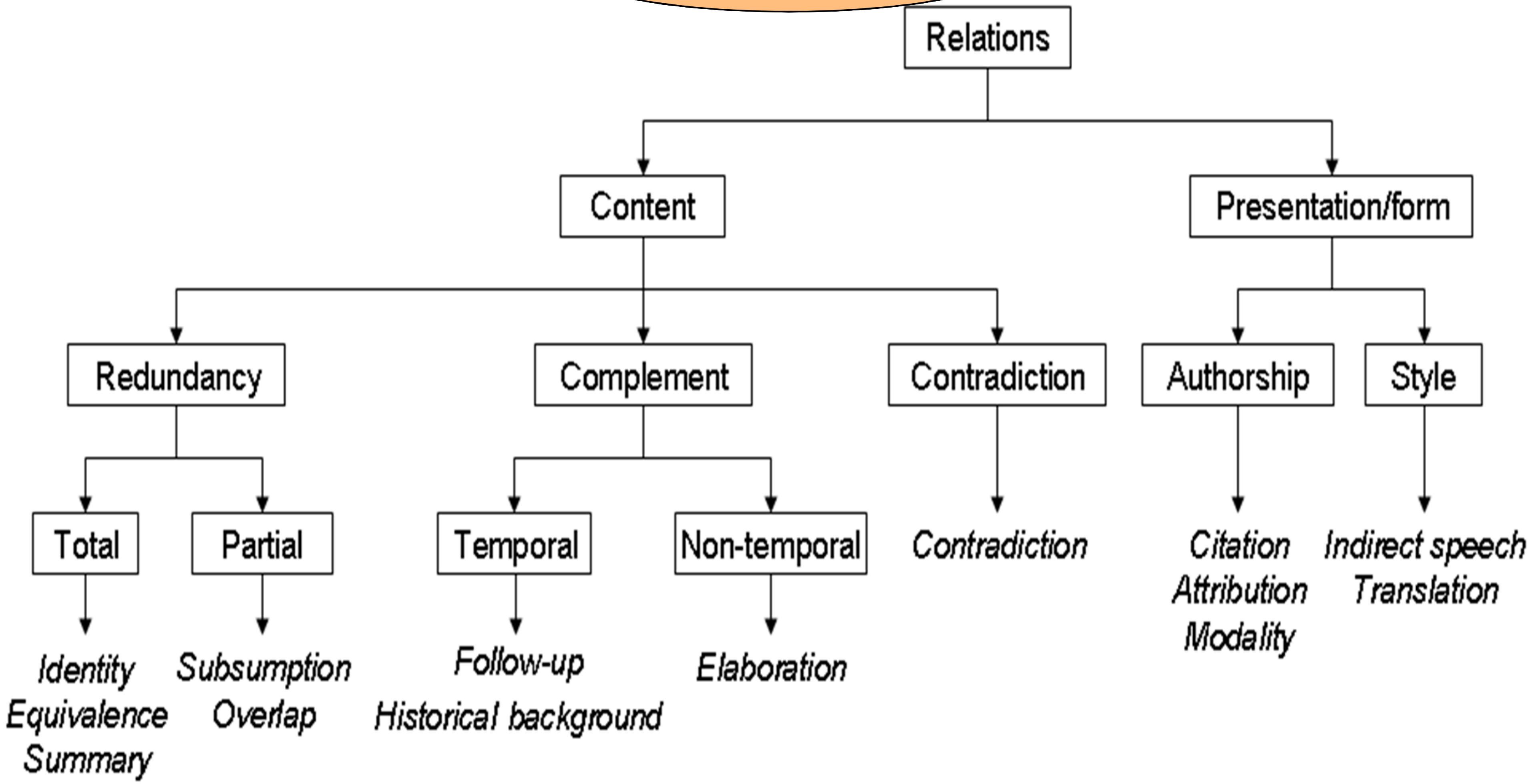
Our experiments: 2nd annotation

- **Typology of relations**



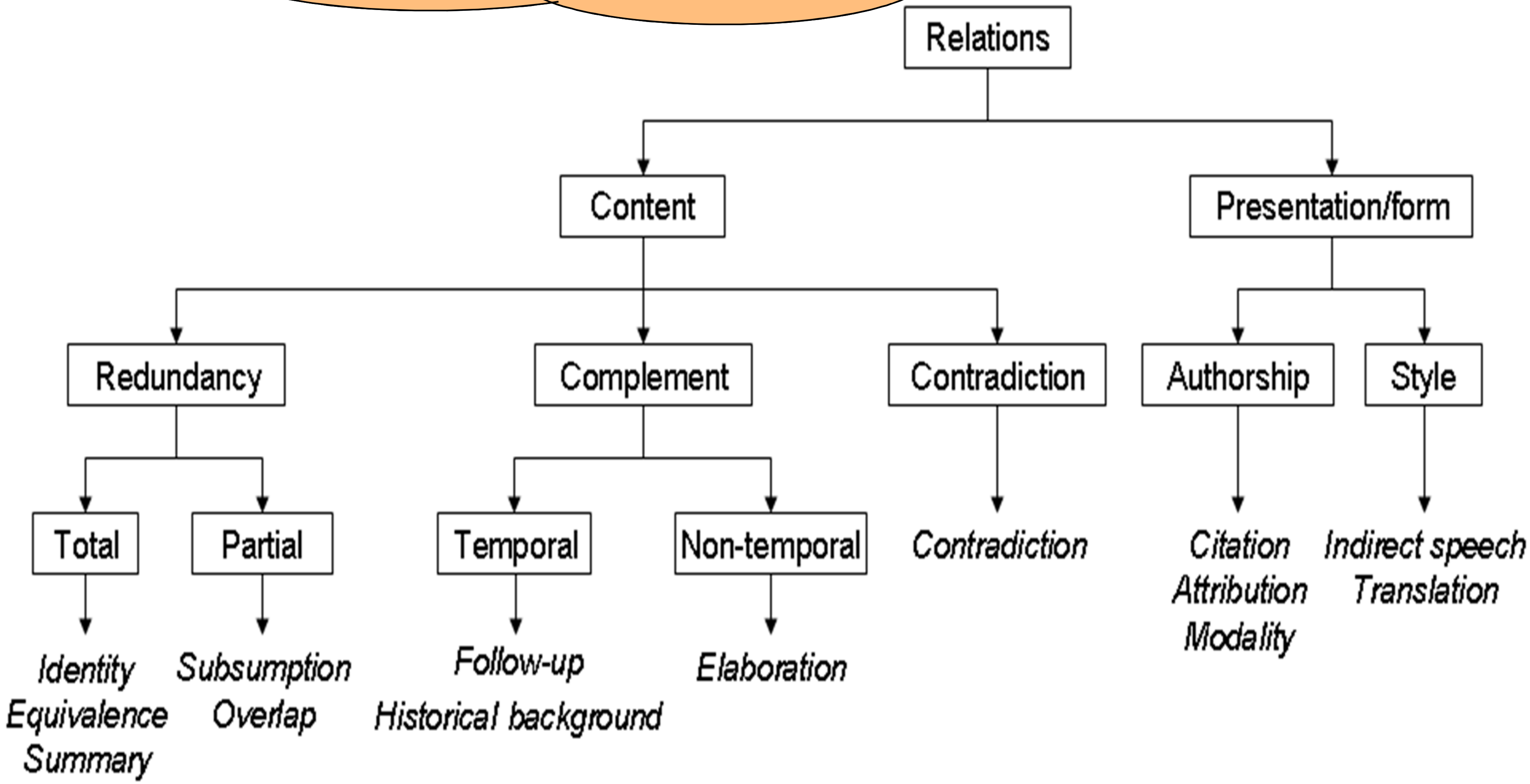
Our experiments: 2nd annotation

It is not possible that 2 content relations happen for the same information piece



Our experiments: 2nd annotation

Presentation/form relations usually happen with some content relation

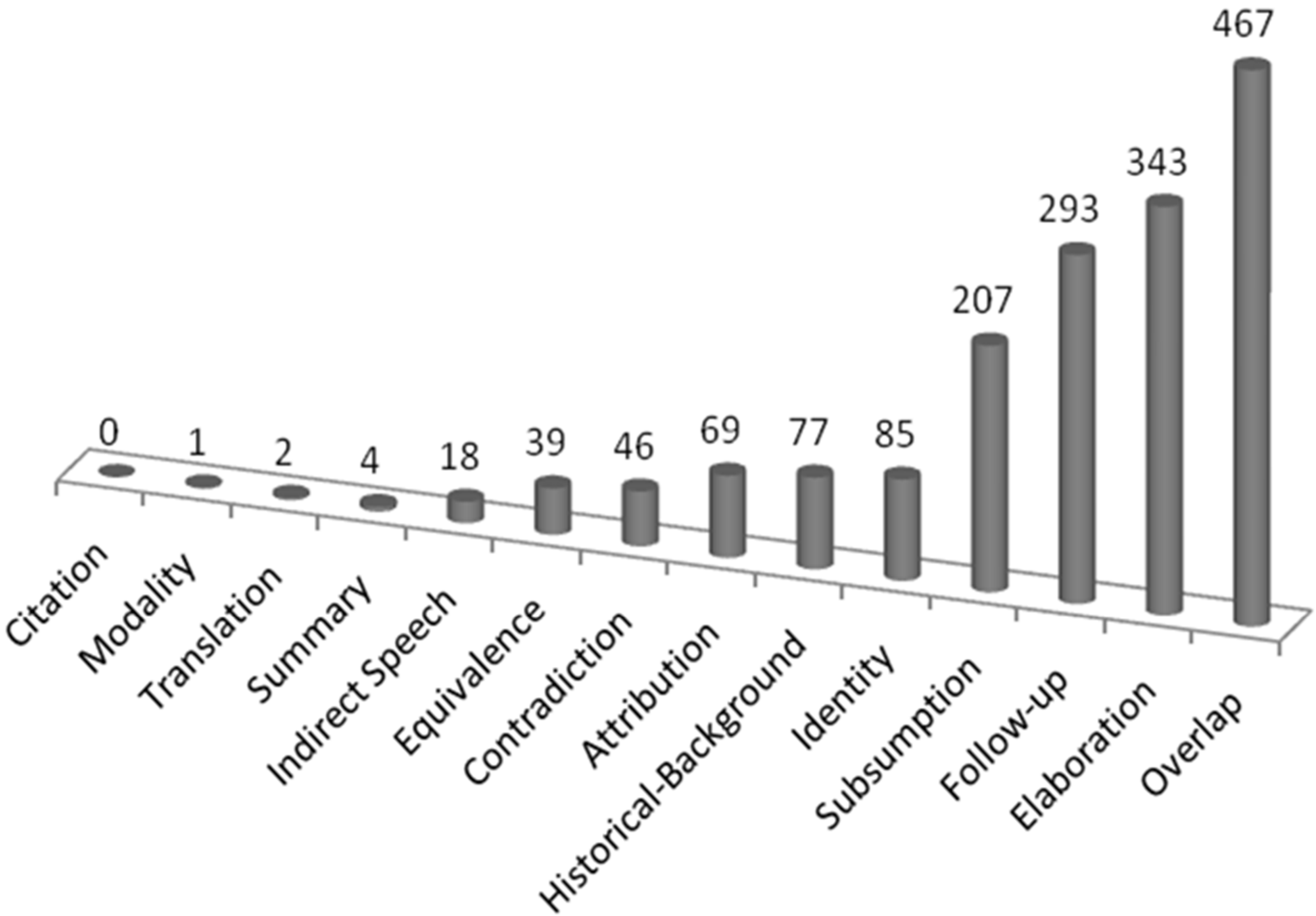


Our experiments: 2nd annotation

- Annotation step with 4 computational linguistics
  - 1-hour daily sections during 3-4 months

- Kappa periodically measured

Our experiments: 2nd annotation



Our experiments: 2nd annotation

- Annotation agreement

	Kappa	*Percentage agreement*		
		Full	Partial	Null
Relations	0.51	0.54	0.27	0.18
Directionality	0.45	0.58	0.27	0.14
Relations categories	0.61	0.70	0.21	0.09

Our experiments: 2nd annotation

- Annotation agreement

	Kappa	*Percentage agreement*		
		Full	Partial	Null
Relations	0.51	0.54	0.27	0.18
Directionality	0.45	0.58	0.27	0.14
Relations categories	0.61	0.70	0.21	0.09

80% of full or partial agreement vs. 58% for English

kappa 96% better than the original annotation for Portuguese

Our experiments: parsing

- Problem modeled as a machine learning task
  - Learning instance: segment pair codified as a set of features
    - Simple features
  - Classes: CST relations

Our experiments: parsing

- Features
  - Difference of segments size
  - Number of common words in the segments
  - Same segments?
  - Position of segments in their texts
  - Number of nouns in the segments
  - Number of verbs in the segments
  - Number of adjectives in the segments
  - Number of adverbs in the segments
  - Number of numerals in the segments

Our experiments: parsing

- WEKA (Witten and Frank, 2005)
  - J48, naïve-bayes, SVM
    - 10-fold cross-validation

- Data: only <u>content relations</u> from CSTNews
  - 1.561 instances
    - Unbalanced data: SMOTE (Chawla et al., 2002)

      - Using the presentation/form relations would generate a multi-label classification problem

Our experiments: parsing

- Results: 0.44 average F-Measure
  - Versus 0.29 for English

Confusion matrix

	A	B	C	D	E	F	G	H	I
Subsumption (A)	105	20	49	15	4	7	1	6	0
Elaboration (B)	27	119	115	56	17	5	0	3	1
Overlap (C)	52	96	204	81	7	14	1	11	1
Follow-up (D)	25	56	83	95	7	22	1	4	0
Historical B. (E)	9	22	12	10	91	7	0	3	0

Our experiments: parsing

- Portuguese vs. English

	English	Portuguese
Subsumption	0.05	0.47
Overlap	0.43	0.42
Equivalence	0.34	0.48
Elaboration	0.24	0.35
Follow-up	0.39	0.33

- Differences in results
  - Better corpus for Portuguese, slightly different versions of CST, language differences

Multidocument parsing

- Questions to answer
  - Which multidocument phenomena happen in news texts? Which ones are more frequent?
    - Ok
  - Are we able to grasp them? How good we are?
    - Ok, not perfect, but hard to be better
  - Is it possible to automate this task?
    - Possibly yes, with a more knowledge-based approach

Multidocument parsing

- Questions to answer
  - Which multidocument phenomena happen in news texts? Which ones are more frequent?
    - Ok
  - Are we able to grasp them? How good we are?
    - Ok, not perfect, but hard to be better
  - Is it possible to automate this task?
    - Possibly yes, with a more knowledge-based approach

Future work

Thank you

www.nilc.icmc.usp.br
taspardo@icmc.usp.br