Discourse Parsing

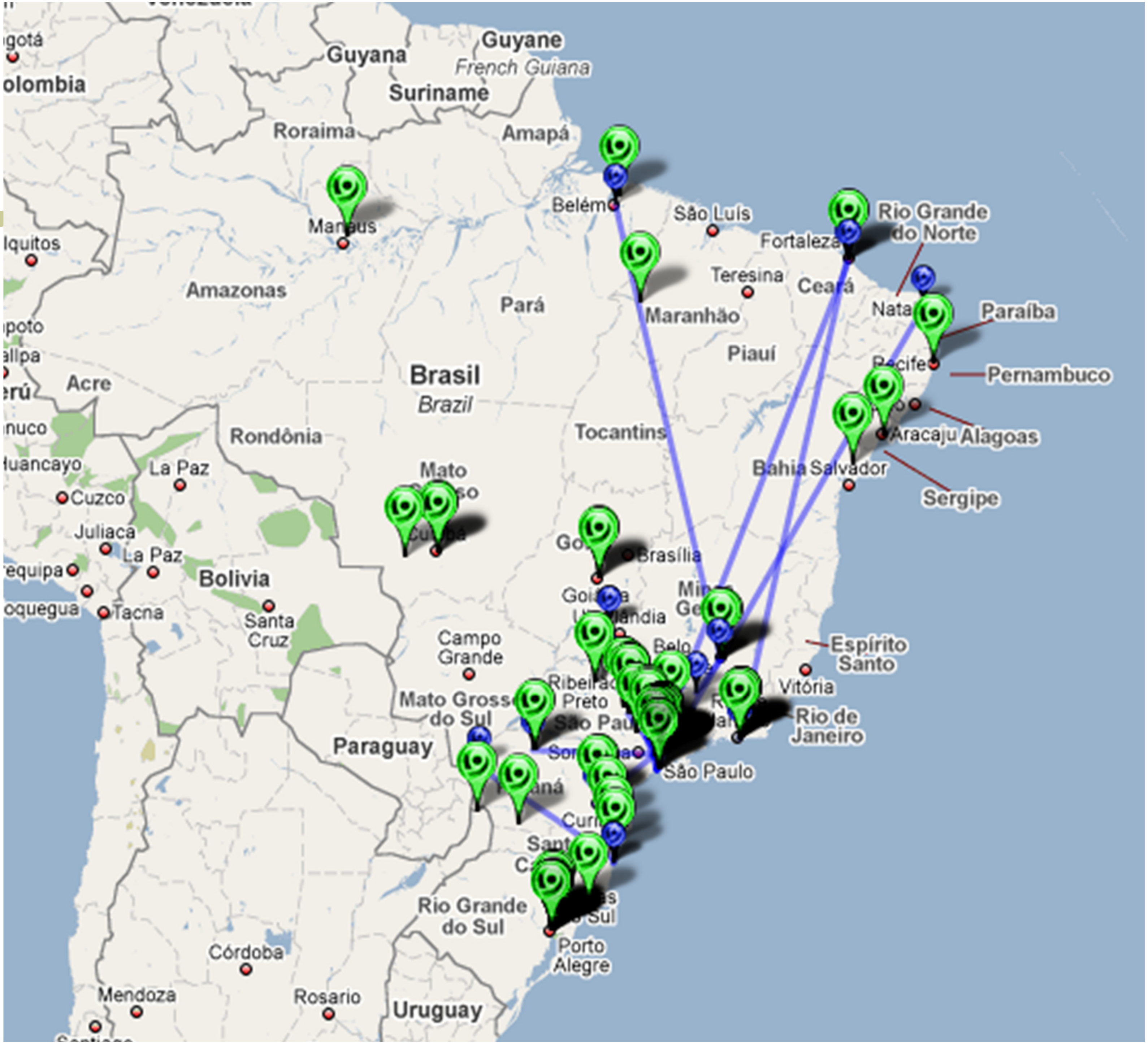
Thiago A. S. Pardo

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

NILC

- Biggest NLP group in Brazil
- Since 1993
  - Grammar checking (MS), writing support tools, machine translation and summarization
- Today: some big funded projects
  - Text simplification
  - Computational terminology
  - **Multidocument summarization**

NLP in Brazil



NILC

- Works for **Brazilian Portuguese** (mainly), Spanish and English
  - Resources: several corpora (the biggest one for Brazilian Portuguese), wordnets, grammars, etc.
  - Tools: POS tagger, syntactical parsers, **discourse parser**, NER, text alignment, etc.
  - Applications: machine translation, **summarization**, simplification, writing support tools, etc.

NILC

- 14 professors from 3 main universities
  - Computer scientists, linguists, and one physicist

- More than 50 students
  - Undergraduate, MSc, PhD, and pos-doc

Outline

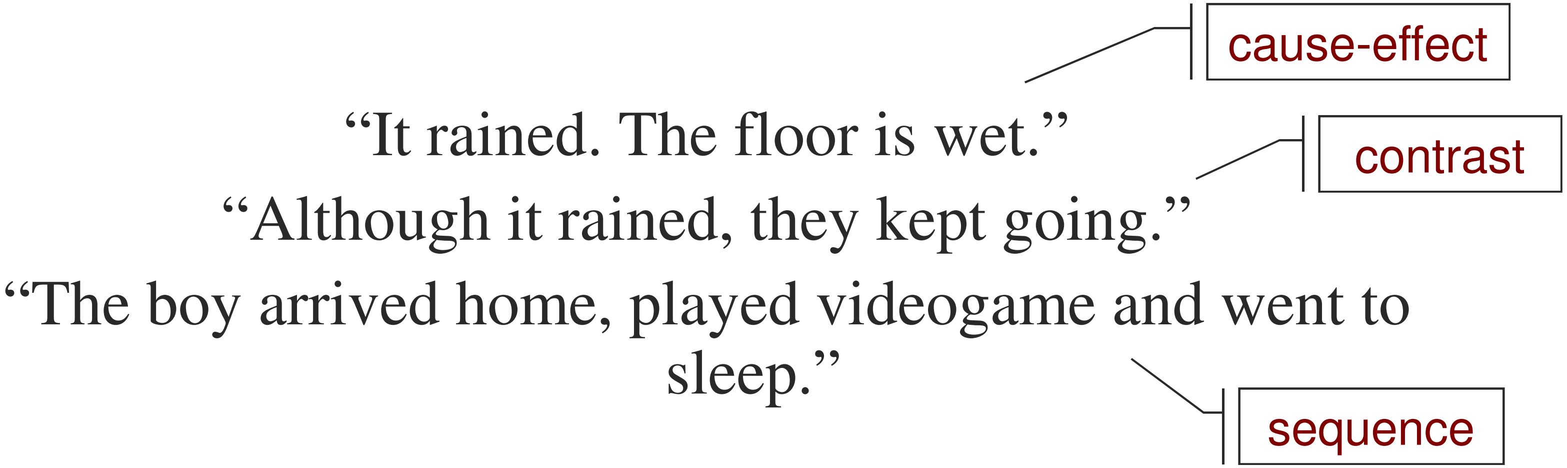
- Single document discourse parser
- Multidocument discourse parser
- Summarization experiences
  - Single and multidocument

Introduction

- Discourse analysis (Marcu, 2000)
  - Uncover the discourse structure of texts, i.e., how propositions of a text are related
    - Propositions: content units of a text, its smallest meaningful 'parts'
    - In general, propositions are expressed by simple clauses

Introduction

- A coherent text have a complex underlying discourse structure



cause-effect

"It rained. The floor is wet."

contrast

"Although it rained, they kept going."

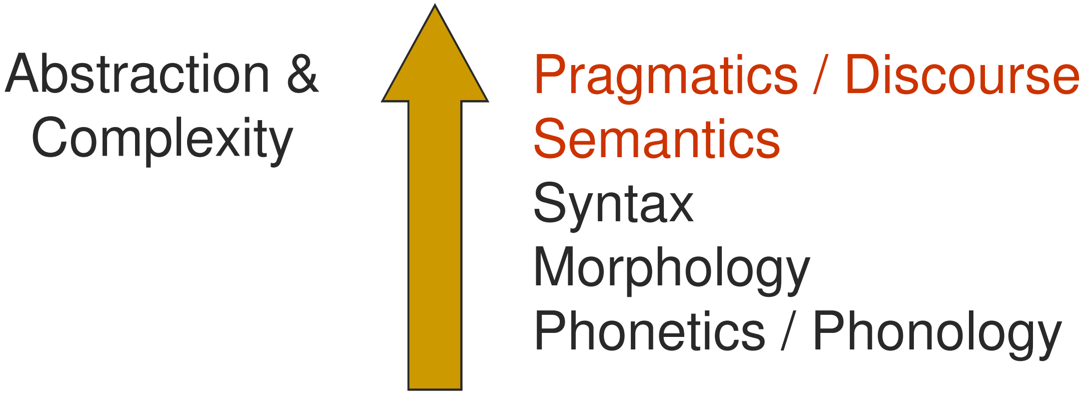
"The boy arrived home, played videogame and went to sleep."

sequence

- Relational analysis (Moore and Pollack, 1992; Moser and Moore, 1996)

Discourse

- Knowledge levels in NLP (Jurafsky and Martin, 2000)



Abstraction & Complexity

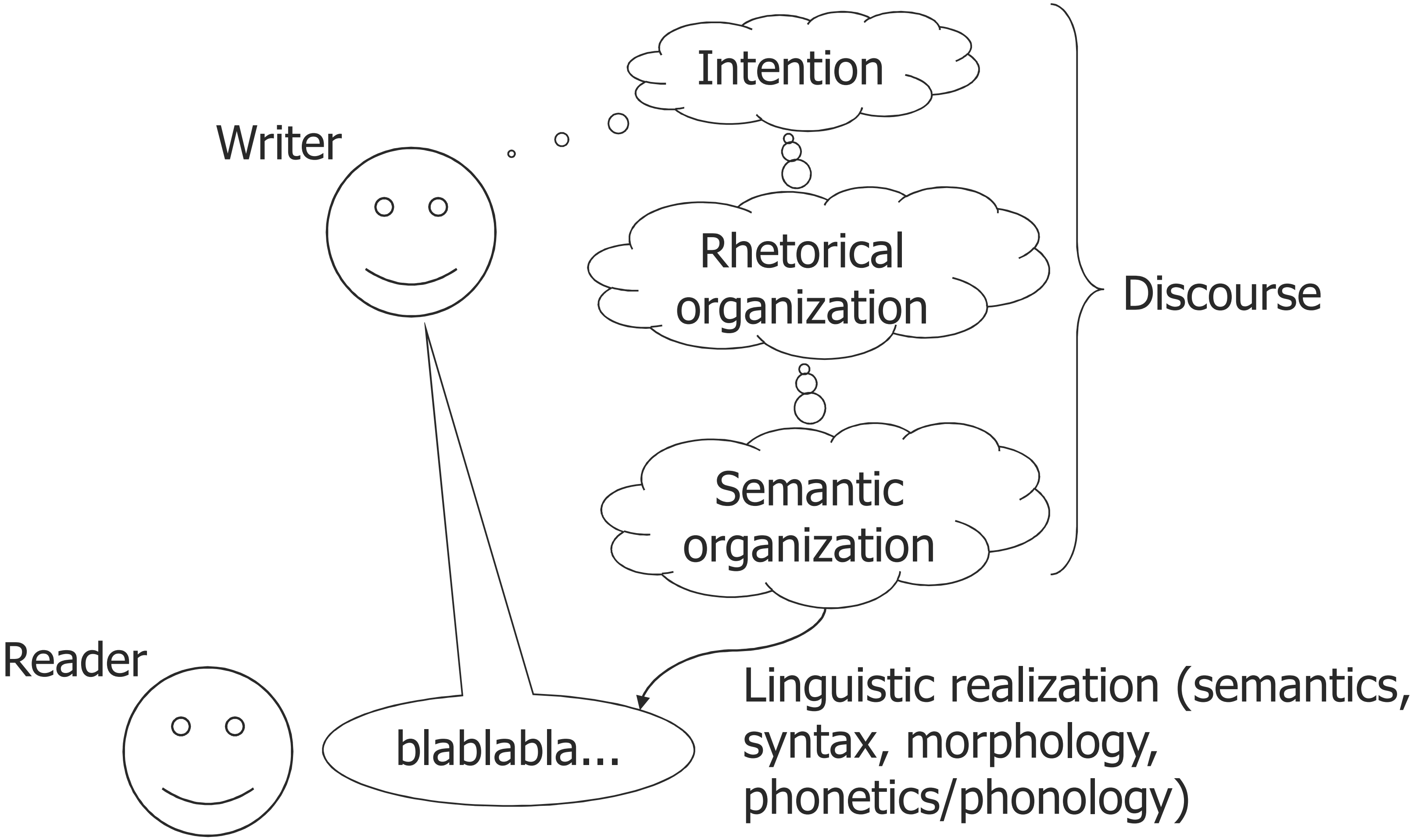
Pragmatics / Discourse
Semantics
Syntax
Morphology
Phonetics / Phonology

- Communicative situation (Koch and Travaglia, 2002): writer and reader

DiZer – DIscourse analyZER

- First automatic discourse analyzer for Brazilian Portuguese
  - Rhetoric
    - The way a text is organized in order to achieve its objective
    - Functional organization of the text (Mann and Thompson, 1987)
    - “Touchable” part of pragmatics (Hovy, 1988)

Rhetoric and the functional language



Discourse theories

- Grosz and Sidner (1986): intentions
- Mann and Thompson (1987): rhetoric
- Jordan (1992) and Kehler (2002): semantics
- Moore and Pollack (1992), Moore and Paris (1993), Korelsky and Kittredge (1993), Moser and Moore (1996), Rino (1996), Marcu (1999, 2000), etc.: mapping among the discourse levels

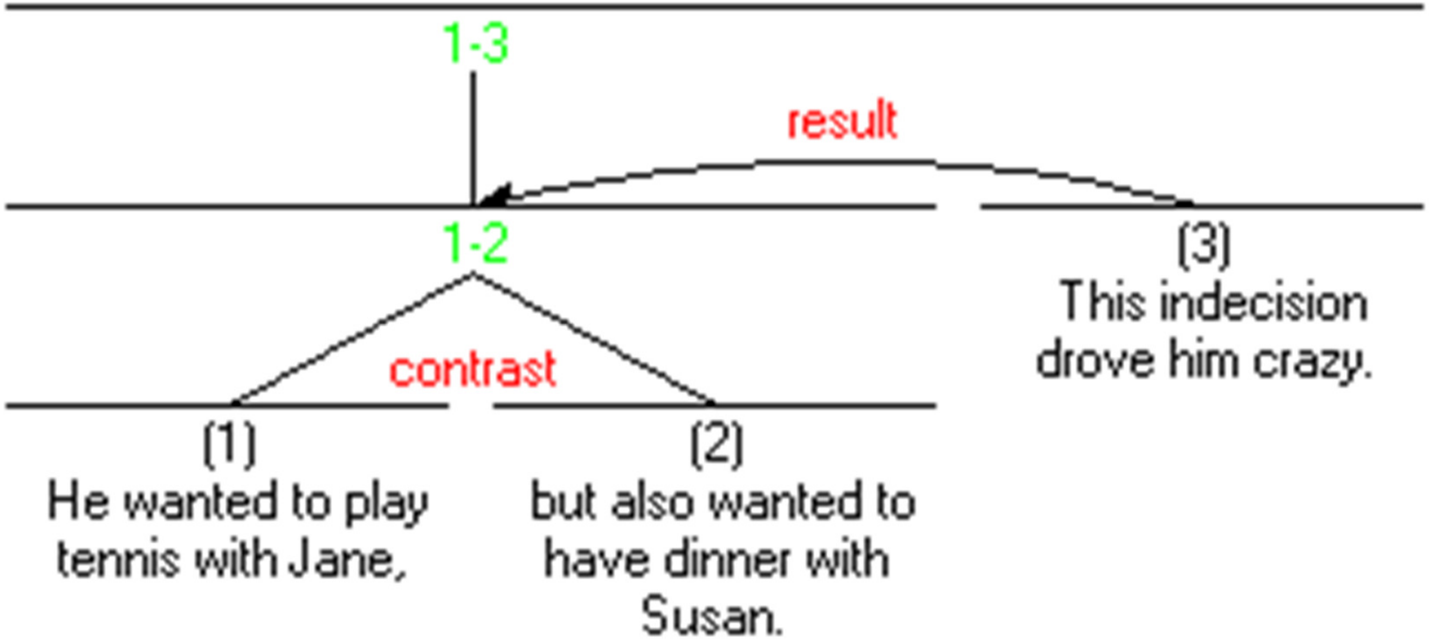
Introduction: RST

- **RST – Rhetorical Structure Theory** (Mann and Thompson, 1987)
  - One of the most used discourse theories in Computational Linguistics

- Main characteristics
  - Relates propositions by rhetorical relations
  - Attributes importance status to each proposition
    - Nucleus: most important proposition in the relation
    - Satellite: complementary information to the nucleus
  - Discourse structures are hierarchical tree-shaped structures

Introduction: example

- The arrow leaves from the satellite and points to the nucleus of the relation
- Some relations are multinuclear: contrast


1-3
result
1-2
(3)
This indecision drove him crazy.
contrast
(1)
He wanted to play tennis with Jane,
(2)
but also wanted to have dinner with Susan.

Motivation

- Few researches and resources for Portuguese
  - No discourse analyzer for this language
  - The “jumping a level” phenomenon

Motivation

- Very useful for NLP
  - Anaphora resolution (Cristea et al., 1998; Schauer, 2000; Seno, 2005)
  - Text summarization (Rino, 1996; O'Donnel, 1997; Marcu, 2000; da Cunha et al., 2009; Uzêda et al., 2010)
  - Machine translation (Marcu et al., 2000)
  - Essay scoring tools (Burstein et al., 2003)
  - Text generation (Moore and Paris, 1993; Rino, 1996)
  - Question answering (Bosma, 2005)

Motivation

- Some discourse analyzers already available for English (Marcu, 2000; Soricut and Marcu, 2003) and Japanese (Sumita et al., 1992)

- None for Portuguese
  - To our knowledge, DiZer (Pardo et al., 2008) was the first one

Project decisions

- Segmentation
  - Phrases vs. clauses vs. sentences vs. paragraphs
- Relation set
  - Generic vs. specific relations
- Research approach
  - Symbolic (linguistic knowledge) vs. statistical
- Text genre and type, etc.

DiZer development

- Knowledge-based approach
  - Corpus study for identifying how discourse relations are signaled in texts
    - Discourse markers
      - “However”, “therefore”, “in order to”, etc.
    - Indicative phrases and words
      - “The results are…”, “The purpose of this work is…”, etc.

DiZer: corpus

- **100 scientific texts**
  - Taken from introduction sections of Computer Science Theses
    - c.a. 53.000 words and 1.350 sentences

- Reasons for choosing these texts
  - Supposedly well written
  - More superficial makers available
  - Other works in discourse analysis for Portuguese have used the same sort of text (Feltrim et al., 2003; Pardo and Rino, 2002)

DiZer: corpus annotation

- The corpus was manually annotated
  - RSTTool (O’Donnel, 1997)
    - Edition environment, computational facilities
  - Discourse annotation manual (Carlson and Marcu, 2001)
    - Developed for English, but equally applicable for Portuguese, since RST is language independent
    - Consistent annotation, as noise-free as possible

- Only one annotator, expert in RST
  - For consistence in annotation
  - For time limitation

DiZer: relations set

- 32 relations: added ones to the original set in bold

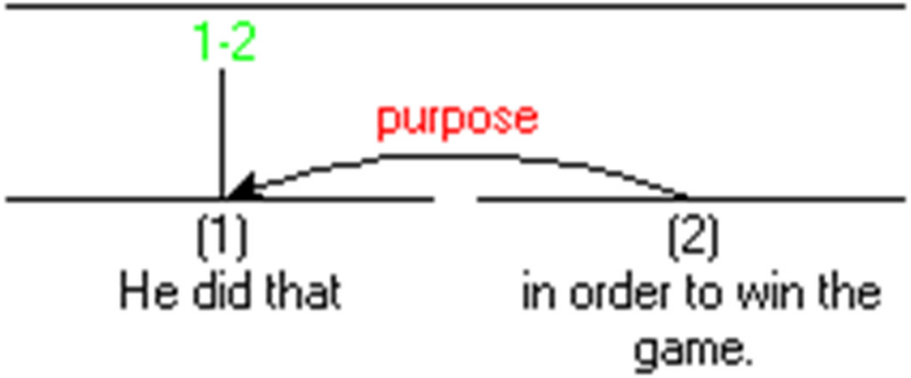
antithesis	contrast	justify	purpose
attribution	elaboration	list	restatement
background	enablement	means	**same-unit**
circumstance	evaluation	motivation	sequence
comparison	evidence	non-vol-cause	solutionhood
concession	**explanation**	non-vol-result	summary
conclusion	interpretation	otherwise	vol-cause
condition	joint	**parenthetical**	vol-result

DiZer: corpus analysis

- More than 750 discourse analysis patterns
  - Codify the correspondence between textual markers and discourse relations

- They constitute DiZer main information repository

DiZer: example of pattern



Corpus analysis

Relation	purpose
Order of segments	nucleus before satellite
Marker in 1st segment	---
Position of Marker	---
Marker in 2nd segment	*in order to*
Position of Marker	beginning

DiZer: patterns

- They can also incorporate morphosyntactic information and user-defined knowledge (genre-specific)
  - “The purpose of this work”, “The aim of these projects”, etc.

Relation	purpose
Order of segments	nucleus before satellite
Marker1 in 1st segment	---
Position of Marker1	---
Marker2 in 2nd segment	*ART purposeClass of PRON workClass*
Position of Marker2	beginning

DiZer: patterns

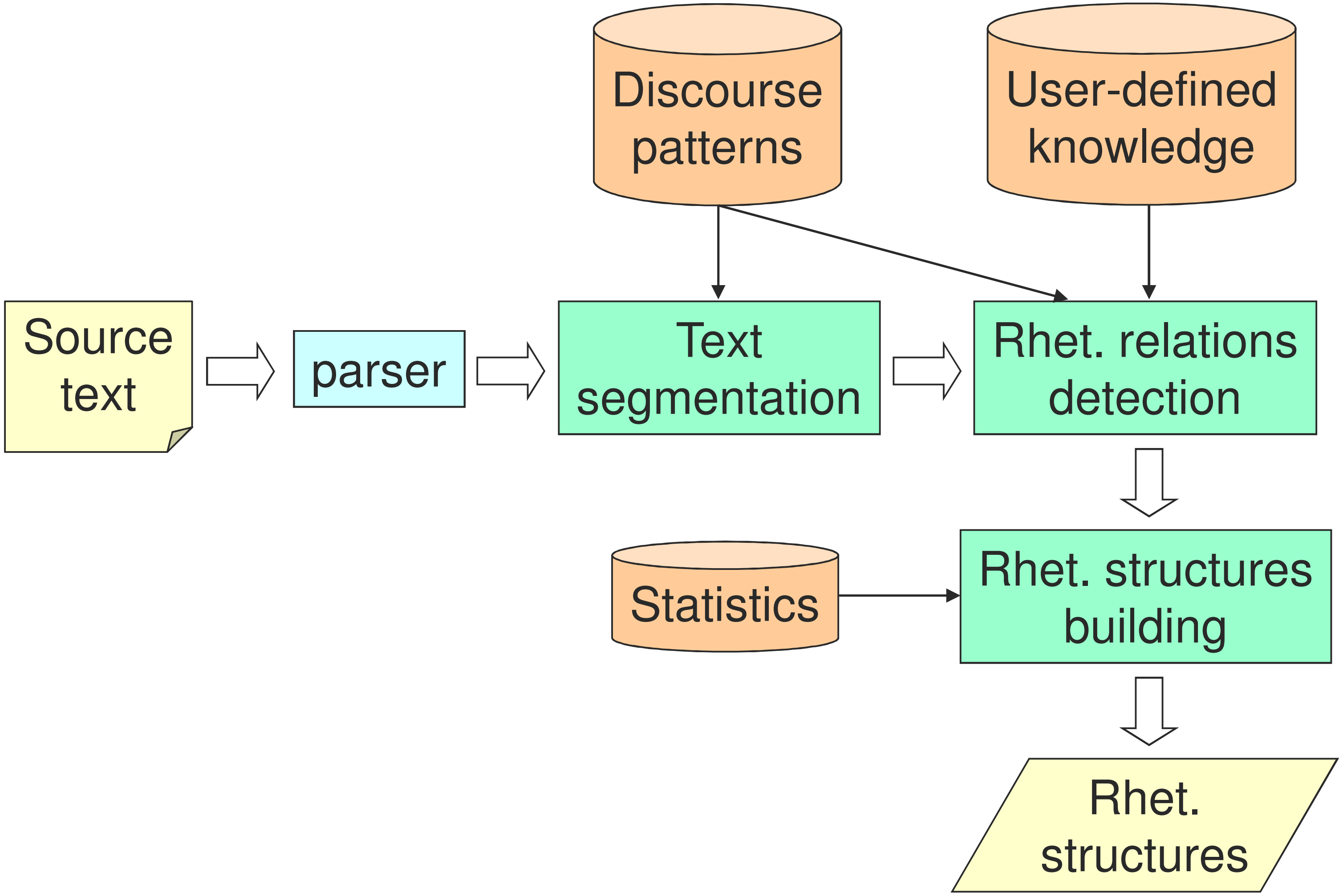
- They can also incorporate morphosyntactic information and user-defined knowledge (genre-specific)
  - “The purpose of this work”, “The aim of these projects”, etc.

Relation	purpo
Order of segments	nucle tellite
Marker1 in 1st segment	---
Position of Marker1	---
Marker2 in 2nd segment	*ART purposeClass of PRON workClass*
Position of Marker2	beginning

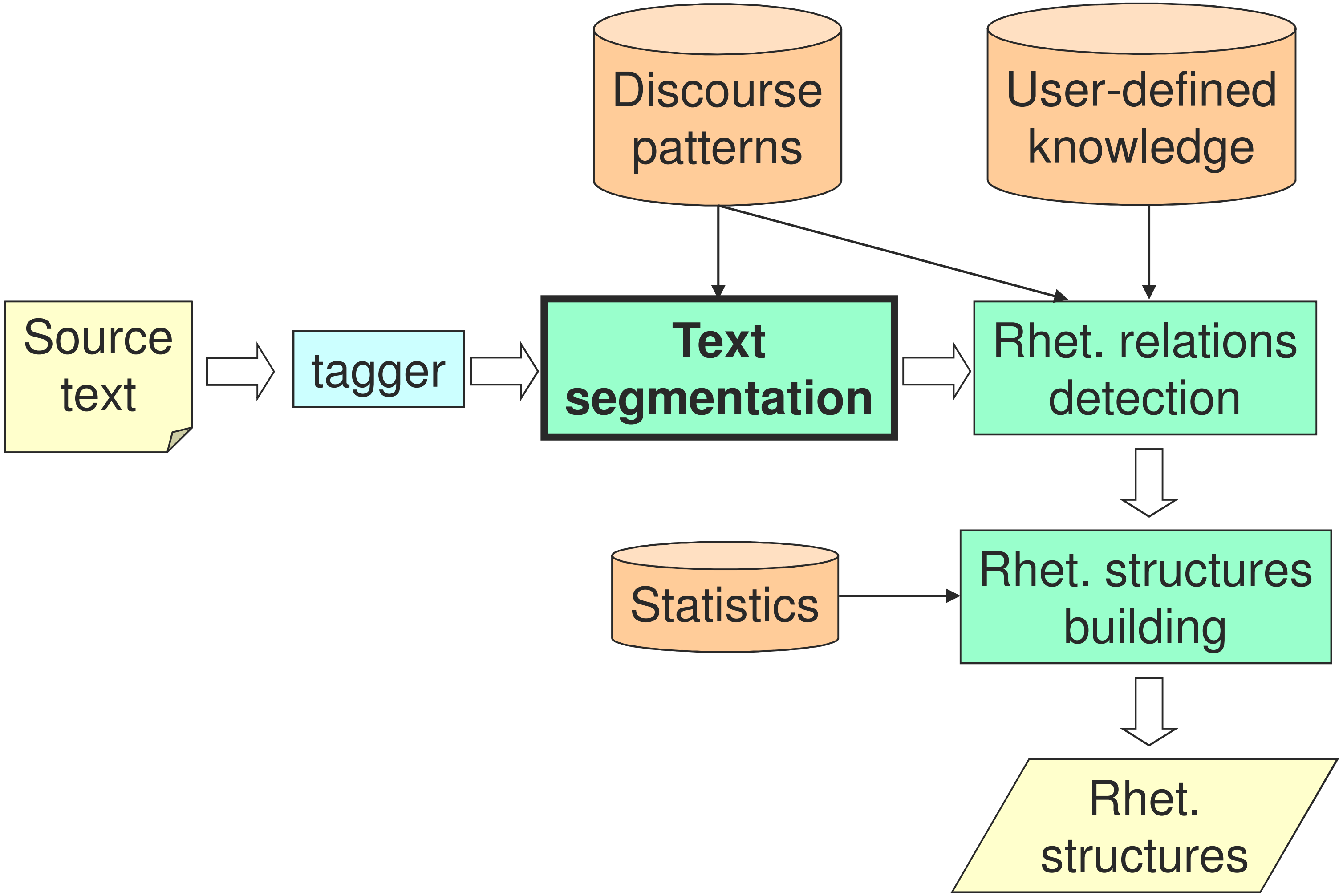
purpose
aim
objective
…

work
project
research
…

DiZer: architecture



DiZer: architecture



DiZer: text segmentation

- Tries to determine the simple clauses
  - Simple punctuation-based rules
    - Comma, dot, interrogation and exclamation signals
    - Abbreviation list
  - Verification of strong discourse markers presence
    - Use of discourse analysis patterns
  - Verification of verb presence in the detected segments
    - Use of POS tags

DiZer: text segmentation

- Syntactical-based rules
  - “Segment the text in the boundaries of relative clauses”
  - “Segment the text in coordinative and subordinate conjunctions”
  - Etc.

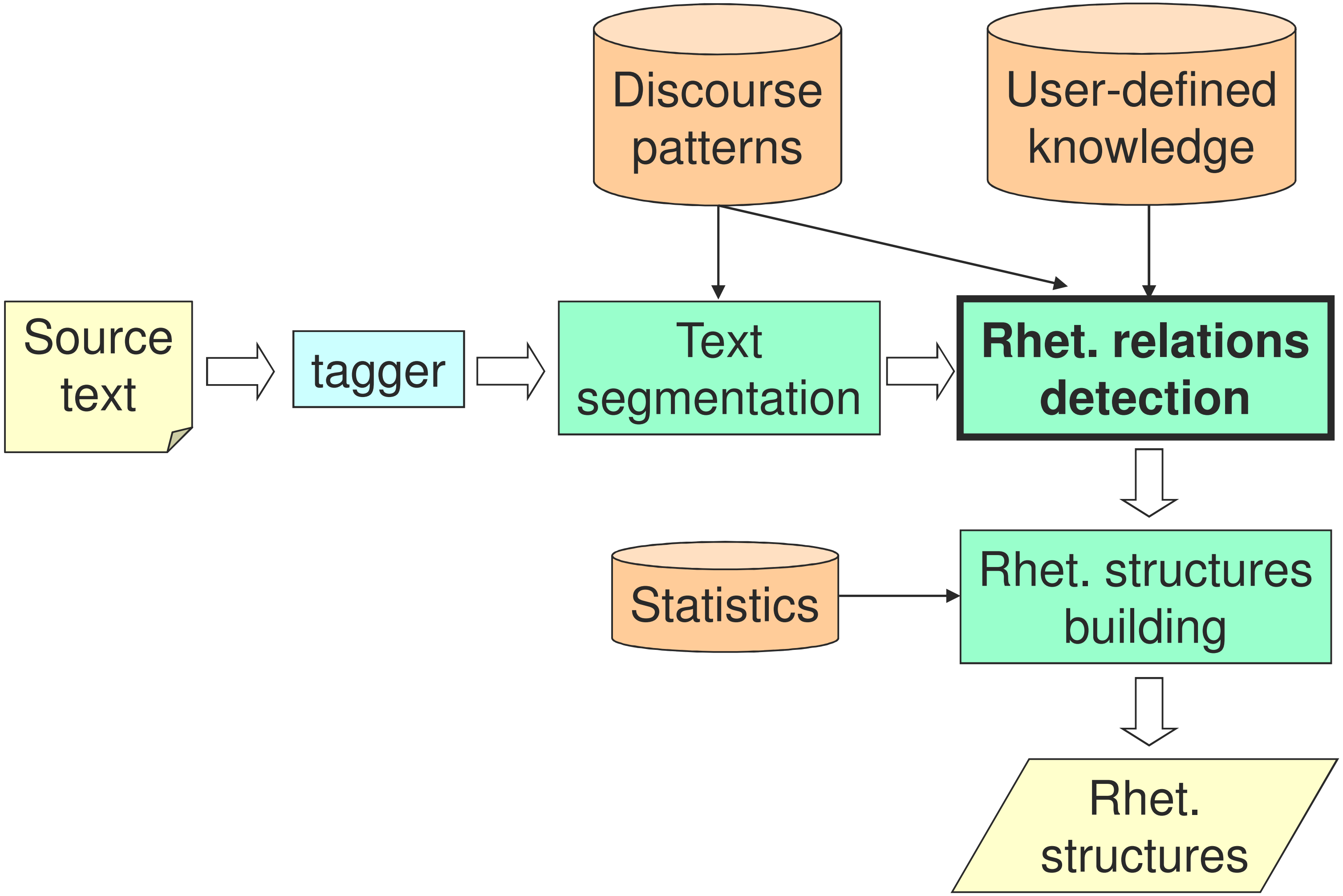
DiZer: text segmentation

*He wanted to play tennis with Jane**, but** also wanted to have dinner with Susan**.** This indecision drove him crazy**.***

⇩

[1] He wanted to play tennis with Jane
[2] but also wanted to have dinner with Susan.
[3] This indecision drove him crazy.

DiZer: architecture



DiZer: rhet. relations detection

- Pattern-matching process between discourse patterns and segments
  - All possible relations are detected
  - If no patterns are found, a default elaboration relation is hypothesized to occur
    - Elaboration is the most frequent relation observed in the corpus, since it is too generic

- Output of this step
  - A set of possible rhetorical relations between propositions

DiZer: rhet. relations detection

[1] He wanted to play tennis with Jane
[2] but also wanted to have dinner with Susan.
[3] This indecision drove him crazy.

⇩

rhetorical_relation(contrast, 1, 2)
rhetorical_relation(result, 3, [1-2])

DiZer: rhet. relations detection

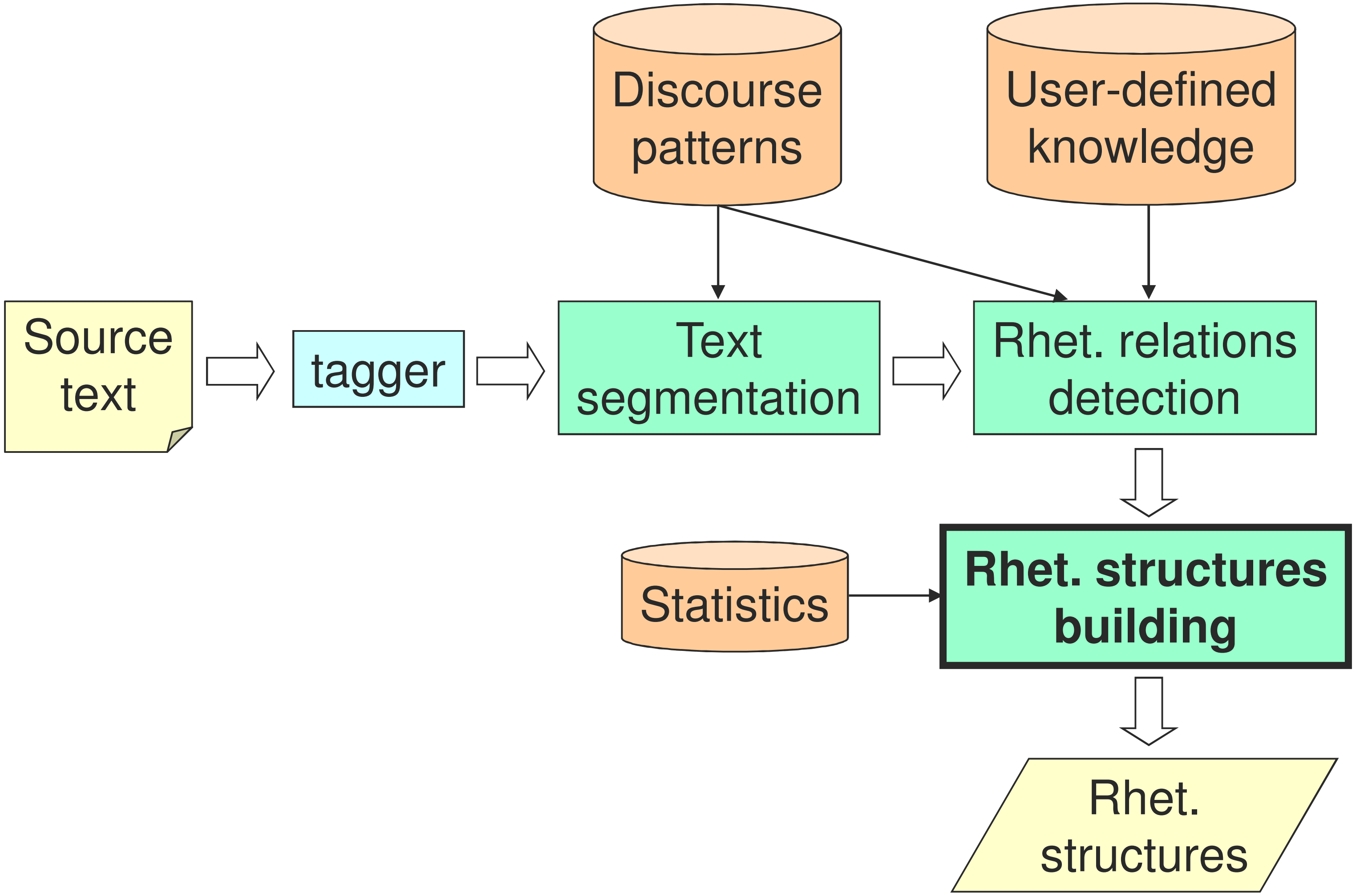
- Analysis is carried out incrementally
  - First, adjacent segments inside a sentence are related
  - Then, adjacent sentences inside a paragraph are related
  - Finally, adjacent paragraphs are related

- Justification for this strategy
  - Writers tend to put together related information
  - Makes computational processing feasible

DiZer: rhet. relations detection

- Limitations
  - In “actual” discourse analysis, not all the relations are established between adjacent segments
  - Most of segments are not signaled by any markers
    - Result: big amount of elaboration relations

DiZer: architecture


Discourse patterns

User-defined knowledge

Source text

tagger

Text segmentation

Rhet. relations detection

Statistics

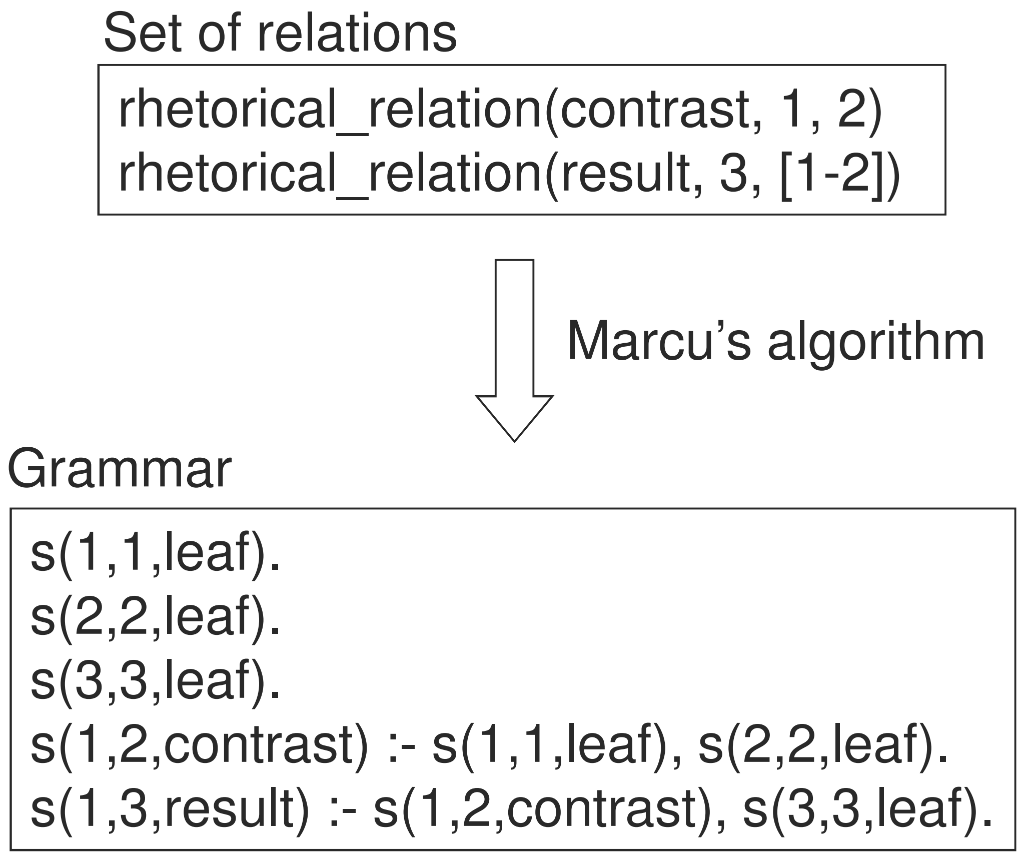
Rhet. structures building

Rhet. structures

DiZer: rhet. structures building

- The rhetorical relations hypothesized before are joined in possible valid rhetorical structures
  - Use of Marcu’s algorithm (1997)
    - It maps the rhet. relations hypothesized into a prolog/DCG grammar
    - The generated grammar produces all possible valid rhetorical structures

DiZer: rhet. structures building



Set of relations

```
rhetorical_relation(contrast, 1, 2)
rhetorical_relation(result, 3, [1-2])
```

Marcu’s algorithm

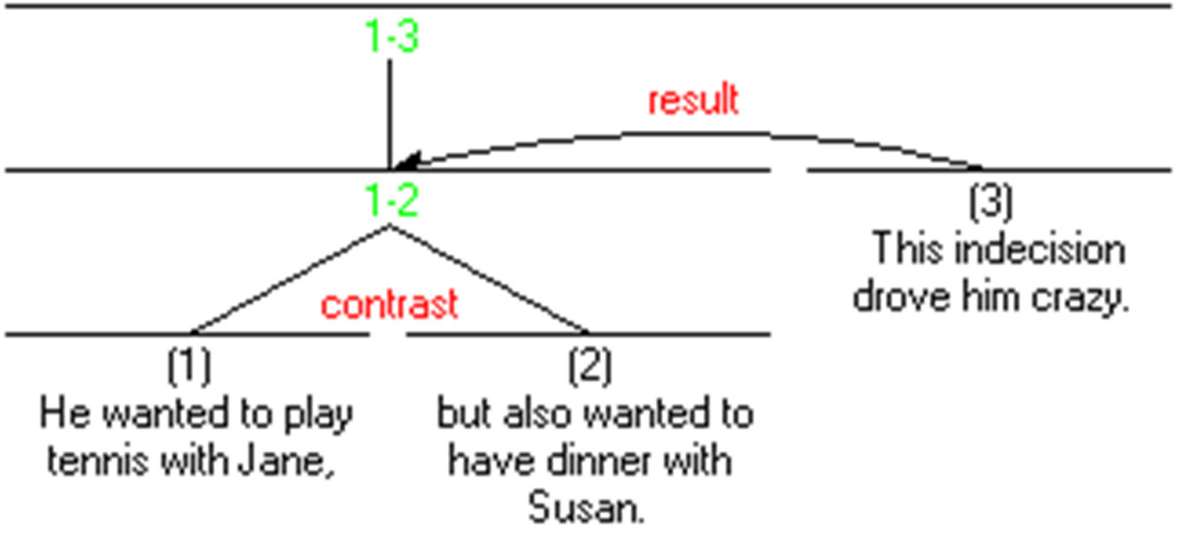
Grammar

```
s(1,1,leaf).
s(2,2,leaf).
s(3,3,leaf).
s(1,2,contrast) :- s(1,1,leaf), s(2,2,leaf).
s(1,3,result) :- s(1,2,contrast), s(3,3,leaf).
```

DiZer: rhet. structures building

```
s(1,1,leaf).
s(2,2,leaf).
s(3,3,leaf).
s(1,2,contrast) :- s(1,1,leaf), s(2,2,leaf).
s(1,3,result) :- s(1,2,contrast), s(3,3,leaf).
```

⇩ Grammar running



(1) He wanted to play tennis with Jane,

(2) but also wanted to have dinner with Susan.

(3) This indecision drove him crazy.

DiZer: rhet. structures building

- The resulting structures are ranked by their **probabilities**
  - Probabilities learned from corpus
    - Probability of a relation node and its children with their nuclearity
    - Simple frequency counts

$$P(t) = \prod_{i=1}^{nro\_rel} P(child_{left}, n_{left}, child_{right}, n_{right} \mid parent_i)$$

Evaluating DiZer performance

- Comparison of DiZer structures with the ones predicted in Rhetalho corpus (Pardo and Seno, 2005)

- Rhetalho
  - Reference corpus with 50 texts
    - Agreement
  - Scientific and news annotated texts
    - 2 experts in RST
    - Annotation protocol

Evaluation

- Selected texts for evaluation
  - **20 scientific texts**
  - **5 news texts**
    - Testing DiZer performance for other text genres
    - Discourse markers are consistently used across different text genres, types and domains

- Methods evaluated
  - DiZer with <u>clausal segmentation</u>
  - DiZer with <u>sentential segmentation</u>
  - <u>Baseline</u> method: sentential segmentation and elaboration relations

Evaluation

- DiZer main tasks
  - Text segmentation
  - Nuclearity determination
  - Relations detection

- Recall, precision and f-measure (%)
  - Recall: how many reference elements are produced
  - Precision: how many produced elements are correct
  - F-measure: combination of recall and precision

Evaluation

- Scientific texts
  - DiZer (with both segmentation methods) outperforms the baseline method

	DiZer clauses			**DiZer sentences**			**Baseline**		
Tasks	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
Segmentation	**57,3**	**56,2**	**56,8**	25,2	41,7	31,4	25,2	41,7	31,4
Nuclearity	**79,7**	**82,3**	**80,9**	39,1	69,5	50,1	32,4	59,5	42,0
Relations	**63,2**	**61,9**	**62,5**	28,7	61,0	39,1	20,7	49,2	29,2

Evaluation

- News texts
  - Only DiZer with clausal segmentation outperforms the baseline method

	DiZer clauses			DiZer sentences			Baseline		
Tasks	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
Segmentation	**48,8**	**54,1**	**51,3**	**9,9**	**20,6**	**13,4**	9,9	20,6	13,4
Nuclearity	**55,8**	**63,5**	**59,4**	**22,3**	**55,3**	**31,8**	28,4	71,3	40,7
Relations	**37,8**	**43,2**	**40,3**	**12,5**	**38,3**	**18,9**	17,6	58,3	27,0

Evaluation

- DiZer and English analyzers performance
  - DiZer presents satisfactory results

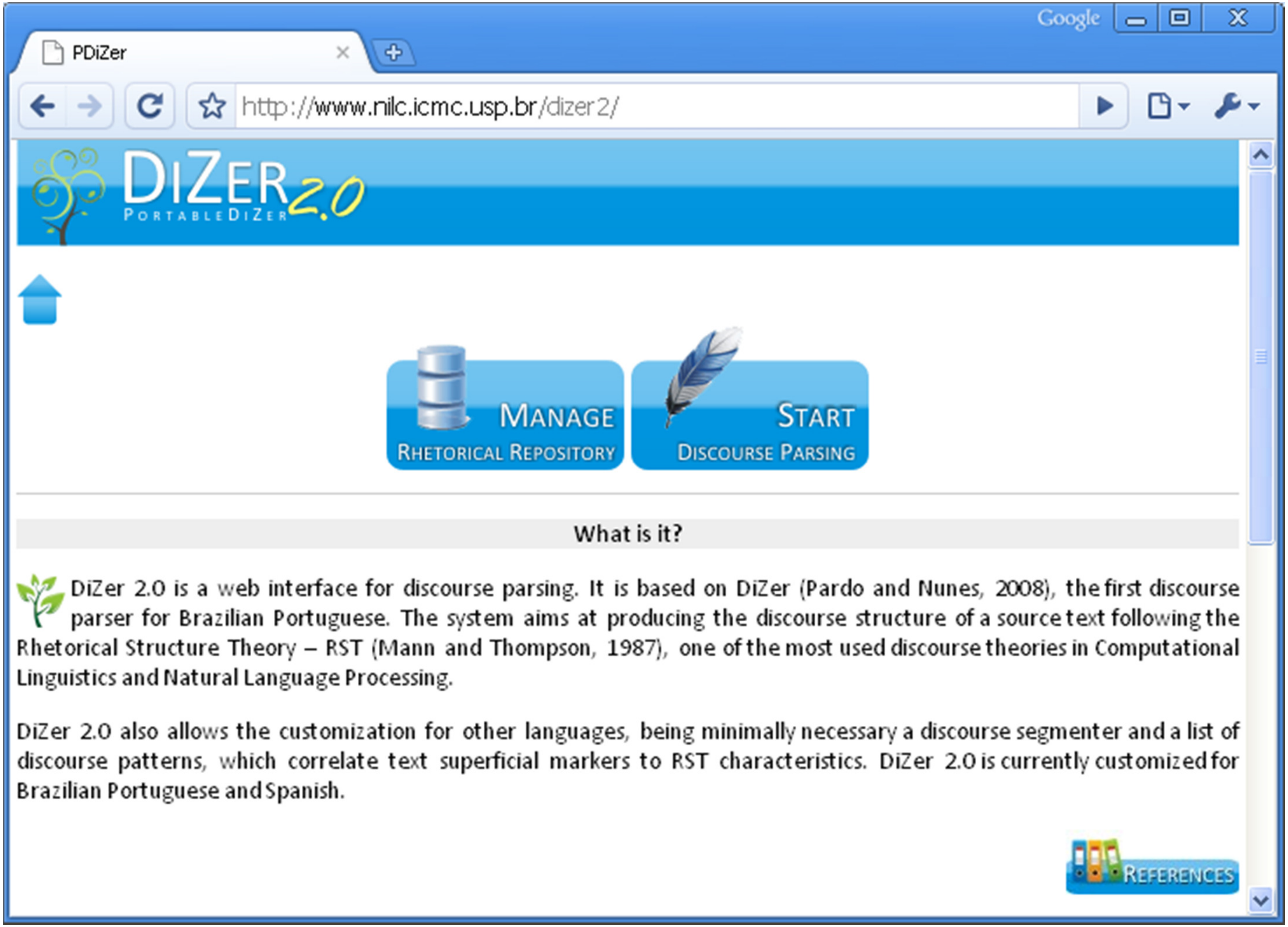
	DiZer	**English analyzers**
Tasks	*F*	*F*
Segmentation	**56,8**	84-97
Nuclearity	**80,9**	63
Relations	**62,5**	49-75

DiZer

- Well… nice, but a “white elephant”
  - Difficult to install
    - Prolog, Perl, C, Delphi
  - Difficult to use
  - Difficult to customize for other languages
  - Etc.

DiZer 2.0

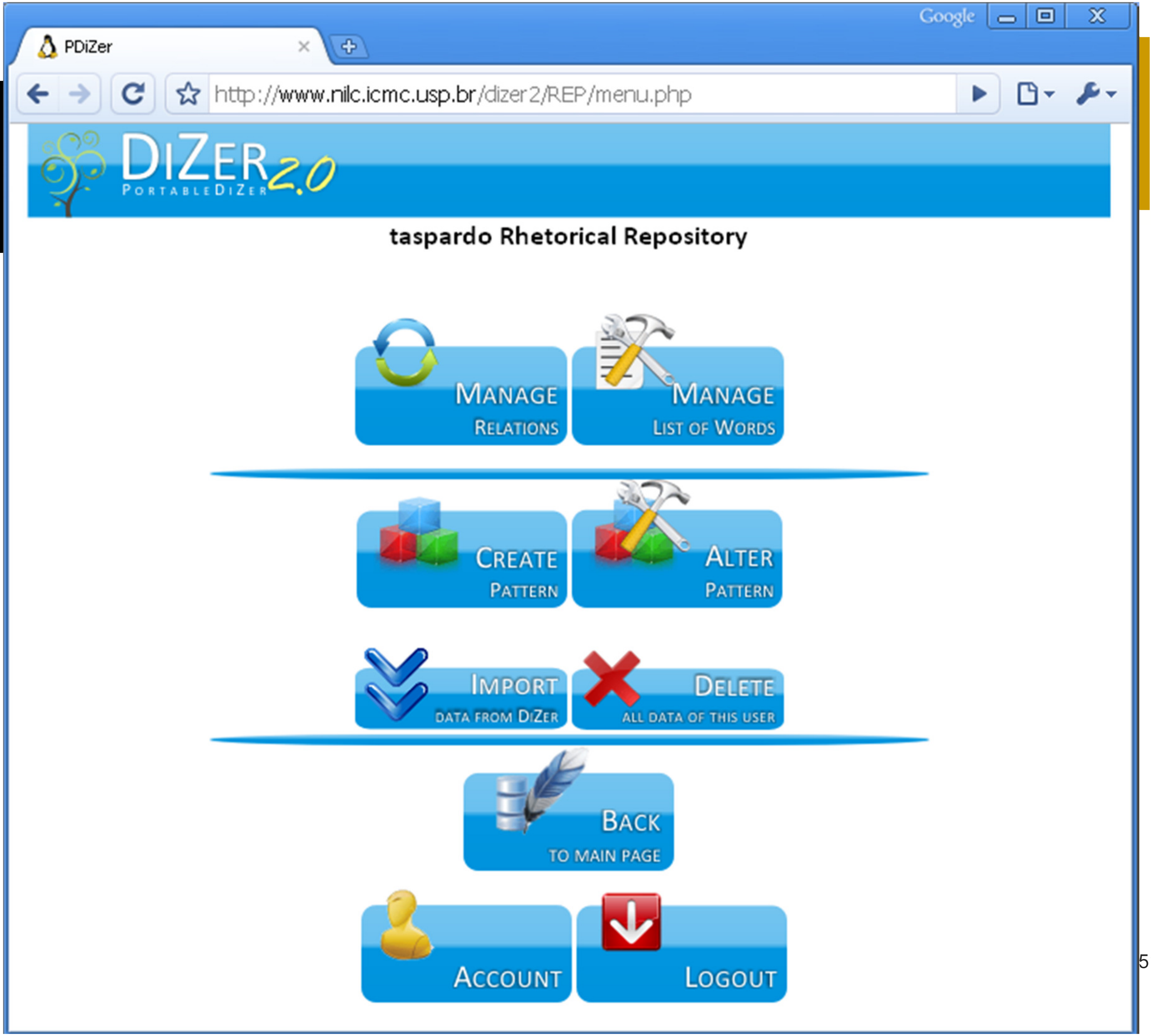
- Web interface
  - Not necessary to install anything
- Easy to use and customize
- Light version
- Collaboration with IULA and TALNE
  - Iria da Cunha Fanego
  - Juan-Manuel Torres-Moreno, Eric SanJuan



What is it?

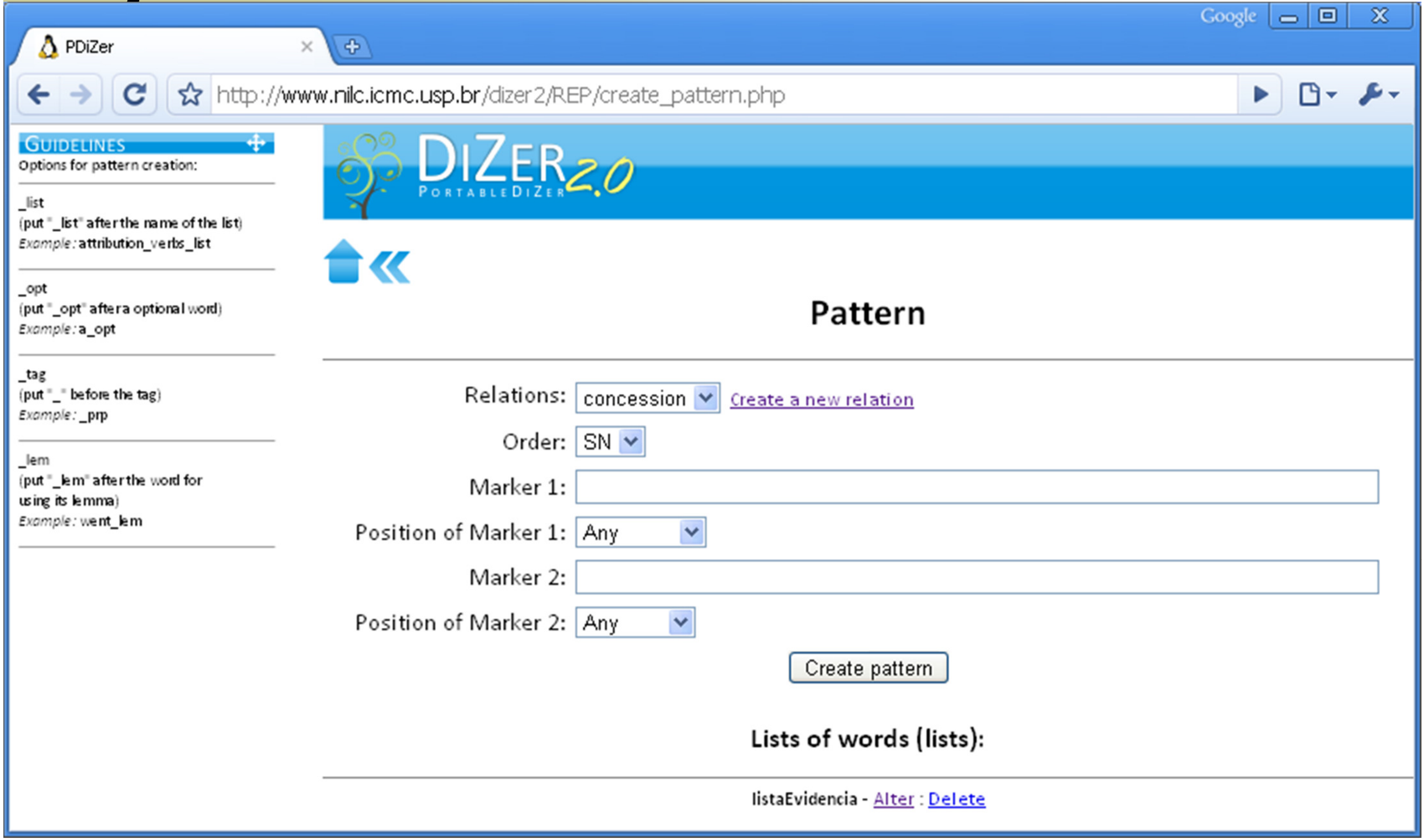
DiZer 2.0 is a web interface for discourse parsing. It is based on DiZer (Pardo and Nunes, 2008), the first discourse parser for Brazilian Portuguese. The system aims at producing the discourse structure of a source text following the Rhetorical Structure Theory – RST (Mann and Thompson, 1987), one of the most used discourse theories in Computational Linguistics and Natural Language Processing.

DiZer 2.0 also allows the customization for other languages, being minimally necessary a discourse segmenter and a list of discourse patterns, which correlate text superficial markers to RST characteristics. DiZer 2.0 is currently customized for Brazilian Portuguese and Spanish.



taspardo Rhetorical Repository

Creating a pattern



GUIDELINES

Options for pattern creation:

_list
(put "_list" after the name of the list)
Example: attribution_verbs_list

_opt
(put "_opt" after a optional word)
Example: a_opt

_tag
(put "_" before the tag)
Example: _prp

_lem
(put "_lem" after the word for using its lemma)
Example: went_lem

Pattern

Relations: concession Create a new relation

Order: SN

Marker 1:

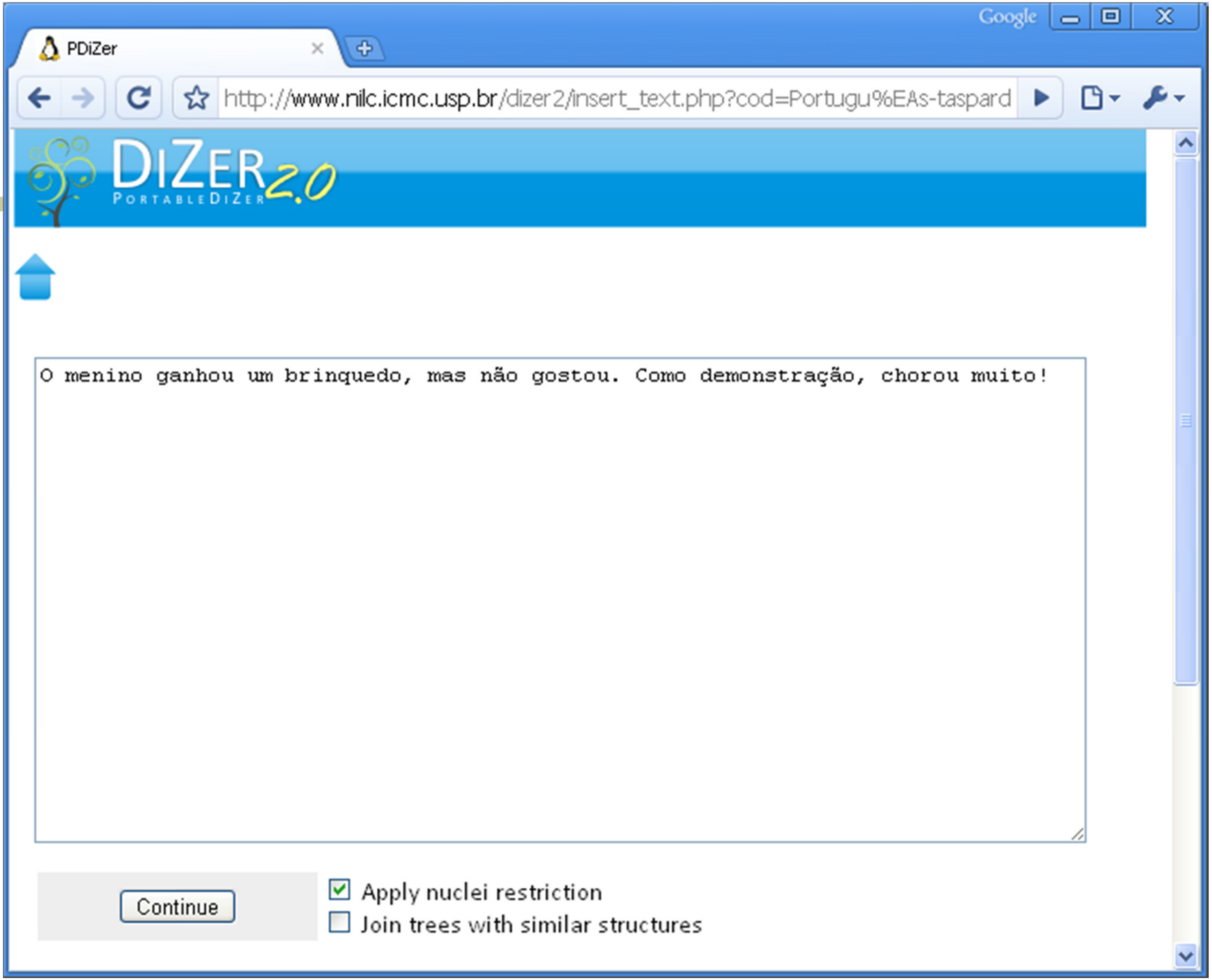
Position of Marker 1: Any

Marker 2:

Position of Marker 2: Any

Lists of words (lists):

listaEvidencia - Alter : Delete

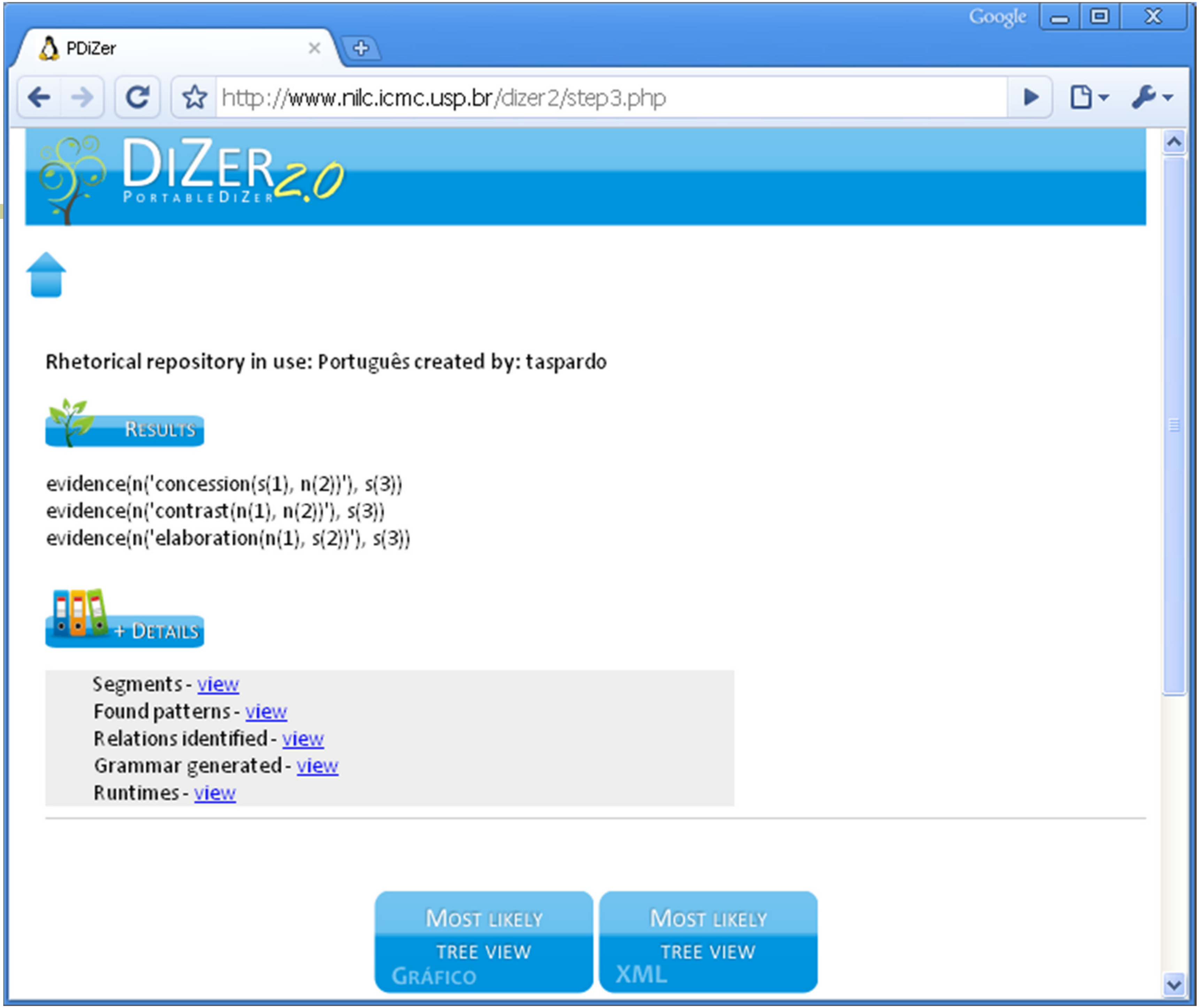


PDiZer
http://www.nilc.icmc.usp.br/dizer2/insert_text.php?cod=Portugu%EAs-taspard
DiZer 2.0
PortableDiZer

O menino ganhou um brinquedo, mas não gostou. Como demonstração, chorou muito!

☑ Apply nuclei restriction

□ Join trees with similar structures



Rhetorical repository in use: Português created by: taspardo

```
evidence(n('concession(s(1), n(2))'), s(3))
evidence(n('contrast(n(1), n(2))'), s(3))
evidence(n('elaboration(n(1), s(2))'), s(3))
```

Segments - view
Found patterns - view
Relations identified - view
Grammar generated - view
Runtimes - view

DiZer 2.0

- Web interface
  - Not necessary to install anything

- Easy to use and customize

- Light version
  - Web made it worse
  - But there is room to improve

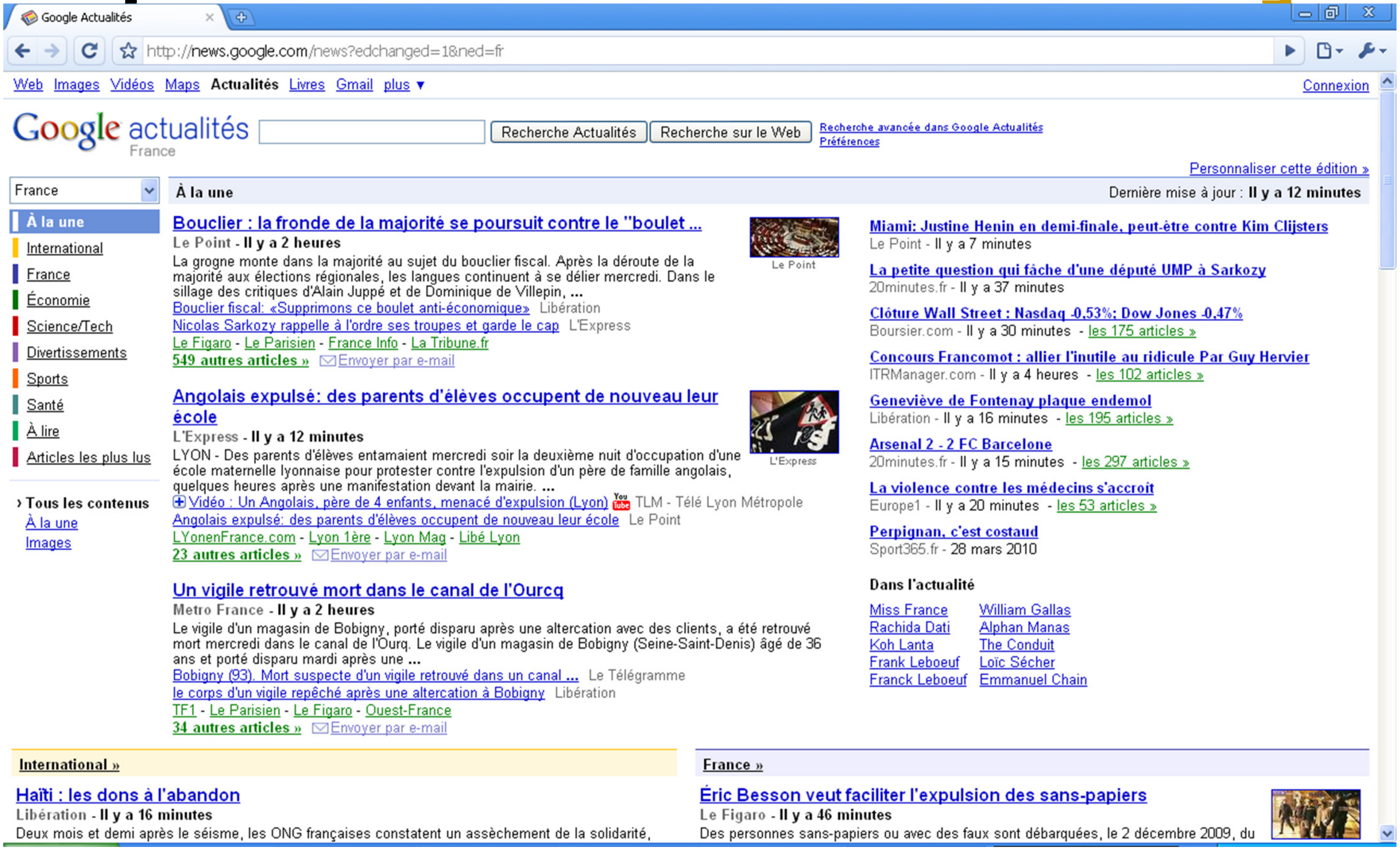
DiZer 2.0

- Brazilian Portuguese
- Beta version for Spanish
  - State of the art discourse segmentation, basic discourse patterns
- Intentions for French and Basque

More recently

- **Web** and the **information explosion** era
  - Too many documents to read and grasp the information
  - 800 exabytes of new information in 2009
    - 3 times more in 2012
  - Situation: a person wants to know about the last world economical crisis

Google News



À la une — Dernière mise à jour : **Il y a 12 minutes**

Bouclier : la fronde de la majorité se poursuit contre le "boulet ...
Le Point - **Il y a 2 heures**
La grogne monte dans la majorité au sujet du bouclier fiscal. Après la déroute de la majorité aux élections régionales, les langues continuent à se délier mercredi. Dans le sillage des critiques d'Alain Juppé et de Dominique de Villepin, **...**
Bouclier fiscal: «Supprimons ce boulet anti-économique» Libération
Nicolas Sarkozy rappelle à l'ordre ses troupes et garde le cap L'Express
Le Figaro - Le Parisien - France Info - La Tribune.fr
549 autres articles » ✉ Envoyer par e-mail

Angolais expulsé: des parents d'élèves occupent de nouveau leur école
L'Express - **Il y a 12 minutes**
LYON - Des parents d'élèves entamaient mercredi soir la deuxième nuit d'occupation d'une école maternelle lyonnaise pour protester contre l'expulsion d'un père de famille angolais, quelques heures après une manifestation devant la mairie. **...**
⊞ Vidéo : Un Angolais, père de 4 enfants, menacé d'expulsion (Lyon) TLM - Télé Lyon Métropole
Angolais expulsé: des parents d'élèves occupent de nouveau leur école Le Point
LYonenFrance.com - Lyon 1ère - Lyon Mag - Libé Lyon
23 autres articles » ✉ Envoyer par e-mail

Un vigile retrouvé mort dans le canal de l'Ourcq
Metro France - **Il y a 2 heures**
Le vigile d'un magasin de Bobigny, porté disparu après une altercation avec des clients, a été retrouvé mort mercredi dans le canal de l'Ourq. Le vigile d'un magasin de Bobigny (Seine-Saint-Denis) âgé de 36 ans et porté disparu mardi après une **...**
Bobigny (93). Mort suspecte d'un vigile retrouvé dans un canal ... Le Télégramme
le corps d'un vigile repêché après une altercation à Bobigny Libération
TF1 - Le Parisien - Le Figaro - Ouest-France
34 autres articles » ✉ Envoyer par e-mail

Miami: Justine Henin en demi-finale, peut-être contre Kim Clijsters
Le Point - Il y a 7 minutes

La petite question qui fâche d'une député UMP à Sarkozy
20minutes.fr - Il y a 37 minutes

Clôture Wall Street : Nasdaq -0,53%; Dow Jones -0,47%
Boursier.com - Il y a 30 minutes - les 175 articles »

Concours Francomot : allier l'inutile au ridicule Par Guy Hervier
ITRManager.com - Il y a 4 heures - les 102 articles »

Geneviève de Fontenay plaque endemol
Libération - Il y a 16 minutes - les 195 articles »

Arsenal 2 - 2 FC Barcelone
20minutes.fr - Il y a 15 minutes - les 297 articles »

La violence contre les médecins s'accroît
Europe1 - Il y a 20 minutes - les 53 articles »

Perpignan, c'est costaud
Sport365.fr - 28 mars 2010

Dans l'actualité

Miss France	William Gallas
Rachida Dati	Alphan Manas
Koh Lanta	The Conduit
Frank Leboeuf	Loïc Sécher
Franck Leboeuf	Emmanuel Chain

International »

Haïti : les dons à l'abandon
Libération - **Il y a 16 minutes**
Deux mois et demi après le séisme, les ONG françaises constatent un assèchement de la solidarité,

France »

Éric Besson veut faciliter l'expulsion des sans-papiers
Le Figaro - **Il y a 46 minutes**
Des personnes sans-papiers ou avec des faux sont débarquées, le 2 décembre 2009, du

Multidocument scenario

- Still an unreasonable amount of information
- Several subtopics
- Different perspectives and focuses
- Different styles and sources
- Redundant, complementary and contradictory information
- Different time and event ordering

Multidocument scenario

- To automatically deal with this world, some organization is necessary
- Multidocument discourse models

A bit of history

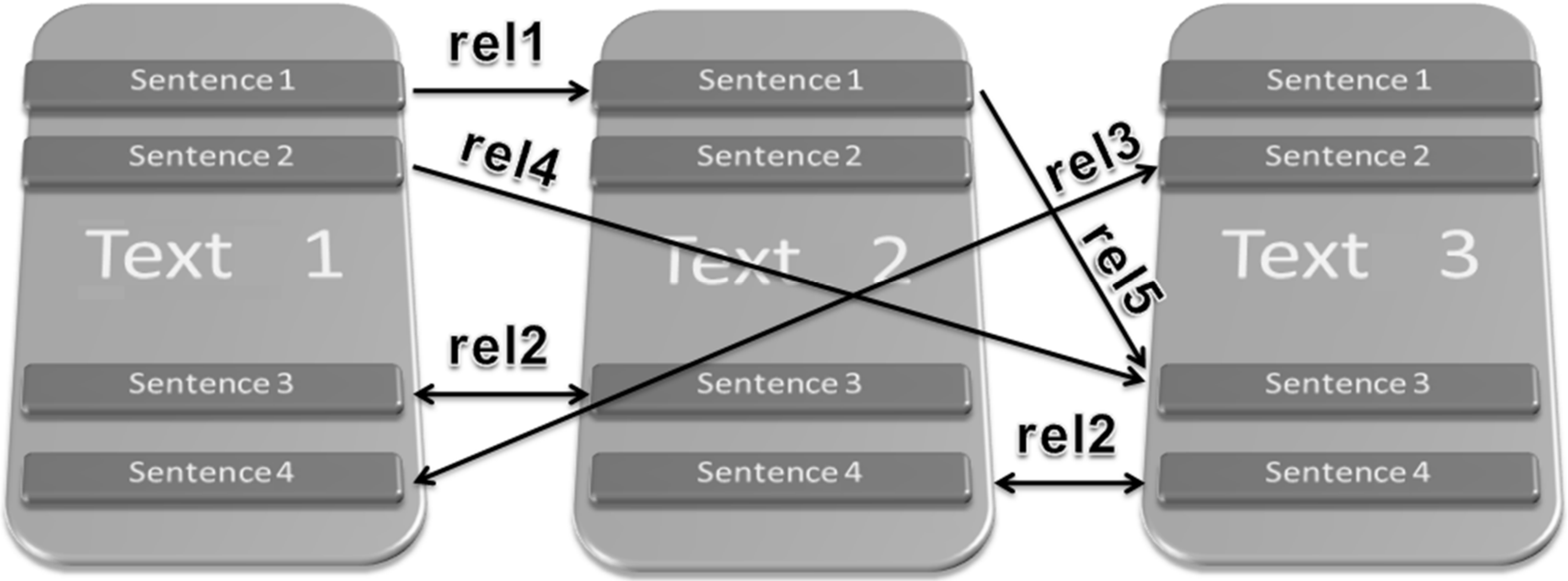
- Trigg and the TextNet system (1983, 1986)
- RST (Mann and Thompson, 1987)
- Mckeown and Radev (1995): SUMMONS and summarization operators
- Radev (2000): **CST** (*Cross-document Structure Theory*)
- Afantenos et al. (2004) and criticism of the model
- Success in multidocument applications (Radev et al., 2000, 2001; Zhang et al., 2002; Afantenos et al., 2004, 2007)

CST

- Cross-document Structure Theory
  - Multidocument discourse theory
  - 24 relations for documents on related topics
  - Complementary data structures
    - Multidocument cube and graph

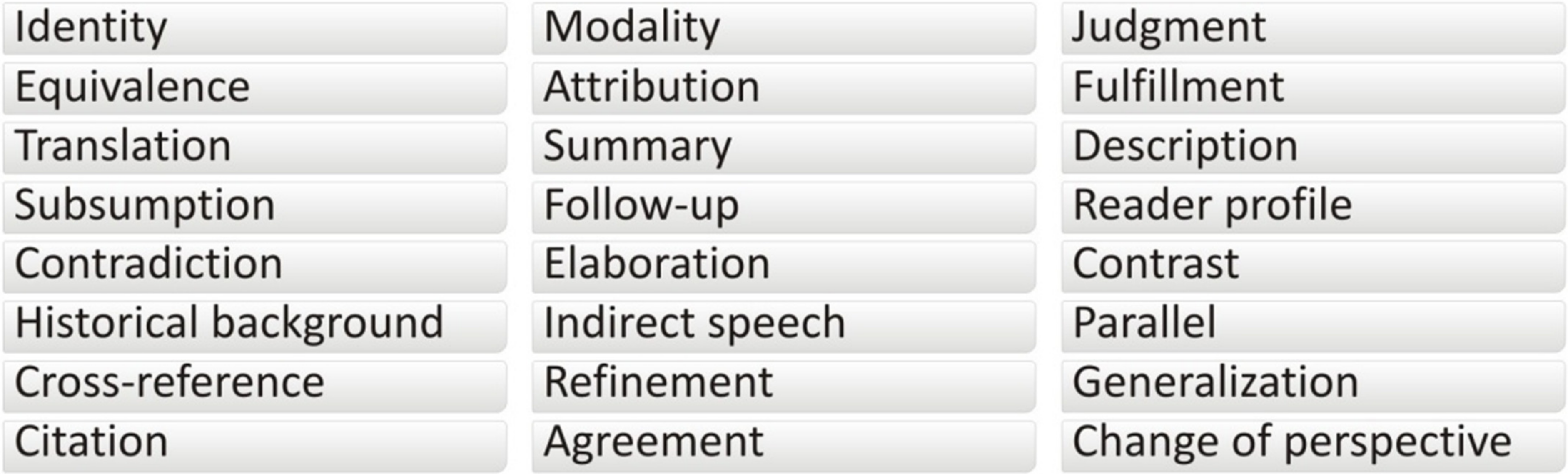
CST

- Multidocument structuring
  - Relations among text spans across documents



CST

- Original relations
  - Low annotation agreement, ambiguity



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

Example

- *Contradiction*, *overlap*, *historical background* (←)

An airplane accident in Bukavu, east of Democratic Republic of Congo, killed 13 people this Thursday in the afternoon.

At least 17 people died after an airplane fell down at Democratic Republic of Congo. Congo has a history of more than 30 airplane tragedies.

CST parsing

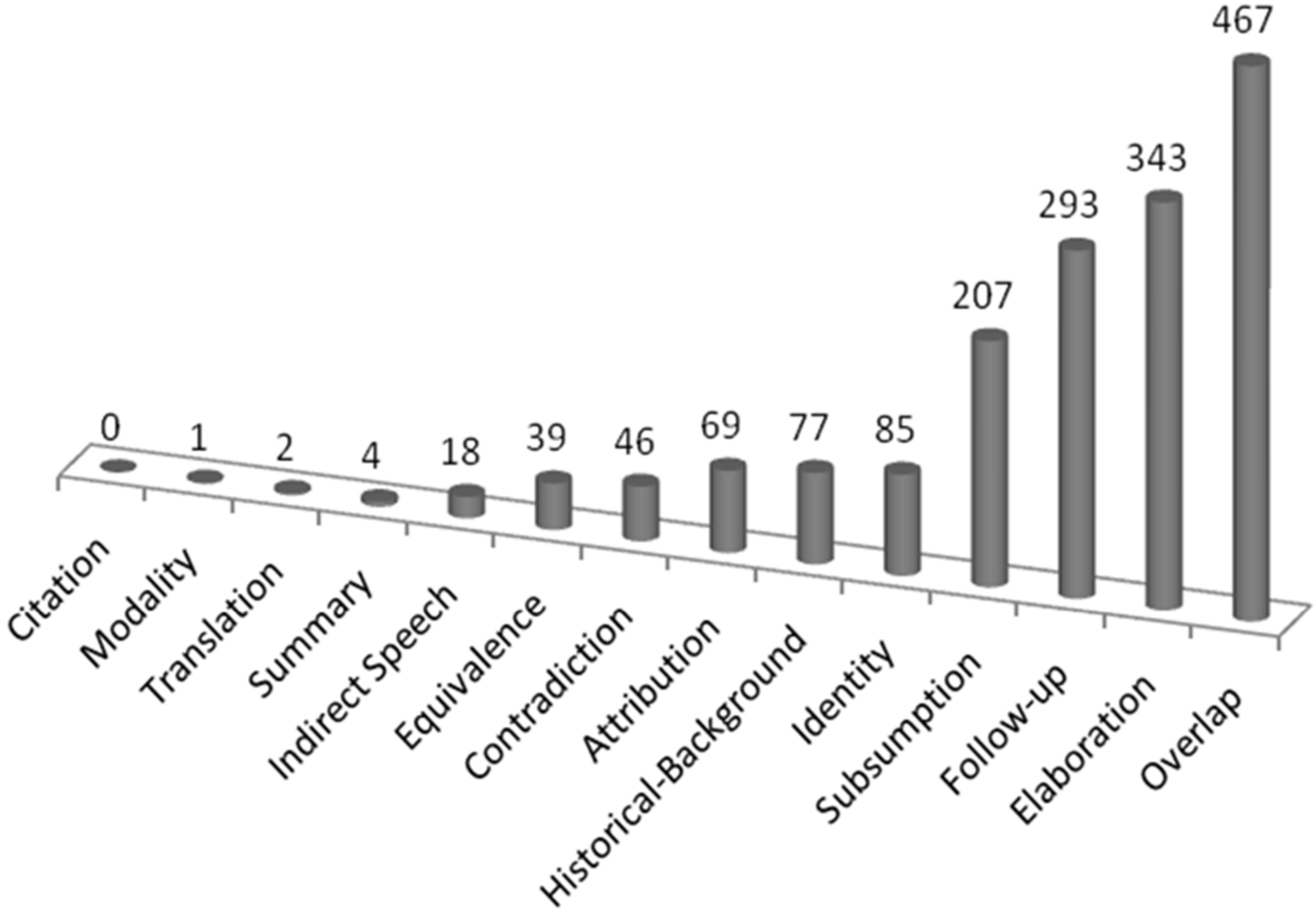
- One single CST parser for English
  Zhang et al. (2003, 2004)
  - Bad results
    - 25% precision

- For Brazilian Portuguese
  - Better corpus annotation
  - CST refinement
  - First tests with machine learning

CST parser for Portuguese

- Corpus annotation
  - 50 clusters of texts on related topics
    - 2 or 3 texts in each cluster
  - Several months of training before annotating the corpus
  - Slight modifications for some relations
  - 81% total or partial agreement among 4 humans
  - Kappa = 0.55 (vs 0.25 for English)

Corpus


0
1
2
4
18
39
46
69
77
85
207
293
343
467
Citation
Modality
Translation
Summary
Indirect Speech
Equivalence
Contradiction
Attribution
Historical-Background
Identity
Subsumption
Follow-up
Elaboration
Overlap

CSTTool

- First machine learning experiments (WEKA)
  - Extraction of shallow attributes from every related sentence pair
    - Size, POS, position, number of nouns and verbs, etc.
  - Class: CST relation
  - Results
    - 41% precision with J48 for all the relations (vs 25% for English)
    - 77% precision with J48 for content relations group

Summarization

- RST for summarization (N vs S)
  - Better than classical summarization methods
  - The content selection method does not really matter

- CST for summarization (#relations)
  - Better than famous superficial summarization methods
  - Improve the superficial methods
    - Very simple strategies tested!

Discourse Parsing

- www.nilc.icmc.usp.br

FIM