PorSimples: Simplification of Portuguese Texts for Digital Inclusion and Accessibility

Sandra M. Aluísio (Coordinator)

http://caravelas.icmc.usp.br/wiki/

2nd Joint MSR-FAPESP Workshop – FAPESP 19/11/2008

Agenda

- Goals
- Practical Applications
- Innovation Aspects
- How we approached the challenges in this project
- Results
- PorSimples in Numbers

PorSimples Main Goal

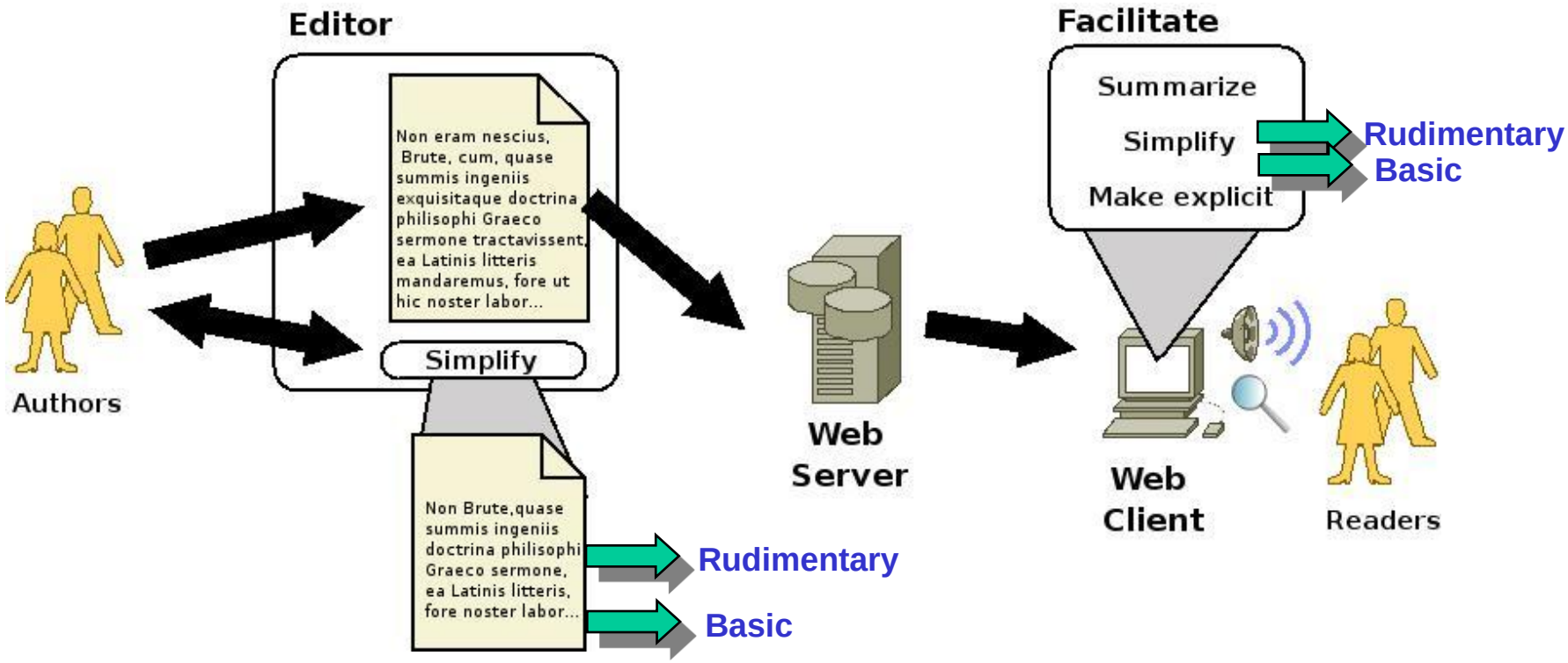
- to help poor literacy readers
  - **rudimentary level***: able only to find explicit information in short texts – e.g. advertisement or short letters;
  - **basic level***: can find information in slightly longer texts and also make simple inferences

to understand informative documents available on the web produced by

- the Brazilian government or
- relevant news agencies

* Literacy levels identified by The National Indicator of Functional Literacy (**INAF**), computed annually since 2001 by the research institution *IBOPE Opinião*

Universal Web Access: information available for readers at the basic and rudimentary literacy levels



(i) an on-line authoring system to help producing simplified texts (Simplify)

(ii) a system to allow people to read Web content (post-processing system) (Facilitate)

Possible users

- **poor literacy readers**
- **children learning to read** texts of different genres
- **people** with other **cognitive disabilities** caused by
  - medical conditions or interventions
  - aphasia and dyslexia
  - traumatic brain injuries, strokes and aneurysms
- **hearing-impaired people** who communicate with each other using sign languages like LIBRAS (Brazilian Sign Language),
  - since the structural differences between LIBRAS and Portuguese make it difficult to understand complex texts
- **people** undertaking **Distance Education**
  - in which text understandability is of great importance

Possible Clients

- Government
- Journalists
- Websites Developers
- School books publishers
- Subtitlers
- Teachers and students in bilingual education and other language-learning contexts
  - texts are usually manually adapted;
  - this is a time consuming and sometimes challenging task
- Developers of NLP (Natural Language Processing) tools
  - e.g., a parser is more likely to get a correct structure for a simple sentence than for a complex one
  - Information Extraction Tools benefit from a simple sentences text
- HCI (Human Computer Interaction) researchers
  - to ensure W3C guidelines
  - to develop assistive technologies, such as screen readers

Technology and methods

- Natural Language Processing (NLP)
  - Automatic Summarization
  - Automatic discourse analysis
  - Text Simplification (TS)

- Human Computer Interaction (HCI)
  - Building Web accessible systems for {language, vision, hearing, age}-impaired readers
  - W3C guidelines

NLP technology and methods

- Text Simplification (TS)
  - aims to maximize **understanding** of written texts through simplification of their linguistic structure by
    - replacing words only understood by a small number of people with more usual words
    - breaking down and changing the sentence syntactic structure
  - TS changes cause an impact on the text structure!
    - The usefulness of syntactic simplification can be undermined if the rewritten text lacks cohesion.
      - e. g., we need to detect and fix pronominal links that have been broken by TS operations

- NLP methods of investigation and evaluation:
  - corpus based, mainly
    - to facilitate porting to other text genres; to cater for other users (scalability)

  (i) we learn a task from a corpus and evaluate it using a corpus (intrinsically)

  (ii) evaluate and tune it with real users (this is time consuming)

Related Work

- TS is a relatively new area (first projects date from early 90s)
  - so there is room for developing new resources and methods, mainly those based on corpus.
  - Some projects consider only
    - **syntactic knowledge** to approach TS, using both rule-based systems and rules learned from a corpus (Chandrasekar, Doran and Srinivas, 1996; Chandrasekar, Srinivas, 1997)
  - Others tackle the generation of simplified texts by focusing on choices at
    - the **discourse level**, trying to answer what choices are most appropriate for people with poor literacy (Williams, 2004; Siddharthan, 2003; 2006)
  - The PSET (Practical Simplification of English Texts) project
    - investigated how **lexical-level and syntactic level** choices affect readability for a special kind of readers – **aphasics** – without considering discourse choices (Devlin, and Unthank, 2006)
  - There is a project which used a corpus to learn where to drop and where to simplify
    - but have used Siddharthan's syntactic simplifier to transform the resulting fragments into complete, grammatical sentences (Petersen and Ostendorf, 2007)
- There is no TS for Brazilian Portuguese (BP)

Innovation Aspects

Plain Language Guidelines

TS systems for English

Corpus Analysis of simple accounts in BP

1. A very detailed Manual for BP Syntactic Simplification was created:

- to implement rules of a **rule-based text simplification system**
- to guide human annotators to simplify texts (corpus creation to learn from a corpus)

- The 6 linguistic constructs (we have expanded on the number of constructs):
  - (1) apposition, (2) relative clauses, (3) subordinate clauses,
  - (4) coordinate clauses, (5) sentences with non-finite verbs, and (6) passive voice

2. We distinguish 2 levels of simplification:

- **Generating natural texts** (the manual says what and how to do BUT it does not say when to do simplifications ➔ we should learn this from corpus)

- **Generating strongly simplified texts** (follows manual rules for each text sentence)

3. Our TS approach is conservative; use of Text Summarization to make text short

Innovation Aspects

4. Fostering a new interdisciplinary research area

Joining efforts with large group of students & senior researchers:

- to study written text comprehension problems
- to deliver in **two years** (December 2009)*
  - A modular architecture for TS
    - **Syntactical module**
    - Lexical module
    - Textual module
  - to be used in FACILITATE and SIMPLIFY systems
- to continue approaching TS for 3 years or more in order to properly address the bottlenecks of TS:
  - Lexical and Textual aspects of TS (we have 2 PhDs addressing these aspects)
  - Other text genre besides the informative one
  - Other users and systems
  - Readability scores for Portuguese (related to the 'reading age' of the text) better than those already available
- and publish consolidated results in selected journals

* As scholarships started in January 2008

First Year Results (1/2)

- Text Simplification
  1. BP Corpus Analysis of simple accounts in BP, supported by AIC tool
  2. Simplification Annotation Editor which will evolve to SIMPLIFY
  3. An Original-Simplified Parallel Corpora - 104 newspaper texts and their simplified versions
  4. the XCES annotation standard developed to register the simplification information
  5. Portal of Parallel Corpora to store and query the original or simplified texts
  6. Rule-based syntactic simplification system for poor literacy readers at rudimentary level (coming soon)

- Measures of readability
  1. Study of more than 60 Coh-Metrix measures*
  2. Resources and Tools' implementation & adaptation for BP

* http://cohmetrix.memphis.edu/cohmetrixpr/index.html

First Year Results (2/2)

- Summarization
  1. Pilot evaluation of summarization methods
- Tools
  1. Discourse Analysis
     - Detection and management of redundant and common sense information
     - New segmentation module and a version for web
     - New evaluation tool/method and a version for web
  2. Online version and distribution of the BP Wordnet (TeP 2.0)
- HCI studies
  1. Accessibility in web applications and
  2. Methods of evaluation in the context of textual comprehension

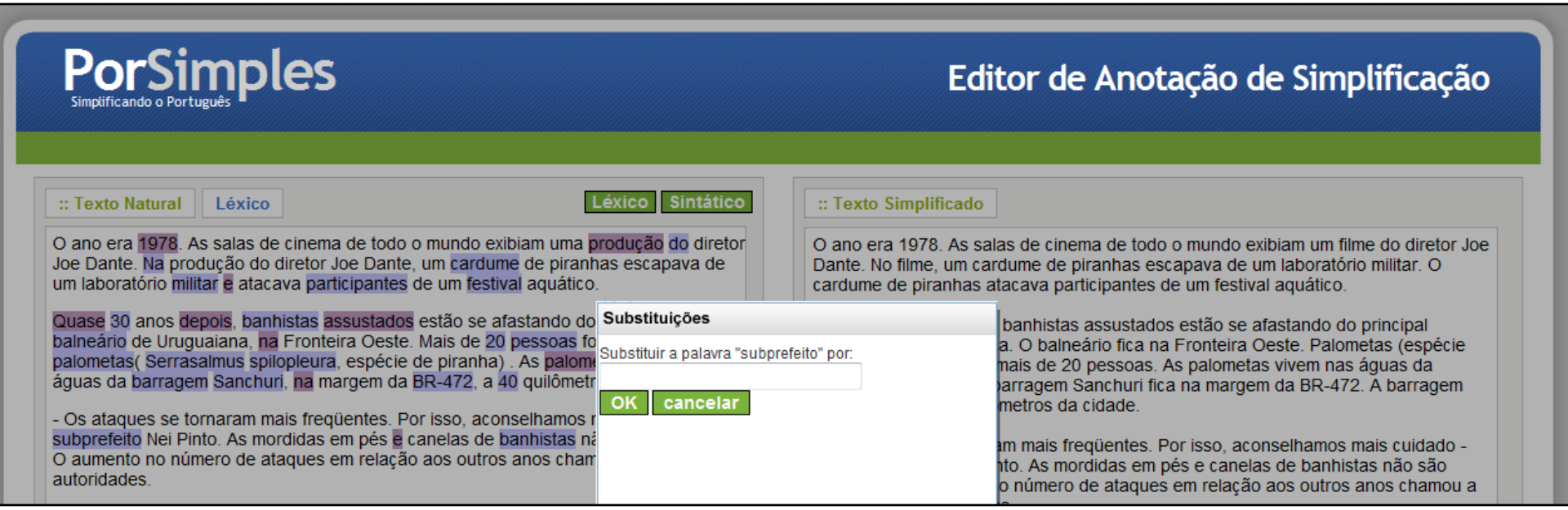
Validation of PorSimples Modules

- Tests with real users for the Summarization (preliminary results)
- Testing Simplification methods (*Facilitate*) with target users (coming soon)
- Tests with users (authors) for *Simplify (next April)*
- Tests with the *Make Explicit* module (the last one)

Methods & Tools for assessing text understanding are being developed

Screenshot of the Simplification Annotation Editor (in the *Sintático* mode): semi-automatic

Automatic Alignment



Word Substitution Interface

Text Summarization

- New and traditional text summarization systems:
  - were implemented (others were ready and available for use),
  - tested for text simplification purposes, and
  - are ready for use

    - Keywords method and 3 variations (Pereira et al., 2002)
    - Location method (Baxendale, 1958)
    - TextRank (a variation of Google PageRank) and a variation of it (Mihalcea and Tarau, 2004)
    - SuPor-2 (Leite and Rino, 2006)
    - GistSumm (Pardo et al., 2003)

Text Summarization: Pilot Experiment (1/2)

- Some possibilities for the reader:

1. Only the summary
2. Text with only the main sentence in bold
3. Text with all the important sentences in bold
4. Text with paragraph headlines
5. Text with highlights (CNN)
6. Text with irrelevant/redundant pieces of information removed (it is smaller than the original text, but it is too big to be a summary)

Text Summarization: Pilot Experiment (2/2)

- The comprehensive evaluation of 9 traditional and state-of-the-art summarization techniques
  - helped us to choose one for the experiment

- Strategies 1, 2 and 3 were evaluated with 20 people (cleaning staff) from ICMC-USP:
  - Up to 2 years in school
  - **4 and 5 years in school: our main focus in PorSimples**
  - 8 years in school
  - 2nd grade complete (more than 10 years in school)

- Preliminary results:
  - Up to 2 years of school: nothing helps (some people did not participate when they saw what the experiment was about; others gave up in the middle for being tired of reading)
  - **4 and 5 years of school: summary helps, but information in bold does not (hypothesis: it is one further information to process)**
  - 8 years of school: all the important sentences in bold help
  - 2nd grade complete: main sentence in bold helps

Text Summarization/ Discourse Analysis

- Detection and management of redundant and common sense information*

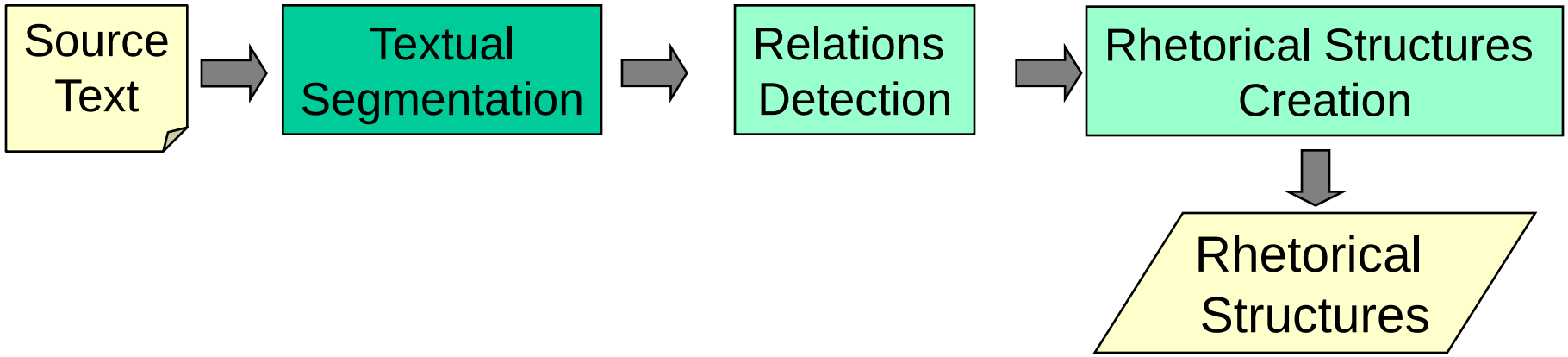
	Poor literacy reader - rudimentary level	**More advanced reader – basic level**
Redundancy	Should be kept	Might be removed with some specified rate
Common sense	Might be inserted	Should be removed

* ***Open Mind Common Sense*** *no Brasil* project

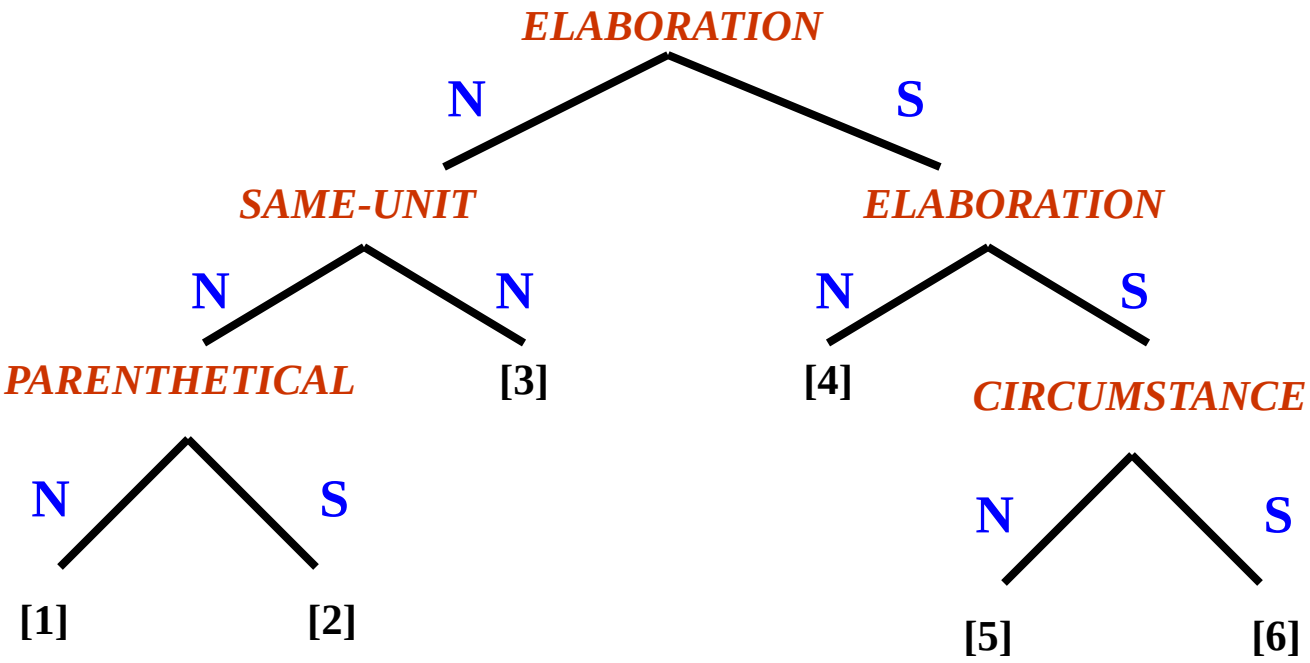
http://commonsense.media.mit.edu/

Discourse Analysis for *Make Explicit* Module

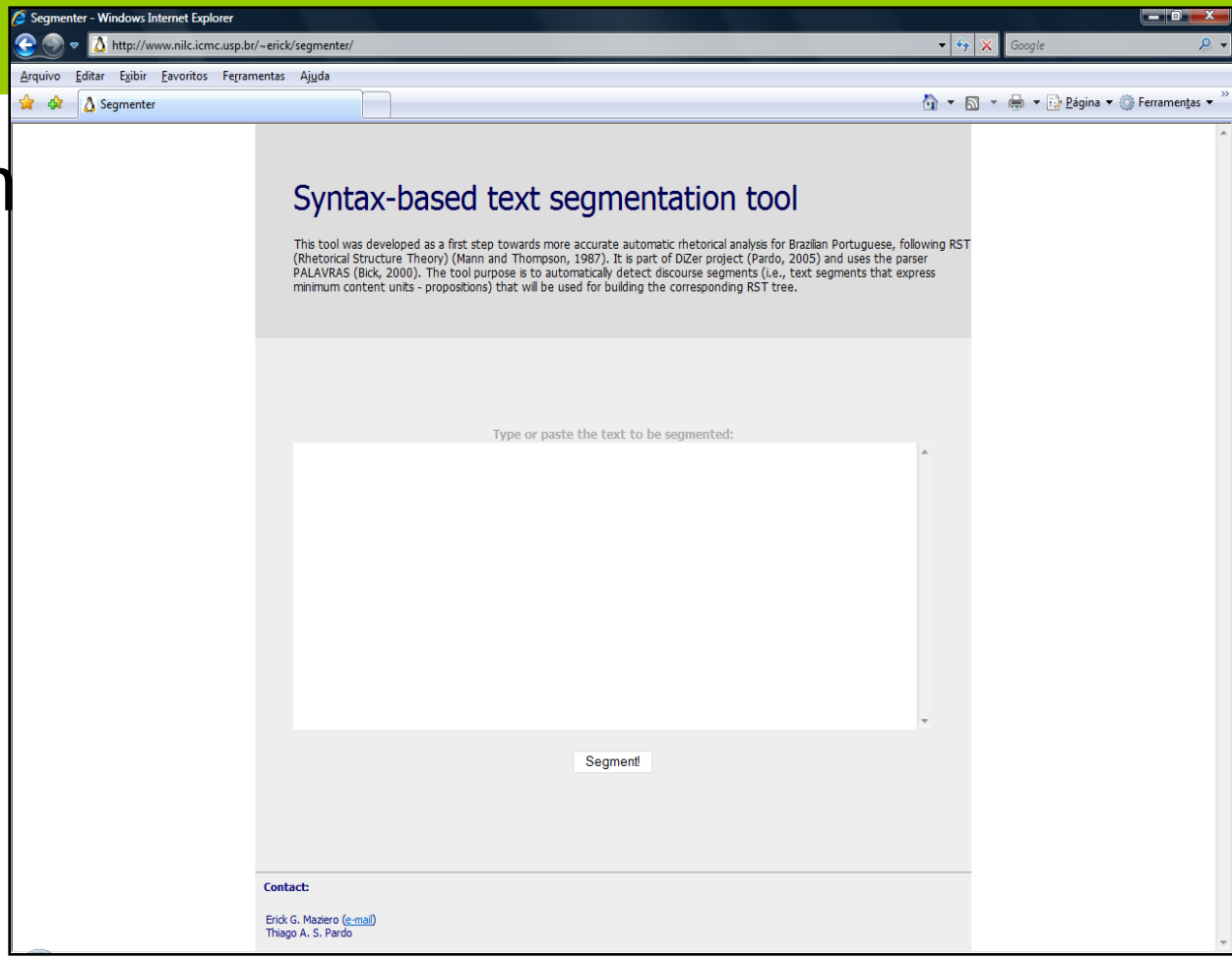
- DiZer (Pardo and Nunes, 2006) for scientific texts
  - Relationship between the parts of the text, following Rhetorical Structure Theory (Mann and Thompson, 1987)

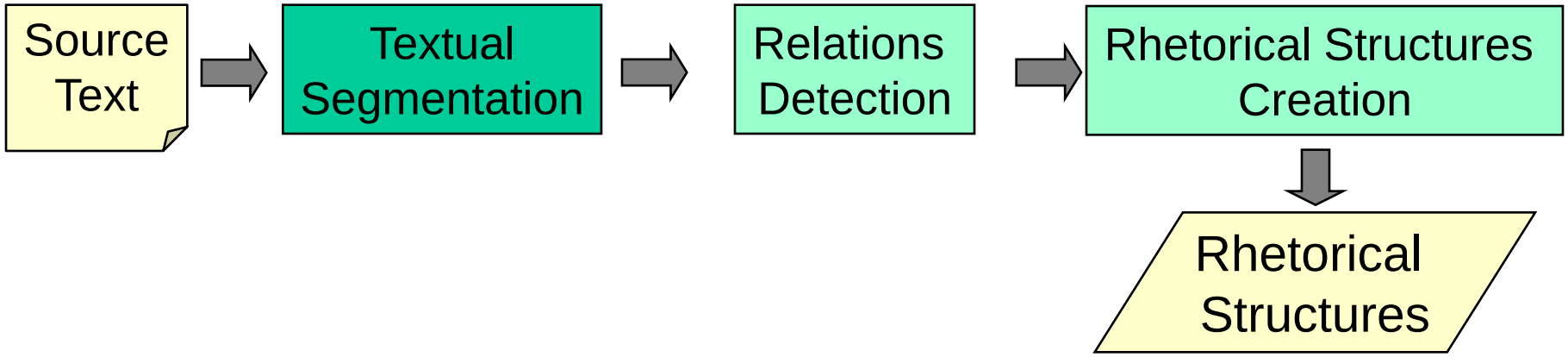


[1] A empresa Produtos Pirata Indústria e Comércio Ltda., de Contagem [2] (na região metropolitana de Belo Horizonte), [3] deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente. [4] Os ganhos são atribuídos pela diretoria da fábrica à nova filosofia [5] que vem sendo implantada na empresa desde outubro do ano passado, [6] quando a Pirata se iniciou no Programa Sebrae de Qualidade Total.

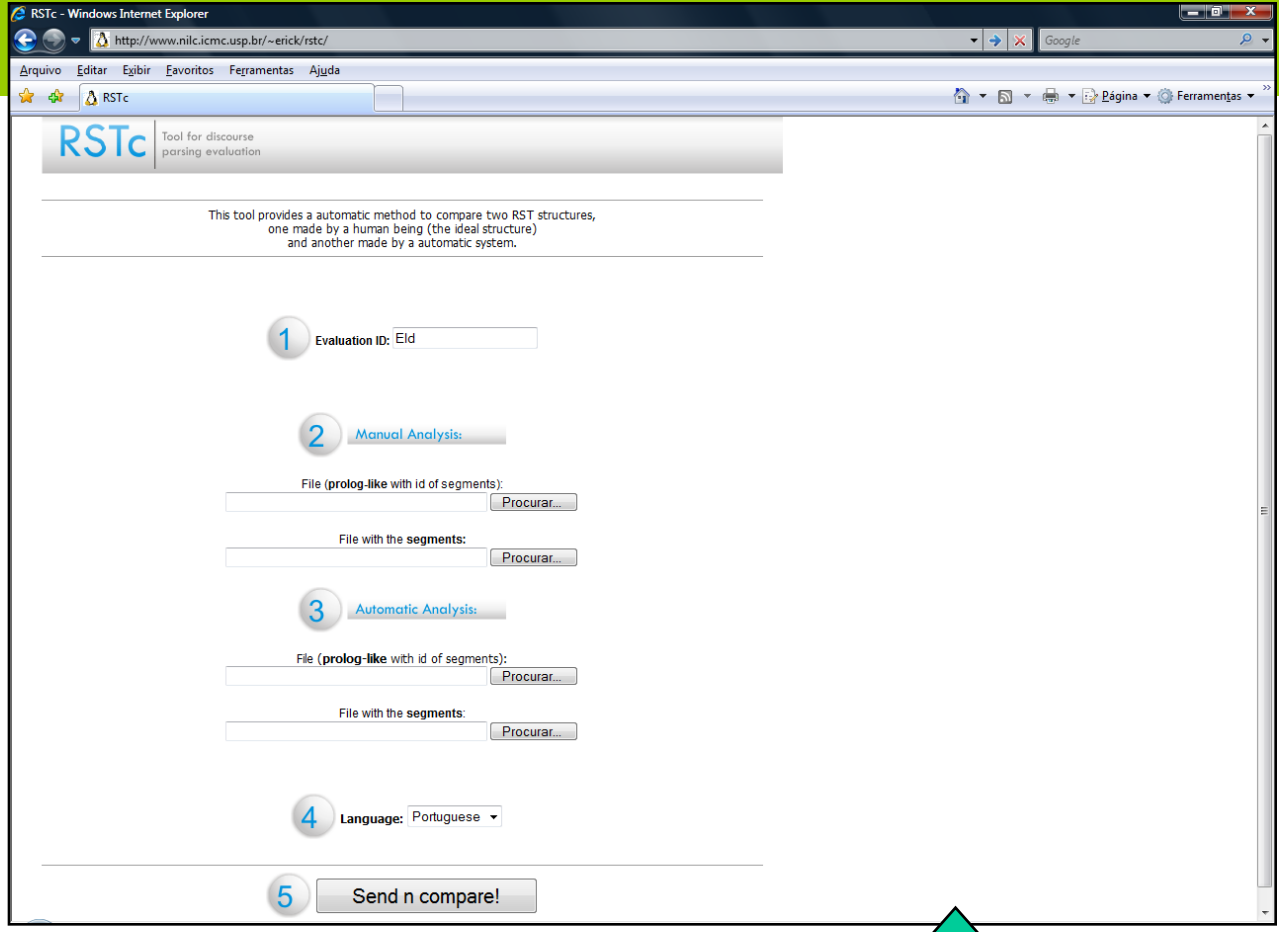


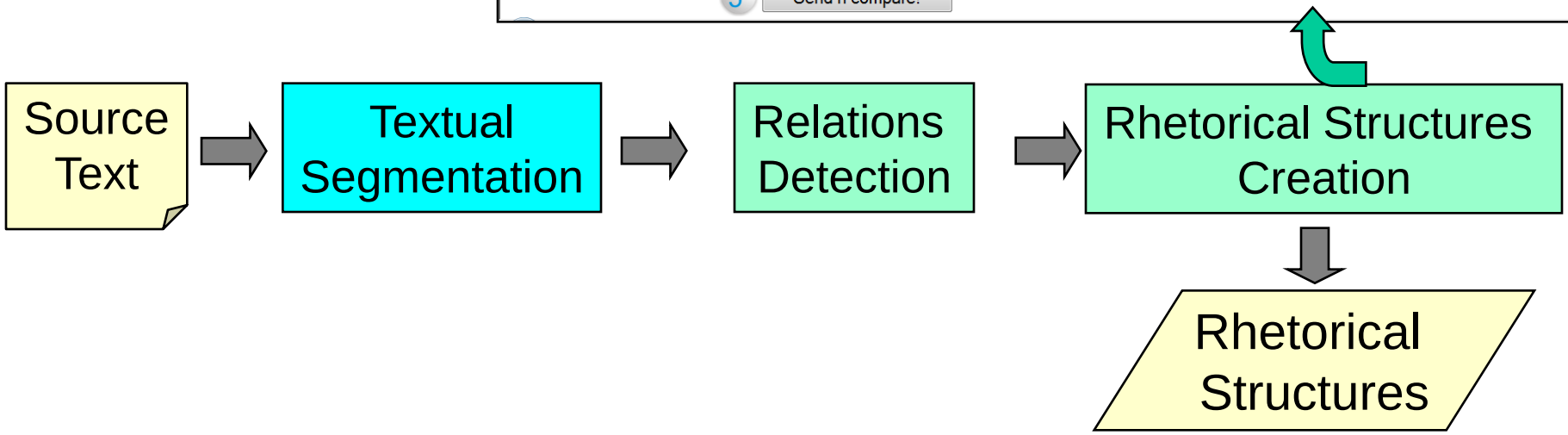
New segmentation module and a version for web





New evaluation tool/method and a version for web





Studies about **X+V technology**

- X+V is required to develop multimodal interfaces.
  - It is an XML standard useful to implement visual Web interfaces (XHTML) along with voice (by dialogs and speech recognition) interaction

- Voice interfaces (text-to-speech) can be considered in PorSimples tools for persons with low level of literacy

- Unfortunately, speech recognition and dialog systems are available only for English

Studies about accessibility in web applications, in the **Web 2.0 context and *Rich Internet Applications (RIA)***

- RIA is a concept that joins Widgets with AJAX (Asynchronous Javascript and XML) in web applications
  - Therefore, an algorithm was developed – using WAI (Web Accessibility Initiative) guidelines
    - to perform evaluation of Javascript code (for Widgets in the web) in a web application;
    - it was intended to help developers to locate in the code, the possible points of failure during validation of the guidelines.
  - These studies could show **advantages** and **disadvantages** in adopting AJAX components and *Widgets,* with multimodal interfaces.
- In conclusion, providing accessible and multimodal-based interfaces in the web is promising for future developments in PorSimples; this is an advance in state of-the-art in terms of Web accessibility.

Studies about traditional methods of evaluation in the context of textual comprehension

- Studies about the *Immediate Recall Protocol* method, especially proposed to evaluate textual comprehension

- **Qweb –** a web application to provide means to build questionnaires about textual comprehension
    1. It should be accessible
    2. It could include several types of questions
    3. It should help the author of questions to formulate them in an appropriate way
    4. It could include voice capture to get the answers;
    5. The two requirements above (3 and 4) are advances at the frontier of the area.

PorSimples in Numbers

- Team: 15 students
- Publications: 6 papers and several Technical Reports; a submitted paper to *CICLing 2009*
- Research Collaborators: 9 senior researchers from several areas:
  - Psycholinguistics
  - Statistics
  - Natural language processing and
  - Human language interaction

Our Team

- 6 students supported by MSR-Fapesp
  - 4 Undergrad, 1 MSc. and 1 PostDoc
- 2 PhD students:
  - Lexical simplification using Textual Entailment – to find patterns of substitution like *"X found a solution to Y" → "X solved Y"*
  - Anaphora Resolution to keep cohesive the simplified text
- 4 undergrad students:
  - Customizing TS to help children learning to read
    - 8-11 (using the corpus Para seu Filho Ler – Zero Hora newspaper and
    - 12-15 (using the corpus Ciência Hoje para Crianças developed by **Instituto Ciência Hoje** (ICH) of Brasileira para o Progresso da Ciência (SBPC))
  - Portal of Parallel Corpora of Simplified Texts
  - Experimental evaluation of a Portuguese Simplification System involving language impaired users
- 3 MSc. students:
  - Module Make Explicit of Facilitate System
  - Normalization of Technical Manuals using TS methods
  - TS method for sign languages like LIBRAS (Brazilian Sign Language) users

Papers Publications

- SANDRA ALUÍSIO, LUCIA SPECIA, THIAGO PARDO, ERICK MAZIERO, RENATA FORTES. "Towards Brazilian Portuguese Automatic Text Simplification Systems. " In the proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008), pp. 240-248.
- SANDRA ALUÍSIO, LUCIA SPECIA, THIAGO PARDO, ERICK MAZIERO, HELENA DE M. CASELI, RENATA FORTES. "A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems " In the proceedings of The 26th ACM Symposium on Design of Communication (SIGDOC 2008), pp. 15-22.
- PAULO MARGARIDO, THIAGO PARDO, GABRIEL ANTONIO, VINÍCIUS FUENTES, RACHEL AIRES, SANDRA ALUÍSIO, RENATA FORTES "Automatic Summarization for Text Simplification: Improving Text Comprehension by Functional Illiteracy Readers. " Publicado nos Proceedings Online do TIL 2008 (http://www.inf.ufes.br/webmedia2008/webmedia2008_wtil2008.html).
- PAULO R. A. MARGARIDO, THIAGO A. S. PARDO E SANDRA M. ALUÍSIO. Sumarização Automática para Simplificação de Textos: Experimentos e Lições Aprendidas " Publicado nos Proceedings em CD-ROM do IHC - UAI 2008. (http://www.inf.pucrs.br/ihc2008/pt-br/)
- MAZIERO, E.G., PARDO, T.A.S., DI FELIPPO, A.; DIAS DA SILVA, B. (2008). A Base de Dados Lexical e a Interface Web do TeP 2.0 – Thesaurus Eletrônico para o Português do Brasil. Publicado nos Proceedings Online do TIL 2008 (http://www.inf.ufes.br/webmedia2008/webmedia2008_wtil2008.html).
- FREIRE, A. P. ; RUSSO, C. M. ; FORTES, R. P. M. "A Survey on the Accessibility Awareness of People Involved in Web Development Projects in Brazil" accepted. In: 5th International Cross-Disciplinary Conference on Web Accessibility - W4A (http://www.w4a.info/2008/). April 21-22, 2008, Beijing, China. Co-Located with the 17th International World Wide Web Conference.

Collaborators

- Lucia Specia (Xerox Research Centre Europe)
- Maria da Graça Pimentel (ICMC-USP)
- Maria das Graças Volpe Nunes (ICMC-USP)
- Renata Fortes (ICMC-USP)
- Thiago Pardo (ICMC-USP)

- Maria Luiza Cunha Linha (UFMG, Departamento de Letras) since August 2008
- Helena Caseli (UFSCar, Departamento de Computação) since September 2008
- Milene Selbach Silveira (PUCRS, Faculdade de Informática), since September 2008
- Mariana Curi (ICMC-USP, Estatística), since November 2008)

References on TS

- Chandrasekar R., Doran C. and Srinivas, B.: Motivations and Methods for Text Simplification. COLING 1996, pp. 1041-1044. (1996)
- Chandrasekar, R., Srinivas, B.: Automatic induction of rules for text simplification. Knowledge-Based Systems, 10, 183–190. (1997)
- Devlin, S. and Unthank, G.: Helping aphasic people process online information. In the Proceedings of the ACM SIGACCESS 2006, Conference on Computers and Accessibility, pp. 225-226. (2006)
- Siddharthan, A. Syntactic Simplification and Text Cohesion. PhD Thesis. University of Cambridge (2003)
- Siddharthan, A. Syntactic simplification and text cohesion. Research on Language & Computation, 4(1):77-109, 2006.
- Williams, S.: Natural Language Generation (NLG) of discourse relations for different reading levels. PhD Thesis, University of Aberdeen. (2004)
- Petersen, S. E., Ostendorf, M.: Text Simplification for Language Learners: A Corpus Analysis. 2007. In Proceedings of the Speech and Language Technology for Education Workshop (Pennsylvania, USA, October 1-3, 2007). SLaTE-2007. Carnegie Mellon University and ISCA Archive, http://www.isca-speech.org/archive/slate_2007. 69-72.

Thanks!