

## Yes, user! compiling a corpus according to what the user wants

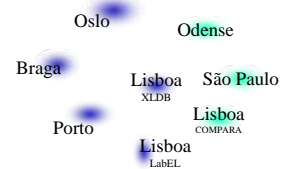
Rachel Aires  
Diana Santos  
Sandra Aluísio  
Linguateca & USP

## Linguateca, a project for Portuguese

- A distributed resource center for Portuguese language technology
- First node at SINTEF ICT, Oslo, started in 2000 (work at SINTEF started 1998 as the *Computational Processing of Portuguese* project)

### IRE model

- Information
  - Resources
  - Evaluation
- [www.linguateca.pt](http://www.linguateca.pt)



## Structure of the presentation

- The corpus at a glance
- The general motivation for the study that led to its creation
- The hypotheses of the study
- Empirical assessment of the hypotheses
- User-based assessment of the hypotheses
- Further work (in progress)
- Related studies

## Yes, user!

- A corpus of 1,703 Brazilian Web pages classified according to **users'** goals in seven different (but not necessarily mutually exclusive) kinds
- Publicly available from [www.linguateca.pt/Repositorio/yesuser.html](http://www.linguateca.pt/Repositorio/yesuser.html)
- **Plus:** Several domain specific binary corpora in Portuguese of 200 Web pages each (100 positive, 100 negative), developed independently by several users (native speakers of Portuguese): Y1, Y2, Y3, ...
  - Japanese cooking; fado; military aviation; law; philosophy of language; fishing; surrealism; history; etc.

## Problem setting: NLP in IR

- Or more specifically: Portuguese processing for IR in Portuguese...
- The general question is: Does it help?
- Of the many questions/subproblems that could be chosen, we (Rachel's PhD studies) concentrated on the specific one:
- Can NLP improve the **presentation** of the results to the user of a Web search engine?
- (In the usual IR tripartite model, there are three phases in IR to which NLP could be added: search, indexing and presentation). The last one has been the less studied as to NLP contribution.

## Problem setting: shortcomings of WebIR

- The problem is no longer to find results, but to navigate among their massive number in a meaningful way
- Web users are extremely varied and use the Web for all kinds of purposes and activities, so standard user testing in a usability lab with a sample of the population is obviously out of the question
- These two factors have led researchers to
  - investigate other properties of Web documents other than their content (such as style, language quality, text genre, authority, novelty, maintenance/update rate)
  - offer personalization solutions for specific kinds of search, for specific kinds of users, for specific kinds of content, as well as try to study the users and their behaviour using inobtrusive techniques
- The work reported here is in this line...

## The main hypotheses

- It is possible to distinguish among different users' needs
- In general, different pages satisfy different users' needs
- It is possible in a majority of cases to evaluate whether a particular page or site was designed to satisfy (or not) a given user's need
- Users are aware of their needs (if asked to distinguish among several)
- It is possible to automatically classify pages according to user's needs
- Users have higher satisfaction when getting results' pages clustered around user's needs

Technical/philosophical      User-related

## The technical hypotheses

- The three first have received some confirmation by the compilation of the Yes, user! corpus
  - a typology of seven user's needs was devised after careful qualitative consideration of the logs of a real Web search engine (todobr)
  - 4 different people looked at Web pages and classified them according to a set of precise guidelines with examples:
    - although there was considerable overlap (many pages satisfied more than one need, no pages satisfied at once every need)
- The fifth hypothesis was put to test by developing automatic classifiers for texts satisfying each user need and testing them with ten cross-fold validation methodology in this corpus
  - promising preliminary results in Aires et al. (2004) with a smaller corpus
  - precision > 70% in the present corpus (Aires et al., SIGIR style workshop 2005)

## The user-related hypotheses

- Are being tested in two different ways:
- Questionnaire to prospective users:
  - Devising and delivering it to graduate students
  - Analysing the 63 answers
- User testing a prototype of a desktop metasearcher, *Leva e Traz*, to see whether users understand the concepts and whether their satisfaction with the search improves
  - User-testing is in progress at the moment
  - The prototype is working

Before we show a screenshot of it, though, we have to say a little more:

## Other questions we wanted to investigate

- Would traditional genres (as used e.g. by Karlgren or Stamatatos et al.) be more understandable to users?
- Or just kinds of texts, in the sense of traditional names like "novel", "biography", "scientific paper", "textbook", etc., would be more in agreement with what the user wants?

*Aluísio et al. (2003) had developed a genre typology, which included also "text types", for Brazilian Portuguese Web (texts) for use in the Lácio-Web corpus and we had also been able to develop automatic classifiers for them*
- Would users spend half a day creating binary corpora, in order to improve their search results in the future? Or were they satisfied with current Web search engine's performance? *So we just asked users*

## Leva-e-traz, desktop metasearcher prototype



## Related work

- We were inspired by Karlgren's (2000) studies, and this work can be seen as a validation and confirmation of his original hypotheses, for Portuguese
- Other people have investigated automatic genre classification for other languages, as Stamatatos et al. (2000) for Modern Greek
- We were inspired by Biber's (1988) empirical investigation of co-occurrence patterns among linguistic features, and, in fact, most of our features were directly inspired by (adapted from) those used by Biber
- There is a huge effort in personalization and user adaptation. Our work in binary corpora can be related to the adaptive IE system of Ciravegna & Wilks (2003)
- Finally, previous work in detecting users goals (Aires & Aluísio, 2002), in proposing evaluation contests to characterize the Web in Portuguese (Aires et al., 2003), and measuring its size (Aires & Santos, 2002) is also relevant

## Further work

- We have parsed Yes,user! with the PALAVRAS parser (Bick, 2000), in order to investigate other relevant features for the classification in seven user needs.
- Although we do not expect in the near future to have a parser quick enough to do this in real time, this is a relevant question to investigate
- If our seven needs classification proves useful, we may add it later to the indexing phase of a search engine, and in that case one might consider the possibility of parsing the contents of the pages before indexing
- We also consider the hypothesis of adding further structural clues (such as provided by HTML and presence or absence of images) to our classifiers in order to improve the classification precision

## Concluding remarks

- It is feasible to classify Web pages according to the kinds of goals users have, as well as according to genre or type of text dimensions
- We have created **Yes, user!**, a corpus in Brazilian Portuguese classified according to these criteria that we make available for other researchers to use and improve
- We have also created a set of small corpora, also available, in order to test adaptation to user's specific interests
- We are currently testing whether these kinds of classification can be put to advantage in real user's life by using a metasearcher which provides the results classifying according to these parameters and do user testing