A Gramateca

seus objetivos e alguns exemplos

Diana Santos

d.s.m.santos@ilos.uio.no

10 de outubro de 2014

O que é a/uma gramática?

Como todas as palavras de uma língua, a sua definição não é simples. E, como em todos os ramos do conhecimento, há diversas perspetivas e opiniões e delimitações. É mais fácil saber usar do que saber definir. E o que é uma definição?
Para os efeitos da Gramateca, *gramática* é a forma como a língua funciona. Por isso pode incluir questões por outros chamadas pragmáticas, estudos de género textual, ou culturais.
E como é que se pode estudar e compreender a língua?
Também a este respeito, há opiniões e posições variadas. A resposta da Gramateca é

- através do estudo de grandes quantidades de texto
- através da interpretação dos casos de comunicação em contexto

O que é uma língua?

- Certo, já foi explicado que era um dialeto com um exército – mas de facto não são só metáforas guerreiras que explicam o mundo, e uma partilha cultural através da língua é uma riqueza que não tem preço.
- Por isso é que eu acho que a questão do espaço lusófono é muito mais do que (ou não é sequer) um saudosismo colonial dos portugueses, mas sim uma globalidade a nível da Terra que permite a comunicação e a compreensão de milhões e milhões de pessoas.
- E que existe algo chamado **língua portuguesa** que é maior do que o português do Brasil, ou o português de Portugal, ou o português de Moçambique, por exemplo. E que nesse encadear de ideias, deveríamos todos trabalhar para um português como língua internacional.

O que é a Gramateca

Um laboratório para o estudo da língua portuguesa, pondo ao dispor dos interessados

- todos os corpos que conseguimos
- anotação automática desses corpos
- anotação manual de subconjuntos dos mesmos
- ferramentas de visualização e de exploração dos corpos
- uma plataforma de revisão e de comparação de diferentes análises

Qual a matéria prima para a escrita de gramática?

Vários ingredientes

- A introspeção do gramático, o conhecimento e a reflexão sobre a sua própria língua (ou em casos raros, outra) é, evidentemente, essencial. O que se chama "sensibilidade linguística"... pode ter-se "ouvido" para várias coisas diferentes.
- A possibilidade de consultar outros falantes/comunicantes
- A possibilidade de consultar a população / subpopulações, com base na agregação de muitos falantes independentes
- O cotejo com outros gramáticos ou com outras intuições ou interpretações

A questão inicial...

A questão inicial que despoletou a Gramateca, foi:

- é possível partilhar os materiais sobre os quais se constrói a gramática?
- de forma a validar, repensar, repetir ("replicar") os raciocínios?

Isto porque – dos estudos contrastivos – foi claro que muitas das conclusões dependiam de formas de preparar e contar o material, e que muitas das contagens poderiam depender de factores que os pesquisadores iniciais não tinha considerado.

Um exemplo concreto

Em Santos e Oksefjell (1999) tentámos comparar, com novos dados, dois estudos empíricos

- A tradução do movimento em Slobin (1996,1997)
- A tradução de verbos de perceção em Santos (1998)

E logo nos apercebemos de que seria muito mais prático podermos olhar para o material/julgamentos apresentados, também. Por exemplo

- Slobin estava a referir-se a quaisquer movimentos, ou só movimentos concretos?
- A perceção de movimentos varia conforme a parte de um livro? Ou dependendo do verbo específico?

Outro tipo de afirmações comuns em textos linguísticos

Henriqueta Costa Campos (1984:539):

> *il faut partir d'une analyse systématique et, dans la mesure du possible, exhaustive, de sa valeur la plus fréquente (je ne puise cette affirmation dans aucune étude statistique qui, à ma connaissance, n'éxiste pas. D'ailleurs, une telle étude impliquerait une typologie préalable.) (...)*

Ora a Gramateca permitiria (ou permitirá) verificar qual o valor mais frequente, desde que se comece a marcar valores de *já* numa amostra.

As gramáticas que temos

Existem muitas gramáticas belíssimas e muitíssimo interessantes e bem documentadas que apareceram nos últimos anos:

- Castilho, Ataliba. *Nova gramática do português brasileiro*. Editora Contexto, 2010
- Raposo, Eduardo Buzaglo Paiva et al. *Gramática do português*, 1.o volume, 2013
- Ilari, Rodolfo (org.). *Palavras de classe aberta*. Vol 3. da Gramática do Português Culto Falado no Brasil, Editora Contexto, 2014

Todas elas são – pelo menos – inspiradas em corpos, e abonadas com exemplos variados deles retirados.

Mesmo assim

- Mas, mesmo assim, falta totalmente o elemento quantitativo.
- Assim como a possibilidade de observar mais casos específicos ou problematizar alguns elementos ou análises apresentadas.
- Embora, e isto seja de louvar, existam e estejam acessíveis os corpos mencionados e usados.

O exemplo de Biber e Leech e Johansson etc

Uma gramática para o inglês que serviu de inspiração para este projeto foi Biber et al. (1999).

- Essa gramática foi construída com base num corpo analisado.
- Além de exemplos, apresenta sempre dados quantitativos.

Problema: o corpo e a sua análise não foram nunca tornados públicos.

O exemplo de Sampson

- Ao contrário, Geoffrey Sampson escreveu um livro chamado *English for the computer* (1995) em que descreveu muitos factos interessantes da gramática do inglês, com base num corpo analisado e estudado, o SUSANNE.
- É interessante que ele mais tarde se veio a queixar que as pessoas usaram e citarem muito o SUSANNE, sem fazer caso nem ligar à gramática (livro) que ele tinha escrito.
- De certa forma isto também acontece com a Floresta e com o trabalho de Eckhard Bick, neste caso corporizado no PALAVRAS (Bick, 2000): a maior parte das pessoas usa-o sem refletir ou ligar às generalizações linguísticas ou opiniões gramaticais do autor.

A questão da acessibilidade e da replicação

Mas, mesmo quando os autores se empenham em tornar os seus dados e análises públicos, isso não é tão fácil assim. Digo-o por experiência própria, embora as condições sejam muitíssimo melhores agora que dantes

- Por causa dos direitos de autor
- Por não haver facilidade em disponibilizar conjuntos de dados
- Por não haver forma de os reavaliar e reescrever de uma forma comparável
- Porque novas versões podem invalidar ou confundir os dados anteriores

E para isso concebemos o Rêve, que ainda está em desenvolvimento.

Passando de nome contável para massivo

- Desde a génese e durante a incubação deste projeto, fomos passando da ideia de "mais uma" gramática para uma "forma de fazer gramática", por um lado pulverizada como areia (ou notas de rodapé), por outro como substantivo abstrato.
- Parece-nos agora que a Gramateca é mais uma infra-estrutura para fazer gramática(s) ou estudos de gramática do que algo que produza mais uma nova gramática.

O poder às massas? Não.

- É importante separar esta forma de pensar da filosofia do "crowdsourcing", que usa muita gente com poucos conhecimentos para obter consenso ou informação. Esse tipo de abordagens é interessante para algumas áreas, mas **não** é o que se tenta fazer aqui.
- Quem pretendemos congregar na Gramateca são as pessoas/peritos interessados na gramática e na cultura associada à língua portuguesa, não qualquer um. Pessoas que queiram reanalisar e reinterpretar análises linguísticas, porque têm uma opinião diferente ou porque querem confirmar uma dada análise, ou ir mais além.
- Muitas vezes, verificar a opinião da maioria dos falantes (informantes) é exatamente aquilo que um linguista precisa. Mas fazer perguntas complexas a amigos e conhecidos, e daí concluir sobre algo teórico, parece-me perigoso.

Um exemplo: causalidade, uma das noções mais complexas na filosofia e na ciência. Alguns colegas meus criaram um sítio na rede para anotar a causalidade entre duas orações em português.

Gramática infinita

Depende das pessoas que queiram usar a infraestrutura, e que queiram partilhar, e ou comparar/discutir as análises já feitas.
Permitir uma análise multivariada, o que significa que os investigadores podem trazer mais variáveis para a análise.

> *Muito bem, o senhor X mostrou que o fenómeno A era condicionado/dependente das variáveis B e C, mas eu quero mostrar que a variável D também é importante – ou que a característica E explica as ocorrências muito melhor...*

Alguns problemas metodológicos na linguística

- Existe em geral – comparado com outras áreas do conhecimento – muito menos preocupação em apresentar dados.
- Muitos artigos discutem opiniões ou teorias de outros linguistas, e apresentam um ou outro contra-exemplo, mas raramente – na maior parte das vezs por falta de meios práticos – dão alguma opinião baseada na prática da língua por outros.
- Outro problema (na minha opinão grave) é que muitas vezes os dados são referentes a populações diferentes, como por exemplo os linguistas que argumentam com dados de "todas as línguas" em vez de uma apenas.

Eu não pretendo dizer que a metodologia da linguística deve ser a mesma que a das outras áreas do conhecimento. Mas também acho que a prática pode ser (muito) melhorada.

Um dos problemas do estudo da língua

Provavelmente, este é um problema que assola muitas outras áreas, devido à extrema especialização académica. Mas eu vou ilustrá-lo em relação à linguística, ou estudo da linguagem natural.

- muitas das afirmações feitas baseiam-se em pressupostos teóricos antagónicos
- factos são dependentes das teorias
- quando um estudo tenta usar técnicas da sociologia, ou da estatística, ou da computação, não há suficientes parceristas qualificados, donde a maior parte dos artigos publicados não têm a qualidade necessária
- e provavelmente alguns artigos de qualidade não conseguem ser publicados

Isto não tem nada a ver com o português, é um problema geral, como foi por exemplo descrito em Sproat, Richard. "Last Words: Ancient Symbols, Computational Linguistics, and the Reviewing Practices of the General Science Journals", *Computational Linguistics* 36, 3, September 2010, pp. 587-94.

A Gramateca tenta ser um caminho

- A Gramateca, ao congregar pessoas de diferentes áreas científicas (sobretudo computação e linguística) e teorias (ou seja, qualquer cientista/linguista/estudioso da cultura e/ou das Letras que queira usar o material é bem vindo), tenta evitar a infeliz separação e compartimentação em áreas diferentes e bem guardadas, nas quais é preciso citar as pessoas certas e não desviar demasiado das ideias "católicas".
- Mas claro está que áreas da linguística que exigem: acesso ao cérebro dos falantes, acesso a todas as línguas (ou muitas), acesso aos dados sociológicos dos falantes, acesso ao sinal de fala, não podem fazer muito uso da Gramateca – a não ser que se restrinjam aos subconjuntos que tenham informação necessária, ou que façam materiais com base no que já existe.

O exemplo do Museu da Pessoa

Este é um caso paradigmático, porque foi um corpo que não foi constituído por razões linguísticas, mas que dado estar acessível e corresponder a material oral – o qual é sempre mais difícil de obter e tratar – foi desde cedo incorporado no AC/DC.
Contudo, cedo descobrimos que as transcrições (da parte portuguesa) estavam cheias de erros de ortografia e imprecisões, e que além disso poderia ser interessante – até porque um artigo foi recusado por falta de informação demográfica sobre os falantes – associar informação sobre cada entrevistado.
Isso levou a uma tarefa que tem sido extremamente interessante (mas que trabalha apenas 1% do material...) que é a revisão (e subsequente anotação e quantificação) de "desvios do oral", ou melhor, da variabilidade linguística que é possível apreciar em falantes idosos com baixo grau de escolarização, oriundos principalmente do Norte de Portugal.
(A propósito, gostaríamos que fosse efetuada uma tarefa semelhante nas entrevistas brasileiras também...)

O exemplo do Museu da Pessoa

Com esse estudo, que está ainda em progresso, já conseguimos identificar muitas áreas de variabilidade de que não estávamos sequer conscientes, e que – o que me parece ainda mais interessante – são casos cuja variabilidade está amplamente estudada no português brasileiro mas que não costuma ser identificados no português de Portugal. E alguns casos, claro, já conhecidos por exemplo através de Peres & Móia (1995).

- Variabilidade na concordância em número
- Problemas com orações relativas e com construções de identidade
- Reflexividade e a posição dos clíticos

Uma descrição resumida pode ser encontrada aqui: Taveira, Paula & Diana Santos. "Revisão da parte portuguesa do Museu da Pessoa", `http://www.linguateca.pt/acesso/revisao_mp.html`

O exemplo do Museu da Pessoa como base de estudos quantitativos

O corpo do Museu da Pessoa (devido à sua acessibilidade) tem sido usado em vários estudos, por exemplo para comparar o PB com o PP.
Mas isso demonstra outro problema da Gramateca (e/ou do AC/DC):

- Mas será aceitável comparar velhinhos do Norte de Portugal com jovens do Rio? E entrevistadores jovens da Universidade do Minho com brasileiros adultos com outra formação? (que eu neste momento ignoro qual...)
- Outras entrevistas, feitas por sociolinguistas e com a melhor das intenções, podem ter problemas semelhantes. Estou-me a referir aos materiais compilados no âmbito do projeto Concordancia (`http://www.concordancia.letras.ufrj.br/`), em que reparei que algumas entrevistas feitas na Madeira foram feitas por brasileiros a madeirenses.

Português internacional

Assim como existe uma vasto movimento em torno de "English as a lingua franca" (uma variedade do inglês falada e escrita maioritariamente por não nativos), também me parece importante investirmos num Π, português internacional, para dotar todos os que aprendem a nossa língua sem ser como língua materna ou paterna de um instrumento robusto e reconhecido para a comunicação internacional, que lhes permita comunicar corretamente com portugueses, brasileiros, africanos, oceânicos e asiáticos que usam o português como língua oficial ou de trabalho.
Uma das formas de o fazer, seria exatamente estudar a forma de interação entre os variados grupos de falantes nativos em português, por exemplo como os portugueses adaptam a sua forma de falar e escrever quando em presença de brasileiros, e vice-versa; ver
`http://www.linguateca.pt/Diana/download/VariantesPIGSCP.pdf`
para mais pormenores.

Ciências ou letras: Uma falsa questão

Tão falsa que praticamente todos os grandes pensadores pensaram as duas (ou três, ou quatro, se quisermos incluir a filosofia e a teologia como áreas separadas).
Começando com Pitágoras e Aristóteles, passando por D. Duarte, Damião de Góis, Descartes, Leibniz, Espinosa, Leonardo da Vinci, Abraão Zacuto, Newton, Pedro Nunes, Euler, Bayes, Freud, Yule, Zipf, Egas Moniz, Wittgenstein, Hofstader...
É por isso que agora vou falar de emoções e do corpo humano...

A questão das emoções: três eixos

A questão das emoções e da avaliação

- A primeira função da linguagem (ver Ellis (1993))
- Pode ser vista segundo três eixos: o facto, o locutor e o ouvinte.

– *Ele ameaçou-me!* pode referir-se a zanga/raiva de "ele", pode exprimir o meu "divertimento", e pode inspirar medo a outros que me ouvem.

A questão das emoções

- Nada mais subjetivo que a distinção entre subjetivo e objetivo.
- Contudo, a linguística preocupou-se no seu afã científico excessivamente com a objetividade, com os factos, com a informação, com as condições de verdade (!!) e muito menos com a subjetividade.
- Essa foi deixada para a psicologia, e até a psicolinguística tem olhado pouco para ela.

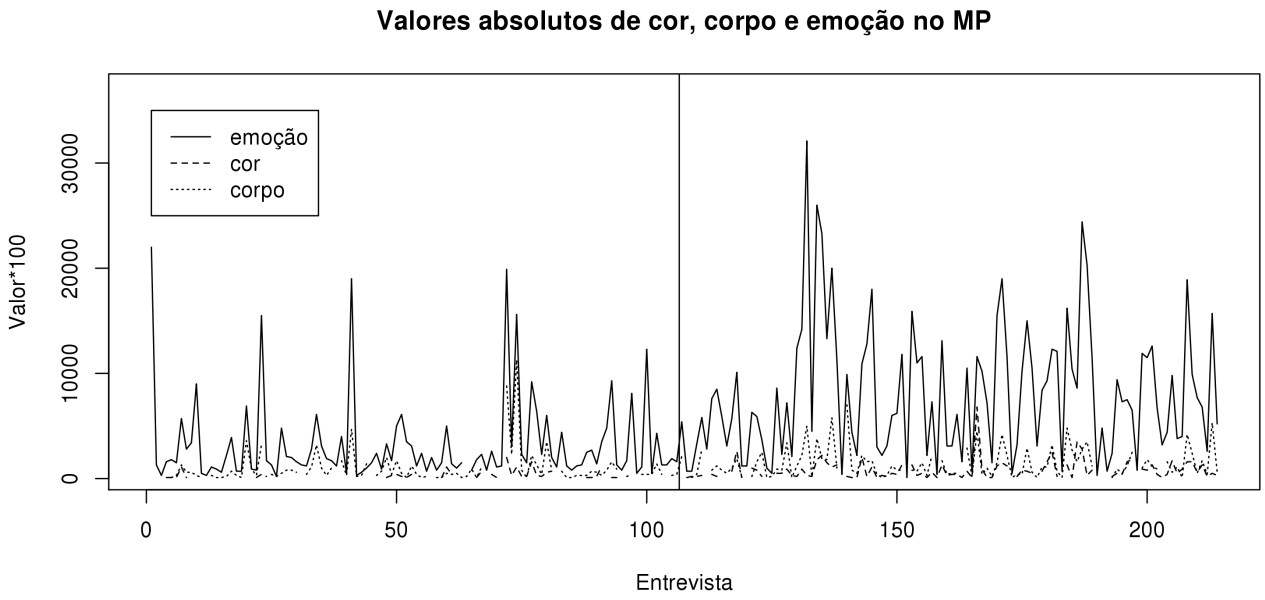
Contudo

Contudo, a questão das emoções é fulcral para compreender o que se diz e o que não se diz, e também aquilo que se pode dizer.

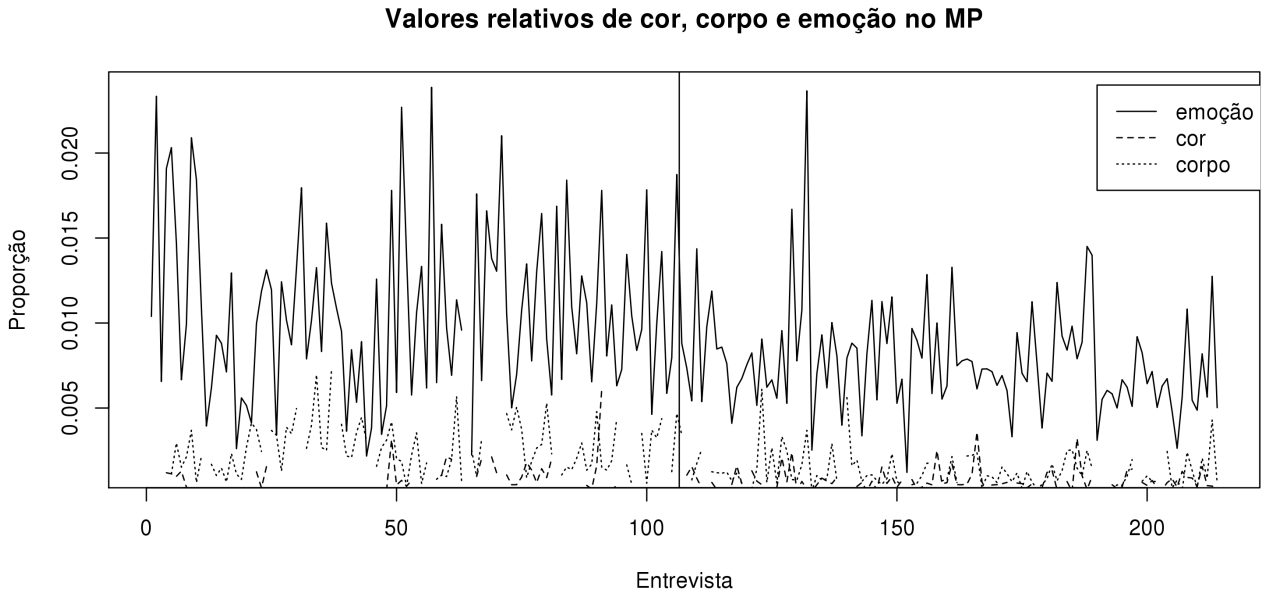
- Que emoções existem na língua portuguesa?
- Que palavras descrevendo emoção existem na língua portuguesa?
- Que formas de exprimir emoção existem na língua portuguesa? (oral e escrita)

Como distinguir emoção de opinião?

Opinião e emoção no Museu da Pessoa I



Opinião e emoção no Museu da Pessoa II



Opinião e emoção

- Uma das primeiras dificuldades é separar emoção de opinião, sensação de postura ou posição, avaliação de descrição.
  (isso é evidente nos nomes que pululam nesta área em inglês: *sentiment analysis, opinion mining, emotional language*...)
- Em questão de estrutura, o português é muito rico

  Quero ir a um bom restaurante Quero ir a um restaurante bom
  O pobre pai não sabia para onde se virar. Ela teve um pai pobre, mas extremoso.
  Meu rico filho! Pobres meninos ricos, que não sabem o que é a vida.

Opinião e emoção

Imaginem que estaríamos interessados em estudar a opinião e emoção mais de perto...

- Esta frase indica uma opinião?
- Se sim, qual a opinião, e sobre o quê?
- E a opinião é positiva ou negativa?

E qual a forma da opinião? Através de uma comparação, ou de uma metáfora? Ou ironia?

Opinião e emoção: o ReLi

Um trabalho muito interessante feito aqui na PUC, e que produziu o ReLi (Freitas et al., 2012, 2014).
Descrição do ReLi: cerca de 1600 recensões a 14 livros, escritos por 7 autores, publicadas na internet por alunos brasileiros, e compiladas e anotadas na PUC em relação a:

- objeto da opinião
- polaridade da opinião
- polaridade total de uma frase

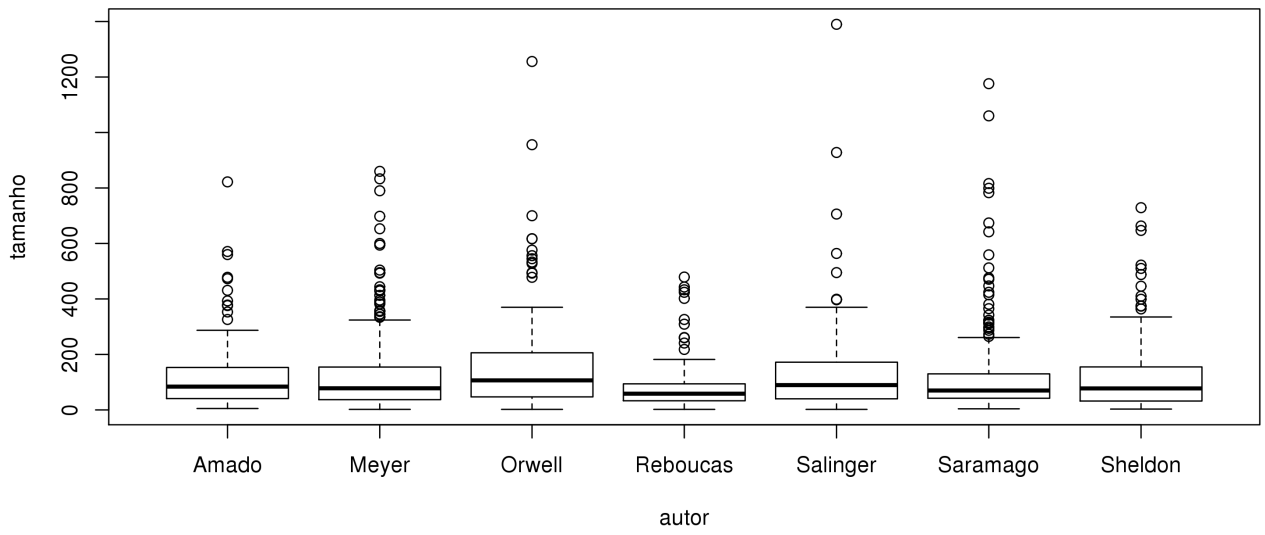
Eu vou tentar mostrar que mesmo com um trabalho tão interessante e tornado acessível para todos, é possível ter vantagem em adicionar outras variáveis. Em particular, a emoção.

Opinião e emoção: revisitando o ReLi

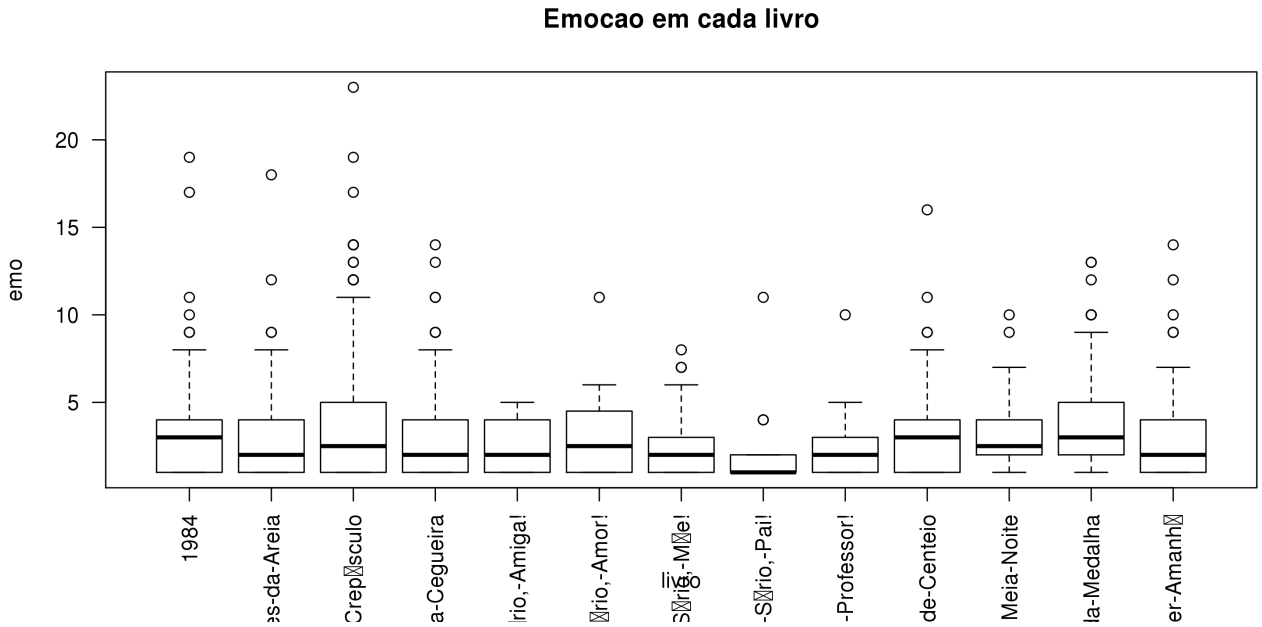
Devido a problemas de compatibilização entre os dois anotadores sintáticos utilizados – por causa de uma linguagem muito pouco padronizada e cheia de emoticões – ainda só conseguimos integrar/compatibilizar cerca de 1300 resenhas.
Relação entre emoções e polaridade:

- Dos 2490 casos considerados como pertencendo à linguagem das emoções,
- apenas 831 (33%) se encontravam em frases com polaridade positiva,
- e 109 (4%) com polaridade negativa.

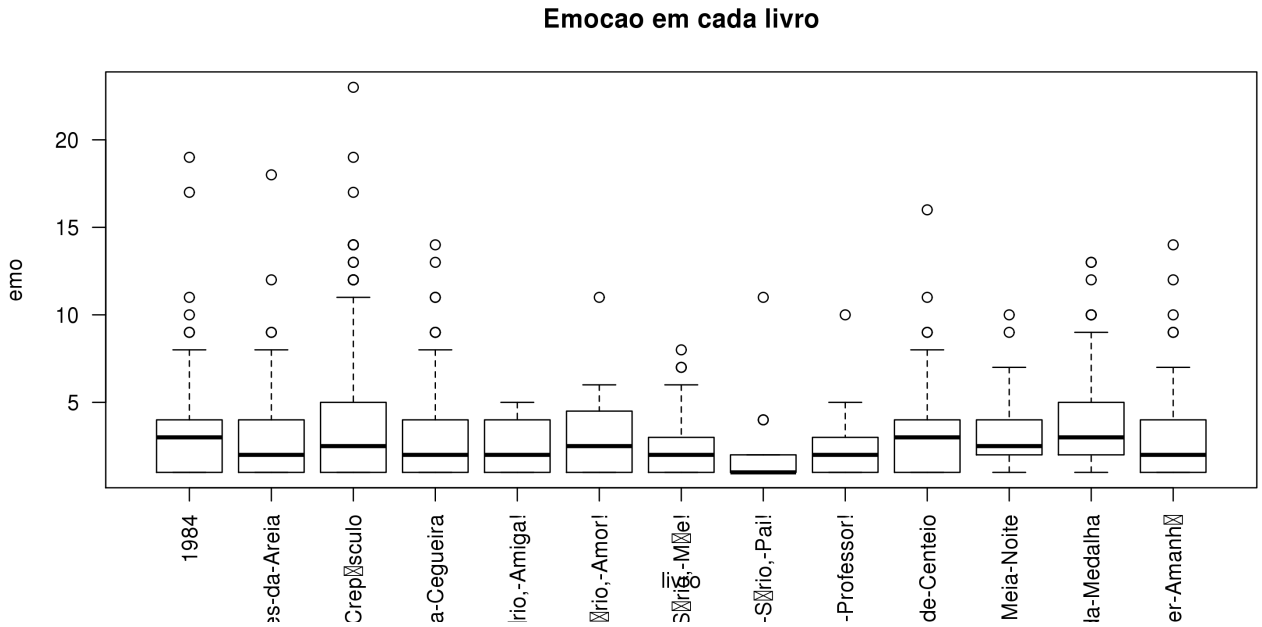
OIhando para o ReLi com as ferramentas da Gramateca I



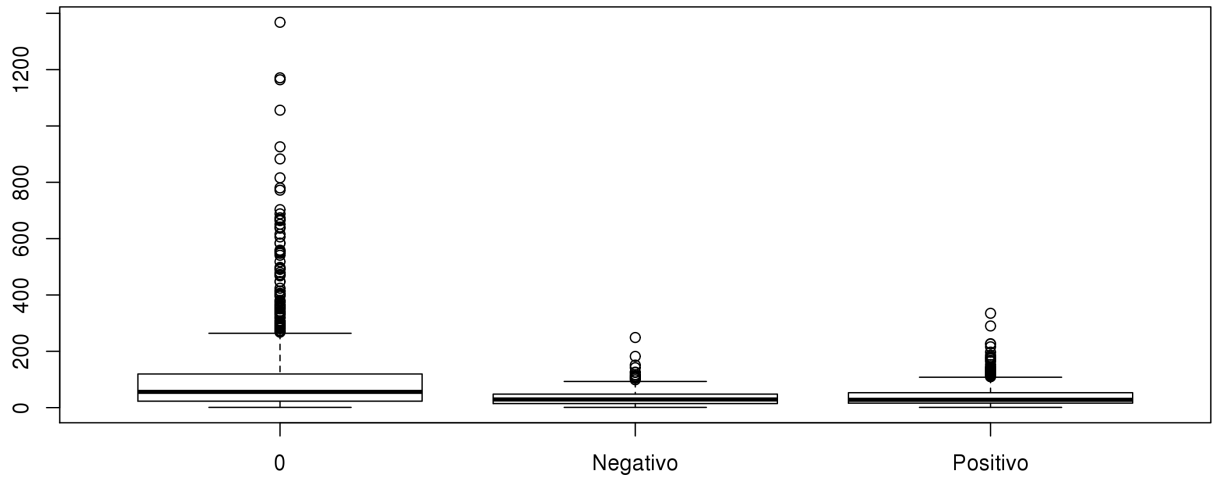
Olhando para o ReLi com as ferramentas da Gramateca II



Olhando para o ReLi com as ferramentas da Gramateca III



OIhando para o ReLi com as ferramentas da Gramateca IV



OIhando para todos os corpos da Gramateca

Para dar alguns aperitivos ao que um pesquisador com interesse pela língua e sem medo da quantidade pode fazer:

- a evolução de um dado campo semântico através dos tempos: a admiração
- a diferença entre autores no uso das cores e/ou dos sentimentos
- o uso da segunda pessoa em diferentes géneros
- a distribuição de nomes geográficos (próprios e comuns) por tema

Conclusão?

Ainda estamos no início.
Todas as sugestões e contribuições são muito bem vindas!

Referências I

- Sampson, Geoffrey. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Clarendon Press, Oxford, 1995.
- Sampson, Geoffrey. "Thoughts on two decades of drawing trees". In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, Kluwer Academic Publishers, 2003, pp. 23-41.