Artigos

Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos

Comparing linguistic annotations in Gramateca: philosophy, tools and examples

Diana Santos[*]
Rui Marques[**]
Cláudia Freitas[***]
Alberto Simões[****]
Cristina Mota[*****]

RESUMO: Neste artigo apresentamos a filosofia geral da Gramateca – um ambiente para fazer uma gramática da língua portuguesa baseada em corpos – e alguns estudos no seu âmbito, nomeadamente o estudo (1) dos conectores condicionais, (2) das palavras referentes ao corpo humano e (3) das emoções na língua. A ênfase é na metodologia, e apresentamos detalhadamente o sistema Rêve para rever e partilhar anotações linguísticas. Ao descrever os vários estudos, indicamos também as metamorfoses e melhorias por que essa ferramenta passou, assim como o tipo de perguntas e de resultados que já conseguimos obter em áreas muito diversas.

PALAVRAS-CHAVE: Corpos. Anotação. Semântica. Ferramentas de *corpora*.

ABSTRACT: This paper presents the general philosophy of Gramateca, for corpus-based Portuguese grammar studies, by reporting on three different studies – conditional connectives, body terms, and emotions – emphasizing methodological aspects. It presents in detail the Rêve system, which allows revising and sharing annotations of Rêve's underlying corpora. While describing the different studies we also report on the improvement of the Rêve tool, and discuss the kinds of questions and results already available for diverse fields.

KEYWORDS: Corpus. Annotation. Semantics. Corpora tools.

1. Apresentação

Neste artigo apresentamos a filosofia geral da Gramateca[1] – um ambiente para fazer gramática com base em corpos[2]. Gramática, aqui, é entendida em sentido amplo e compreende

[*] Universidade de Oslo e Linguateca
[**] Universidade de Lisboa (FLUL)
[***] PUC-Rio e Linguateca
[****] Universidade do Minho e Linguateca
[*****] Linguateca

[1] http://www.linguateca.pt/Gramateca

[2] Utilizamos ao longo do artigo o termo "corpo", como já foi defendido em Santos (2008, p. 43): "Vejamos o próprio exemplo da linguística com corpos: este último objecto tem sido variadamente chamado *corpora* (plural *corpora*), *córpus* (plural *corpora* ou *córpus*), mas parece não ter sido sequer equacionado o uso duma palavra genuinamente portuguesa e semelhante, *corpo*, empregue aliás de forma análoga em linguagem legal: *corpo de delito*. Na acepção mais lata de corpo como colecção de textos, é usada naturalmente a palavra *acervo* no Brasil,

não apenas aspectos (morfo-)sintáticos, mas também áreas relacionadas ao sentido e ao léxico, bem como questões vinculadas aos gêneros textuais e aspectos culturais, por exemplo. A ênfase é na metodologia, visto que em outras publicações cobrimos outros aspetos da Gramateca (Santos, 2014; Simões & Santos, 2014; Santos, 2015). Especificamente, iremos ilustrar como uma das ferramentas desse ambiente, o Rêve, direcionou o estudo relacionado a (a) conectores condicionais; (b) léxico do corpo humano e (c) emoções na língua.

No âmbito da Gramateca, uma das maneiras de se estudar e/ou descrever uma língua é por meio do estudo de grandes quantidades de texto. Uma das fontes de inspiração para este projeto foi a gramática de Biber et al. (1999) para o inglês; mas, passados quinze anos do projeto inicial, é possível ir além: não só porque os corpos em que a Gramateca se baseia são públicos – diferentemente do trabalho de Biber et al. (1999) – e porque as técnicas estatísticas modernas estão bem mais desenvolvidas, mas também porque, em vez de um simples etiquetador, temos acesso à riqueza da anotação morfossintática fornecida pelo PALAVRAS (Bick, 2000), e também a alguma anotação semântica criada pela Linguateca[3], nomeadamente cor (Silva & Santos, 2012), roupa, corpo humano (Freitas, 2013) e emoções (Santos e Mota, 2015).

De fato, o AC/DC[4] – um serviço de acesso a *corpus* e disponibilização de corpos, iniciado pela Linguateca em 1999 – constitui a base sobre a qual se sustenta a Gramateca. Atualmente, graças ao grande número de projetos que nos autorizaram a disponibilizar seus corpos, temos material público de qualidade para dar início ao projeto. Como indicado na página da Gramateca, temos consciência de que uma gramática (ou qualquer estudo) baseado em corpos depende do material em que se baseia. Ainda que seja provável que não tenhamos (ainda) o corpo ideal para escrever **a** gramática, e conscientes de que o material de que dispomos refere-se majoritariamente à língua escrita[5], acreditamos que o que for produzido pode ser relevante como passo inicial de futuros estudos, nem que seja pelo aspeto metodológico.

Enfatizamos o aspecto metolodógico porque a Gramateca é também um laboratório para o estudo da língua portuguesa, no qual estão disponíveis (a) todos os corpos disponibilizados

mas aparentemente não no sentido técnico associado a corpos electrónicos, mais influenciado pelo inglês. Infelizmente o uso não consagrou a possível expansão desse termo, provavelmente por não ter semelhanças suficientes com as designações inglesas/latinas. Proponho assim usar *corpo* e *corpos*, na esperança de que esta portuguesificação (e não aportuguesamento) seja aceite."

[3] A Linguateca é um centro de recursos para o processamento computacional da língua portuguesa; abriga o projeto Gramateca.

[4] Disponível em: http://www.linguateca.pt/ACDC.

[5] Embora essa seja uma limitação, lembramos que, comparativamente, existem muito mais estudos sobre o português oral (do Brasil e de Portugal), como ilustram o projeto NURC e o projeto do Português Fundamental.

pelo AC/DC; (b) a anotação automática desses corpos; (c) ferramentas de visualização e de exploração dos corpos, como a própria interface do AC/DC, e os serviços mais recentes Comparador e Disparador, descritos em Simões & Santos (2014); (d) o Rêve, um serviço para anotação manual de subconjuntos dos corpos associado a uma plataforma de revisão e de comparação de diferentes análises, foco do presente artigo.

A anotação linguística, portanto, é um dos elementos fundamentais da Gramateca, que almeja, além de contribuir com a descrição e o estudo do português, incentivar a prática de compartilhamento (e, consequentemente, a possibilidade de reanálise) dos dados que permitiram os estudos e sistematizações realizados por meio da anotação. Em outras palavras, porque acreditamos que a prática de compartilhamento do material classificado linguisticamente poderá servir de base para mais estudos (e controvérsias) sobre a gramática da língua portuguesa, investimos na viabilização dessa prática. Concordamos, portanto, com Sampson (2001) e Archer (2012), para quem textos contendo material anotado constituem a matéria-prima da maior parte da investigação linguística moderna.

Nossa intenção é contribuir com a metodologia científica no campo da linguística: não só permitir a repetição de uma experiência (o que é uma das propriedades exigidas à metodologia científica), mas também partilhar diferenças de interpretação de um mesmo material. Ou seja, enquanto nas ciências naturais se espera que a mesma experiência leve aos mesmos resultados, nas ciências humanas é não só esperável, mas provável, que haja diferenças na interpretação quando algo é repetido por outros pesquisadores. Estranhamente, tal situação é em geral criticada, em vez de ser vista como uma propriedade que resulta precisamente de lidarmos com um tipo diferente de ciência (ou conhecimento).

Neste artigo (e no projeto da Gramateca) tratamos diretamente dessa questão, criando um ambiente para que diferenças de interpretação possam ser mais do que comentários em artigos ou notas de rodapé na documentação de um corpo anotado. Para justificar a ênfase em um ambiente que promove o compartilhamento e a discussão da anotação de corpos, valemo-nos dos seguintes pressupostos:

(i) entendemos a anotação como uma forma de estudar a língua, e não apenas como uma atividade mecânica capaz de prover material para sistemas que processam a língua automaticamente. A anotação possibilita um estudo linguístico empírico, desenhado à maneira clássica, no qual se criam hipóteses (categorias/etiquetas provisórias) que serão verificadas durante a anotação. Nesse processo, as hipóteses iniciais podem ser confirmadas ou os dados podem levar à reformulação das categorias iniciais, e o processo recomeça. Enfatizamos a anotação como um

procedimento que envolve interpretação, classificação e formalização do fenômeno em foco; e

(ii) aceitamos o conceito de vagueza como fundamental na linguagem natural, como proposto e advogado por Santos (2006). Daí que aceitamos, por princípio, a possibilidade de mais de uma interpretação – e, consequentemente, de classes de anotação – ao mesmo segmento de texto, o que não significa que qualquer interpretação seja aceitável, e aqui invocamos o conceito de "comunidades interpretativas" de Fish (1980). Não consideramos tais diferenças interpretativas anomalias, mas antes visões alternativas que enriquecem a discussão de/sobre uma língua.

Note-se que, de um ponto de vista prático, a discussão sobre as opções linguísticas de anotação do material da Linguateca sempre foi possível, visto que os corpos sobre os quais trabalhamos na interface AC/DC são públicos e, portanto, qualquer interessado pode escolher um subconjunto das análises para rever. No entanto, até agora não havia um serviço que permitisse comparar tais análises, fazer novas análises, propor análises de fenômenos ainda não estudados ou anotados e, além disso, partilhá-las com outros. Agora há, e é disso que trataremos ao longo deste artigo, que está dividido em quatro partes. Na primeira, tratamos de anotação; na segunda, tratamos de dados sobre conectores condicionais. Tais dados, tornados públicos, motivaram a construção do Rêve, ferramenta que permite explorar de maneira simples a anotação como forma de estudo linguístico. Na terceira parte, relatamos uma experiência com o Rêve na qual foram possíveis o refinamento e a validação de um esquema de anotação complexo, o Esqueleto, com ênfase em alguns dos pontos mais controversos dessa proposta de anotação. Na quarta parte, apresentamos uma parcela de um estudo mais amplo sobre as emoções na língua portuguesa, no qual o Rêve atuou como ambiente de validação de uma hipótese.

2. Rêve: por uma outra ênfase na anotação?

Em geral, atividades de anotação são feitas em larga escala, porque são vinculadas a necessidades do processamento automático de uma língua. Quando a anotação é feita privilegiando o estudo linguístico – nosso caso –, a exigência da quantidade perde a relevância. Claro que amostras maiores tendem a apresentar casos novos, trazendo potenciais desafios a um dado esquema de anotação, mas casos particulares podem ser exatamente o que o pesquisador deseja. Porque os interesses são distintos – estudar a língua se sobrepõe à criação de um recurso linguístico para o processamento automático de uma língua (PLN) –, a

metodologia também pode ser outra: poucas frases e poucos – ou mesmo um único – anotador, que almeja tirar proveito da metodologia da anotação para materializar, em segmentos da língua em uso, uma dada classificação do fenômeno em análise.

Esse tipo de atividade-análise, de configuração de cenário de pesquisa, já é possível com o que temos hoje: a compilação de textos para constituir um corpo, a posterior utilização de um editor de textos para marcação e a disponibilização desses dados em alguma página na internet permitem a criação de um ambiente nos moldes do que apresentamos aqui, mas não sem algum trabalho. Assim, o que propomos com o Rêve é possibilitar ao linguista um ambiente de teste de hipóteses *on-line*, gratuito e público, no qual (a) os textos – associados a alguma (boa)[6] anotação linguística – já estão disponíveis para uma seleção simples das frases de interesse; (b) a criação e a in*corpora*ção das classes testadas também são extremamente simples; e (c) a tabulação dos resultados é automática.

Em contrapartida, retomando a ideia de anotação em larga escala, é verdade que a "anotação massiva" tem sido bastante apregoada pelos defensores do "*crowd-sourcing*", que também a aplicam à linguística (Munro et al. 2010). Embora em alguns casos isso seja não só apropriado como a única maneira de o fazer, noutros casos – por exemplo, em Rumshisky (2009) – pensamos que se tem ido longe demais, como aliás, os próprios Rumshisky et al. (2012) parcialmente discutem, apontando os (novos) problemas que esse tipo de tarefa acarreta – por exemplo, devido à falta de confiança nos próprios anotadores. Isto acontece porque a anotação (a tarefa linguística) por meio de um sistema como o Amazon MTTurk[7] é feita por pessoas sem comprometimento com a tarefa, e não por colegas interessados no mesmo tipo de assuntos – peritos – e cujas análises, além disso, são tornadas públicas de forma identificada.

Outra questão muito debatida em projetos associados à anotação humana é a concordância entre anotadores[8], com a ideia implícita de que quantidade (na concordância) pode dizer alguma coisa sobre qualidade. Como Reidsma & Carletta (2008) demonstraram, quando diferentes anotadores "erram" na mesma direção, os erros acabam por transformar-se em concordâncias, o que sem dúvida é um problema. Por outro lado, a discordância na anotação

[6] O analisador PALAVRAS é reconhecidamente um dos melhores e mais ricos para a língua portuguesa.
[7] O Mechanical Turk é um sistema lançado pela empresa americana Amazon em 2005, no qual se contratam pessoas (público em geral, não especialistas) para realizar tarefas intelectuais relativamente simples, mas em casos em que seja preciso fazer muitas vezes para obter volumes quantitativamente satisfatórios. O Mechanical Turk não é necessariamente para pesquisa. Veja em: https://www.mturk.com/mturk/welcome.
[8] Veja-se Tinsley & Weiss (2000) para uma apresentação matemática e Artstein & Poesio (2008) para uma panorâmica extensa no PLN. Contudo, ainda existem bastantes vozes dissidentes, como Powers (2012) e Uebersax (s/d).

não é necessariamente um indicativo de baixa qualidade, sobretudo em áreas onde a interpretação e o conhecimento do contexto e do assunto podem ser mais importantes do que o "palpite".[9] No entanto, achamos que é conveniente quantificar casos em que a nossa própria interpretação vacila, porque isso pode precisamente apontar quer para deficiências da análise proposta, quer para casos de mudança sincrônica em progresso. Por essa razão, essa quantificação é uma das tarefas que propomos.

Para nós, um dos pontos de interesse quando deslocamos o foco da tarefa de anotação da construção de um recurso linguístico-computacional para uma maneira de estudar fenômenos linguísticos é explicitar as diferentes possibilidades de leitura, as diferentes interpretações, isto é, as discordâncias. Interessa-nos também, no estudo de uma língua, a divergência e a alteridade, pois delas também se alimenta a pesquisa, sobretudo nas letras e nos estudos sociais.

3. Conectores condicionais

A ideia base do estudo sobre os conectores condicionais é obter dados quantitativos sobre diferentes tipos de contextos em relação aos conectores condicionais "a + infinitivo" (que é usado com esse sentido sobretudo em português europeu), "se", "caso" e "no caso de". Duzentas ocorrências de cada tipo foram anotadas quanto à sua base modal (epistêmica ou circunstancial), assim como foi marcado se eram "condicionais de enunciação" ou não (cf. Lopes 2009). Simplificadamente, a anotação tem em conta se a relação entre a oração condicional e a frase matriz é uma relação do plano epistémico[10] (como em *se a rua está molhada, [então] choveu*), do plano ontológico (como em *se não houver oxigénio suficiente na água, os peixes acabam por morrer*) ou do plano da enunciação (como em *se tiveres sede, há cerveja no frigorífico*). Esses dados (texto e anotação) foram depois tornados públicos na página da Gramateca.

[9] Um exemplo anedótico é quando a interpretação de um verbo em contexto pode ser significativamente melhorada por conhecer o resto da entrevista ou mesmo conhecer pessoalmente o entrevistado, como descrito em Bacelar do Nascimento et al. (1993).

[10] A distinção entre o plano epistémico, ou dedutivo, o plano ontológico, ou de conteúdo, e o plano da enunciação é observada por Sweetser 1991. Simplificadamente, a construção condicional remete para o plano epistémico se o nexo entre o antecedente e o consequente for um nexo dedutivo (grosso modo, "se p, então q" será equivalente a "a verdade de p permite concluir a verdade de q"); remete para o plano ontológico se o nexo entre antecedente e consequente for próximo do de uma relação de causalidade (grosso modo, "se p, então q" será equivalente a "a situação descrita por p leva à situação descrita por q"); e remete para o plano da enunciação se, simplificadamente, a asserção do antecedente for uma justificação para a asserção do consequente. Para uma explicação mais detalhada sobre a distinção entre estes três planos, ver Sweetser 1991.

Além da anotação relativa ao tipo de condicional, foram também anotados os casos em que os conectores considerados não são operadores condicionais. Por exemplo, "se" pode ser também uma conjunção integrante (como em *ele perguntou se alguém telefonou)* e "a" pode ser um operador aspectual (como em *ele está a cantar*). Nessas ocorrências, a anotação do operador foi marcada como "OUTR". Os resultados teóricos desse estudo foram publicados em Marques (2014).

Esse trabalho anterior de análise, viabilizado no ambiente da Gramateca, foi precioso na possibilidade de permitir fixar os requisitos de especificação de uma ferramenta que permitisse anotar os corpos e torná-los públicos e revisáveis por outros interessados, permitindo tanto uma anotação inicial como anotações alternativas; portanto, anotações múltiplas. Entre as várias lições que aprendemos com esta atividade, destacamos a necessidade de levar em conta erros quer da anotação, quer dos próprios corpos[11], assim como a vantagem de poder obter *a posteriori* mais informação sobre cada ocorrência. Assim nasceu a ferramenta Rêve.

A Figura 1 mostra o Rêve na sua vertente de anotação, em que os revisores podem ter acesso a cada um dos exemplos obtidos e selecionar a categoria que lhes parece correta. Note-se que, de modo a que esses revisores não sejam influenciados por prévias anotações, o sistema não sugere qualquer categoria *por omissão* (ou seja, em inglês, *by default*). Note-se, também, a possibilidade de atribuição de mais de uma classe, simultaneamente, materializando, dessa forma, o conceito de vagueza na língua.

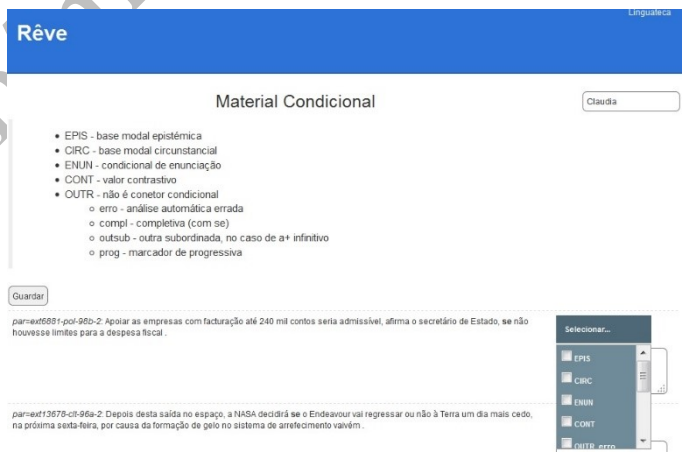
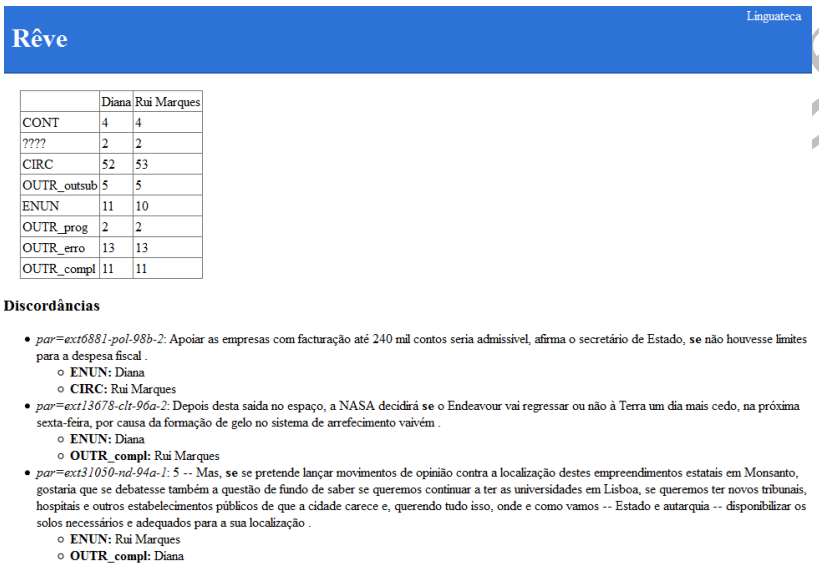


Figura 1. Anotação de exemplos no Rêve.

[11] Mais especificamente: alguns exemplos selecionados automaticamente tinham erros de digitação ou de ortografia, por exemplo.

Posteriormente, os resultados de anotação podem ser analisados e comparados. Se, por um lado, a anotação *concordante* entre diferentes revisores é útil para a obtenção de um recurso anotado; por outro lado, a informação *discordante* é igualmente útil para posterior discussão ou, simplesmente, para a obtenção daquelas situações duvidosas. A Figura 2 mostra o resultado de comparação das anotações relativas aos conectores condicionais, levando em consideração a contabilização independente, isto é, a distribuição de cada classe, por anotador. Veja-se Simões & Santos (2014) para outros exemplos.



Linguateca

Rêve

	Diana	Rui Marques
CONT	4	4
????	2	2
CIRC	52	53
OUTR_outsub	5	5
ENUN	11	10
OUTR_prog	2	2
OUTR_erro	13	13
OUTR_compl	11	11

Discordâncias

- *par=ext6881-pol-98b-2:* Apoiar as empresas com facturação até 240 mil contos seria admissível, afirma o secretário de Estado, **se** não houvesse limites para a despesa fiscal .
  - **ENUN:** Diana
  - **CIRC:** Rui Marques
- *par=ext13678-clt-96a-2:* Depois desta saída no espaço, a NASA decidirá **se** o Endeavour vai regressar ou não à Terra um dia mais cedo, na próxima sexta-feira, por causa da formação de gelo no sistema de arrefecimento vaivém .
  - **ENUN:** Diana
  - **OUTR_compl:** Rui Marques
- *par=ext31050-nd-94a-1:* 5 -- Mas, **se** se pretende lançar movimentos de opinião contra a localização destes empreendimentos estatais em Monsanto, gostaria que se debatesse também a questão de fundo de saber se queremos continuar a ter as universidades em Lisboa, se queremos ter novos tribunais, hospitais e outros estabelecimentos públicos de que a cidade carece e, querendo tudo isso, onde e como vamos -- Estado e autarquia -- disponibilizar os solos necessários e adequados para a sua localização .
  - **ENUN:** Rui Marques
  - **OUTR_compl:** Diana

Figura 2. Comparação de anotações discordantes.

O conjunto total das anotações de um ou de todos os revisores pode ser facilmente exportado para um arquivo de texto com valores separados por caracteres de tabulação, o que permite que esses dados possam ser subsequentemente processados por outras ferramentas de análise estatística, como o R (R Core Team, 2015).

4. Como anotar características de partes do corpo humano: o projeto Esqueleto

Em relação às partes do corpo, a nossa intenção foi usar o Rêve para estudar a possibilidade de consenso sobre algumas questões de interpretação de usos de palavras de partes do corpo humano. Em Freitas *et al*. (2015), foram propostas classes semânticas para as palavras do léxico do corpo humano. Em termos gerais, a proposta consiste em distinguir, por um lado, as palavras do corpo humano que fazem referência, no contexto da frase, ao corpo humano de fato, daquelas que se distribuem por outros campos semânticos. Para esses sentidos não

*corpora*is, propomos treze classes semânticas[12], além da classe genérica "outros". Se algumas classificações correspondem a estabilizações consensuais (por exemplo, a classe "corpo:medida" para *dois **dedos** de pinga*), em outros casos as estabilizações propostas podem não ser tão evidentes. Nesse sentido, o Rêve auxiliou a tarefa de avaliação da classificação proposta e a tarefa de validação dessas mesmas categorias. Para tanto, escolhemos um conjunto de categorias de anotação que nos pareceram especialmente finas e, portanto, com maior potencial para discordância, como: (a) a diferença entre parte de algo ou lugar (virtual), como em *O calibre real é a grandeza medida directamente na **boca** do cano, sem qualquer artefacto, ou no projéctil*.; e (b) a diferença entre expressões que envolvem sentimento ou opinião, como em *Não é daqueles vinhos que se possa **torcer o nariz***). Selecionamos um conjunto de exemplos e pedimos a pessoas interessadas no assunto, mas não necessariamente envolvidas com o projeto de anotação, para os reanotarem, apenas com base nas nossas descrições genéricas das classes envolvidas.

Dois grupos de dados foram assim criados[13]: o Esqueleto 1 e o Esqueleto 2, com cerca de 100 frases cada um, a partir de procura no AC/DC de exemplos que já haviam sido anotados, mas que tinham dado origem a discussões e incertezas, junto com outros casos que pareciam menos complicados. Cada grupo lidava com diferentes classes. Para ver se o contexto da tarefa influenciava o julgamento/escolha dos participantes, alguns exemplos foram incluídos nos dois conjuntos – precisamente 23 exemplos.

Os resultados[14] da anotação, por um lado, confirmaram a complexidade da proposta, o que ficou evidente pela grande discordância; por outro lado, esses mesmos resultados levaram à reformulação de uma das classes – as classes partedeobjeto e membro foram reunidas na classe parte – o que motivou uma segunda rodada de anotação. Na segunda rodada, a concordância obtida subiu para cerca de 85%, um número bastante razoável quando se trata de anotações semânticas. No entanto, para esse cálculo, consideramos também concordância a sobreposição de classes[15], alternativa possível porque já havíamos determinado que uma

[12] As classes semânticas são: sentimento, vegetal, parte, lugar, doença, opinião, posição, animal, movimento, faculdade, grupo, medida e centralidade.
[13] Disponíveis em: http://www.linguateca.pt/reve/stats/8 e http://www.linguateca.pt/reve/stats/10.
[14] Os resultados teóricos provenientes dessas anotações encontram-se em Freitas et al.2015.
[15] Para contabilizar a concordância, que neste caso é parcial, consideramos a sobreposição de classificações como concordante, mas consideramos que cada maneira de classificar corresponde a uma interpretação diferente. O Rêve também calcula outra medida de concordância, a concordância estrita, que só considera concordância se duas (ou mais) classificações/interpretações forem exatamente iguais.

palavra poderia receber mais que uma análise. Assim, a análise da palavra "nervos", na frase abaixo, foi considerada concordante:

> A equipe não controlou os **nervos**, errou passes e insistiu nas bolas altas, e os principais jogadores estiveram muito apáticos.
> **SENTIMENTO:** anotador1, anotador2
> **SENTIMENTO_FACULDADE:** anotador3

A análise das discordâncias, por sua vez, revelou que, em quase todos os casos, uma das categorias selecionadas era a categoria "outros", evidenciando a dificuldade de enquadramento de certas palavras/sentidos nas classes propostas, por um lado, e a sobreposição de sentidos, por outro.

É importante mencionar ainda que casos considerados simples pela equipe do projeto revelaram-se também alvo de controvérsia quando tornados públicos no ambiente do Rêve, indicando que mesmo esses deveriam ser repensados ou ter a sua explicação melhorada.

Especificamente com relação ao Esqueleto, a utilização do Rêve não foi especialmente original , uma vez que uma das intenções era validar o esquema de anotação proposto. Na próxima seção, trataremos de outro tipo de utilização, cujo principal objetivo é ilustrar a divergência.

5. Um estudo das emoções com a Gramateca: os diferentes sentidos de *admirar*

O lado emotivo da língua é uma área que tem recebido algum interesse nos últimos tempos[16], ou seja, o fato de que uma língua não é apenas um veículo de comunicação de fatos, mas também, ou sobretudo, de emoções e de valores. A esse respeito consideramos que foi muito infeliz, para a linguística, a separação logocêntrica entre verdade objetiva e opinião, que pôs a ênfase nos valores de verdade e na informação transmitida[17]. O interesse renovado sobre as emoções na língua não significa que existam soluções ou consenso (ainda), e a Gramateca pretende dar um contributo à investigação desse tema em relação ao português. Usaremos, pois, esse nosso interesse no estudo das anotações emotivas com a intenção de verificar os diferentes sentidos da emoção que podemos representar pela palavra *admirar* e outras que podem significar tanto surpresa quanto apreciação. (Mota e Santos, 2015).

[16] Ver Pang & Lee, 2008, para uma compilação de trabalhos recentes na área.
[17] Opinião essa também corroborada, por exemplo, por Arrojo (1992).

Um dos temas relevantes na análise de sentimentos ou de opiniões é até que ponto essas são consensuais ou se dependem do utilizador/leitor de dada obra ou texto. Em Santos e Mota (2015), estudamos a diferenciação entre duas "emoções" associadas ao lexema *admirar*, uma envolvendo respeito (que consideramos positiva) e outra, surpresa (que pode ser positiva ou negativa). Embora à partida essa diferença seja fácil de compreender, o estudo detalhado e a anotação por várias pessoas diferentes revelaram que não são poucos os casos em que as interpretações podem ser tanto divergentes como ambíguas ou vagas. O exercício com o Rêve foi desenhado para verificar precisamente essa questão: até que ponto uma mesma emoção (admiração) pode carregar dois sentidos com polaridade antagônica? Até que ponto a diferença entre *surpresa* e *respeito* é consensual entre falantes de português? Adicionalmente, seria possível que uma mesma ocorrência em contexto desse margem a interpretações distintas? Nossa aposta era de que a anotação, por diferentes pessoas, de frases selecionadas evidenciasse discordâncias trazendo assim questões teóricas relevantes para o quadro dos estudos de emoção e sentimentos.

Para estudar essas questões, selecionamos 108 casos que continham palavras como *admirar* e *maravilhar.* Desses 108, em 63 houve discordâncias, quantidade considerável. Com relação a *surpresa* e *respeito*, em 20 casos alguma das anotadoras considerou a presença de ambos os sentidos, e apenas eles, na frase (concordância estrita, ver nota 15). Em apenas um caso, todas as anotadoras concordaram quanto à presença de ambos os sentidos, simultaneamente:

> 167581: -- De tudo isso, aventurou Raimundo, o que mais me **admira** é a sua memória: o senhor com efeito tem uma memória de anjo.

Esses dados, facilmente obtidos com o Rêve, corroboram a hipótese inicial sobre a dificuldade, em certos casos, de distinção entre esses sentimentos.

Adicionalmente, a tarefa com o Rêve nos mostrou outros pontos para o estudo dos sentimentos em língua portuguesa:

(1) existem ainda mais sentidos associadas ao lexema *admirar*, o que nos fez redesenhar o experimento e incluir a classe/etiqueta *apreciar*, empregada consensualmente – e exclusivamente[18], em 6 casos, como em:

> 237065: Ele ficou algum tempo a contemplá-la naquela posição, que a fazia mais

[18] Apenas o sentido de "apreciar" foi atribuído.

bonita, e, perdido em saudosas reminiscências da sua mocidade, **admirava** a curva macia dos seios, palpitantes, sob a compressão.

(2) existem casos em que nenhum dos três sentidos está em jogo;

(3) existem casos em que os três sentidos são possíveis. Especificamente, em sete casos, alguma das anotadoras considerou a presença simultânea dos três sentidos, mas não houve concordâncias em quaisquer desses casos; e

(4) não são poucos os casos em que cada avaliador interpretou de uma maneira distinta.

A Figura 3 apresenta a distribuição das classes atribuídas por anotadora, considerando as classificações múltiplas. O cálculo é feito automaticamente pelo Rêve.

Contabilização Combinatória

Nesta contabilização, cada classificação múltipla é considerada como uma categoria independente.

	Claudia	Cristina	Diana
APRECIAR	21	18	9
APRECIAR|NENHUM	1	0	0
APRECIAR|RESP	11	9	20
APRECIAR|RESP|SURP	1	2	4
APRECIAR|SURP	11	2	3
NENHUM	4	3	2
NENHUM|SURP	2	0	0
RESP	17	29	36
RESP|SURP	3	11	6
SURP	25	34	28

Figura 3. Distribuição das discordâncias no estudo das emoções.

Diferentemente do Esqueleto, no caso das emoções, mais do que buscar, com várias rodadas de anotação, a concordância das interpretações em todos os casos, interessou-nos observar a variabilidade de interpretações que palavras de emoção podem ter.

6. Observações finais

O Rêve é uma ferramenta criada para explorar a ideia de pesquisa e aprendizado (humano) por meio da anotação. Está inserido no contexto da Gramateca, que tem como propriedade essencial partilhar publicamente os seus dados e discussões sobre gramática – em sentido amplo – do português. Uma vez públicos, os dados estão disponíveis para que sejam

discutidas correlações e controvérsias, assim, fomentando, por exemplo, o compartilhamento e replicação de experiências.

Partilhamos conhecimento por meio de artigos e capítulos de livros, mas não temos o costume de partilhar dados, especificamente anotações. Como argumenta Xiao (2009) em defesa da anotação, analisar linhas de concordância, classificando-as de acordo com a intuição, é também um processo de anotação, só que um processo implícito. É justamente esse caráter velado, que torna a classificação irrecuperável e, portanto, bem menos confiável que a anotação explícita.

Por isso, o objetivo da ferramenta está menos vinculado ao resultado "convencional" da anotação – o material anotado – que ao processo de anotação propriamente. Pelo mesmo motivo, o Rêve não é, primordialmente, uma ferramenta para medir a concordância entre anotadores, embora isso também seja possível, mas antes um facilitador para a revisão e reanotação de materiais linguísticos variados, cujo objetivo é não apenas torná-los acessíveis, mas também permitir que a sua manipulação e a medição de diferenças conceituais não demandem conhecimentos informáticos muito específicos.

Estamos conscientes de que existem vários ambientes para corrigir anotações ou fazer anotação a partir do zero[19], mas queríamos desenvolver algo que pudesse interagir da melhor maneira possível com a Gramateca, não só quanto à escolha inicial de dados, mas também quanto às informações associadas às frases que já se encontram no AC/DC, e com isso possibilitar o cruzamento de informações diferentes ou estudos de replicação. Neste último caso, pretendemos que, se nova informação – ou simplesmente informação mais atualizada – vier a ser in*corpora*da nos corpos,[20] possamos fazer uso dela no Rêve sem ter de reanotar tudo outra vez. Em outras palavras, após um pesquisador ter passado várias semanas a caracterizar/anotar um conjunto de casos e tê-lo tornado público através da Gramateca e, portanto, passível de ser investigado por outros pesquisadores, também com base em outros critérios, seria certamente desolador que, na próxima versão dos corpos subjacentes à Gramateca, essa informação deixasse de poder ser utilizada ou necessitasse da repetição do trabalho anterior.

[19]Alguns exemplos são o Brat Rapid Annotation Tool (http://brat.nlplab.org), o GATE (https://gate.ac.uk/demos/movies.html#section-1.2.5) e o EtiquetHAREM (http://www.linguateca.pt/aval_conjunta/HAREM/ManualUtilEtiquetHAREM.pdf).

[20] Os problemas que a existência de corpos dinâmicos (melhorando com o tempo) e a anotação de versões antigas desses corpos podem acarretar são discutidos por exemplo por Santos (2014c).

De maneira a garantir a integração do Rêve na Gramateca, um dos nossos critérios de desenho é poder cruzar o estudo da anotação humana com os resultados do processamento automático ou semiautomático. Assim, juntamos estudos qualitativos detalhados, exigindo profundo conhecimento linguístico, a estudos quantitativos abrangendo grandes quantidades de dados, cuja análise humana é impossível.

Referências Bibliográficas

ARCHER, D. *Corpus* annotation: a welcome addition or an interpretation too far? In: TYRKKÖ, J.; KIPIÖ, M.; NEVALAINEN, T.; RISSANEN, M. (Org.). Outposts of historical *corpus* linguistics: from the Helsinki *corpus* to a proliferation of resources. **Studies in Variation, Contacts and Change in English**. Vol. 10, 2012. Disponível em http://www.helsinki.fi/varieng/series/volumes/10/archer

ARROJO, R. **O signo desconstruído**. São Paulo: Pontes, 1992.

ARTSTEIN, R.; POESIO, M. Inter-coder agreement for computational linguistics. **Computational Linguistics**, v. 34, n. 4, p. 555-596, 2008. crossref http://dx.doi.org/10.1162/coli.07-034-R2

BACELAR DO NASCIMENTO, M. F.; MENDES, A.; SANTOS, D. O *corpus* e a classificação sintáctica dos verbos. **Actas do 1.o Encontro de Processamento de Língua Portuguesa (escrita e falada) - EPLP'93** (Lisboa, 25-26 de fevereiro de 1993), pp. 125-129.

CRAGGS, R.; WOOD, M. M. Evaluating Discourse and Dialogue Coding Schemes. **Computational Linguistics** 31, No. 3, September 2005, pp. 289-296. crossref http://dx.doi.org/10.1162/089120105774321109

FISH, S. E.. **Is There A Text in This Class? The Authority of interpretive communities**. Cambridge: Harvard University Press. 1980.

FREITAS, C. de. **Esqueleto**: anotação das palavras do corpo humano. Primeira edição: 15 de novembro de 2013. Disponível em: http://www.linguateca.pt/acesso/Esqueleto/Esqueleto.html

FREITAS, C. de; SANTOS, D.; CARRIÇO, B.; JANSEN, H.; MOTA, C. Investigação do léxico do corpo humano e anotação semântica de *corpus*. **Revista de Estudos da Linguagem** v. 23, n. 3, 2015 (no prelo).

LOPES, A. C. M. Contributos para o estudo de construções condicionais não-canónicas em Português europeu contemporâneo. **Diacrítica, Ciências da Linguagem**, 23/1, pp. 149-169, 2009.

MARQUES, R. Modalidade e condicionais em português. **Revista Virtual de Estudos da Linguagem - ReVEL,** edição especial, vol. 12, n. 8, pp. 106-130, 2014.

MOTA, C.; SANTOS, D. **Emotions in natural language: a broad-coverage perspective**. Janeiro de 2015. Disponível em: http://www.linguateca.pt/acesso/emotionsBC.pdf.

MUNRO, R.; BETHARD, S.; KUPERMAN, V.; TZUYIN LAI, V.; MELNICK, R.; POTTS, C.; SCHNOEBELEN, T.; TILY, H.. Crowdsourcing and language studies: the new generation of linguistic data. In **Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10).** Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 122-130.

PANG, B.; LEE, L.. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval** vol. 2, pp.1-135, 2008. crossref http://dx.doi.org/10.1561/1500000011

POWERS, D. M. W. The Problem with Kappa. In **Proceedings of EACL 2012**, pp. 245-355, 2012.

R Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2015. Disponível em: http://www.R-project.org.

REIDSMA, D.; CARLETTA, J. Reliability Measurement without Limits. **Computational Linguistics**, 34(3), pp. 319-326, 2008. crossref http://dx.doi.org/10.1162/coli.2008.34.3.319

RILOFF, E.; WIEBE, J.; PHILIPS, W. Exploiting subjectivity classification to improve information extraction. **Proc. 20th National Conference on Artificial Intelligence (AAAI-2005)**, volume 3, pp. 1106-1111, 2005.

RUMSHISKY, A. Crowdsourcing Word Sense Definition. In **Proceedings of the Fifth Linguistic Annotation Workshop (LAW-V)**. *ACL-HLT 2011*. Portland, Oregon, 2011.

RUMSHISKY, A.; BOTCHAN, N.; KUSHKULEY, S.; PUSTEJOVSKY, J. Word Sense Inventories by Non-Experts. In CALZOLARI, N.; CHOUKRI, K.; DECLERCK, T.; DOGAN, M. U.; MAEGAARD, B.; MARIANI, J.; ODIJK, J.; PIPERIDIS, S.. **Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)** (Istambul, 23-25 de Maio de 2012).

SAMPSON, G. **Empirical Linguistics**. London: Continuum, 2001.

SANTOS, D. What is natural language? Differences compared to artificial languages, and consequences for natural language processing. Palestra convidada, **SBLP2006** e PROPOR'2006, Itatiaia, RJ, Brasil, 15 de Maio de 2006.

SANTOS, D. Corporizando algumas questões. In: TAGNIN, S. E. O..; VALE, O. (Orgs.). **Avanços da Linguística de *Corpus* no Brasil**. São Paulo: Editora Humanitas/FFLCH/USP, 2008, pp.41-66.

SANTOS, D. Podemos contar com as contas?. In ALUÍSIO, S.; TAGNIN, S. E. O.. **New Language Technologies and Linguistic Research:** A Two-way Road. Cambridge Scholars Publishing, 2014, pp. 194-213.

SANTOS, D. Gramateca: *corpus*-based grammar of Portuguese. In BAPTISTA, J.; MAMEDE, N.; CANDEIAS, S.; PARABONI, I.; PARDO, T. A. S.; VOLPE NUNES, M. das G. **PROPOR 2014**, LNAI 8775. Springer, Heidelberg (2014), pp. 214-219. crossref http://dx.doi.org/10.1007/978-3-319-09761-9_24

SANTOS, D. PoNTE: apontando para corpos de aprendizes de tradução avançados. **Linguamática** 6, 1, 2014, pp. 69-86.

SANTOS, D. Comparando corpos orais (transcritos) e escritos na Gramateca. In: BARDEL, C. **Proceedings from the conference Parler les langues romanes/Parlare le lingue romanze/Hablar las lenguas romances/Falando línguas românicas GSCP 2014**, University Press Università di Napoli L'Orientale, 2015.

SANTOS, D.; MOTA, C. A admiração à luz dos corpos. In SIMÕES, A.; BARREIRO, A.; SANTOS, D; SOUSA-SILVA, R.; TAGNIN, S. E. O. **Linguística, Informática e Tradução:** Mundos que se Cruzam. Homenagem a Belinda Maia, OSLa, Vol 7, No 1, 2015, pp. 57-77.

SILVA, R.; SANTOS, D. **Arco-íris**: notas sobre a anotação do campo semântico da cor em português. Disponível em: http://www.linguateca.pt/acesso/arcoiris.pdf.

SIMÕES, A.; SANTOS, D.. Nos bastidores da Gramateca: uma série de serviços. **Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish**, at PROPOR 2014, São Carlos, Brasil, 9 de outubro de 2014, pp. 97-104, 2014.

SWEETSER, E. **From Etymology to Pragmatics**: Metaphorical and Cultural Aspects of Semantic Structure. Cambridge University Press, 1991.

TINSLEY, H. E. A.; WEISS, D. J. Interrater Reliability and Agreement. In: TINSLEY, H. E. A.; BROWN, S. D. (org), **Handbook of Applied Multivariate Statistics and Mathematical Modeling**, San Diego, CA: Academic Press, 2000, pp. 95-124. crossref http://dx.doi.org/10.1016/B978-012691360-6/50005-7

UEBERSAX, J. **Kappa Coefficients**: A Critical Appraisal. Disponível em: http://www.john-uebersax.com/stat/kappa.htm.

XIAO, R. Theory-driven *corpus* research: Using *corpora* to inform aspect theory. In: LÜDELING, A.; KYTÖ, M. ***Corpus* Linguistics**: An International Handbook. Berlin: De Gruyter, 2009, volume 2, pp. 987-1.008.

Artigo recebido em: 29.03.2015
Artigo aprovado em: 20.04.2015