# Extração automática de candidatos a termos do *Curso de Linguística Geral* com apoio de recursos da Linguística de *Corpus* e do Processamento de Linguagem Natural[1]

## Automatic extraction of term candidates from *Course in General Linguistics* with resources from *Corpus* Linguistics and Natural Language Processing

Maria José Bocorny Finatto[*]
Lucelene Lopes[**]
Alena Ciulla[***]

**RESUMO**: Este trabalho apresenta um estudo em que técnicas de Processamento de Linguagem Natural (PLN) e de Linguística de *Corpus* (LC) são utilizadas para extrair e estruturar termos relacionados a conceitos importantes de Saussure no texto em português do *Curso de Linguística Geral* (CLG). Tomando o CLG como um *corpus*, busca-se um método de representação automática de conteúdo através de ferramentas computacionais. Uma vez submetido ao *parser* PALAVRAS, um etiquetador morfossintático para a língua portuguesa, o *corpus* do CLG é processado pela ferramenta extratora de sintagmas nominais relevantes, denominada ExATOlp, que implementa diversas técnicas de PLN de base linguística e de base estatística. Em seguida, são geradas listas e gráficos hierarquizados dos sintagmas nominais do CLG, elencados pela ferramenta como os mais específicos/relevantes do *corpus* em questão. Esses resultados são comparados com dados gerados pela ferramenta AntConc, ferramenta de acesso livre bastante empregada em trabalhos de LC, aplicada ao mesmo *corpus*. Os resultados mostram o potencial da ferramenta ExATOlp para trabalhos em LC e para o levantamento de dados lexicais para estudos terminológicos, para a mineração de dados e para a geração de ontologias em língua portuguesa.

**ABSTRACT**: This paper presents a study based on Natural Language Processing techniques (PLN) and Corpus Linguistics (CL) approaches to extract terms related to important saussurean concepts in the Brazilian Portuguese edition of the *Course in General Linguistics*. Taking the CGL as a corpus, we aim at an automatic representation method of content through computer tools. Once submitted to the parser PALAVRAS, a morphossyntatic tagger, the corpus is processed by ExATOlp, a tool implementing various linguistic and statistically based NLP techniques. The tool generates hierarchical lists and charts of noun phrases, which are organized according to their specificity / relevance in the target corpus. These lists are then compared to data generated by AntConc - a free access tool quite used in LC approaches - applied to the same corpus. The results show the potential of ExATOlp in works on LC and in colleting lexical data for terminology studies, data mining and generation of ontologies in Portuguese.

[1] Este trabalho é uma extensão de outros relatos de resultados relacionados ao mesmo projeto de pesquisa sobre representação automática de conteúdo de textos científicos, tal como vemos em Ciulla, Finatto (2013).

[*] Docente do PPG-Letras-UFRGS e pesquisadora CNPq.
[**] Professora colaboradora da FACIN-PUCRS e pós-doutoranda DOCFIX-FAPERGS/CAPES.
[***] Professora visitante do Departamento de Linguística, Filologia e Teoria Literária e do PPG- Letras-UFRGS e pós-doutoranda DOCFIX-FAPERGS/CAPES.

**PALAVRAS-CHAVE**: Extração automática de termos. Curso de Linguística Geral. ExATOlp.

**KEYWORDS**: Automatic extraction of terms. Course in General Linguistics. Saussure.

## 1. Introdução

### 1.1 A extração automática de termos em *corpora*

A extração automática de termos, conforme Di Felippo (2013, p. 66) consiste na identificação e na recolha, a partir de um *corpus* especializado, de expressões linguísticas que tenham um potencial terminológico. No Brasil, essa extração tem sido procedida no âmbito do Processamento de Linguagem Natural (PLN), uma especialidade da Ciência da Computação – cm um caráter bastante aplicado, e também no âmbito da Linguística de *Corpus* (LC), que se integra à Linguística Aplicada, com um caráter descritivo e de auxílio a pesquisas de Terminologia e Terminografia.

Em que pesem as enormes vantagens dessa automatização, ainda assim, o verdadeiro estatuto terminológico dessas unidades deverá ser posteriormente confirmado, seja em novos levantamentos ou em novos contrastes em *corpora* ou por especialistas do domínio. Assim, a extração ainda é um trabalho de máquina, que precisaria ser validado, para que haja uma identificação propriamente dita de unidades terminológicas.

A extração automática, cujas melhores margens de acerto encontram-se ainda em torno de 27% (conforme TEIXEIRA, 2011) para *corpora* em português do Brasil (PB), vem se desenvolvendo, especialmente desde a década de 80 para diferentes idiomas. Nela têm sido empregadas abordagens estatísticas e índices como *tf-idf* e *log likelihood*. Vale lembrar que são ainda muito discutidas as medidas ou pontos de corte de frequência ou de distribuição de uma dada expressão ou palavra ao longo de diferentes textos que integram um *corpus* para que uma unidade ou expressão multipalavra possa se enquadrar como um "candidato a termo".

Ferramentas informatizadas para essa extração partem de cálculos robustos do ponto de vista estatístico e matemático, mas, em geral, não fazem uso de informações linguísticas ou de descrições de linguagens especializadas, o que limita a obtenção de melhores resultados, de acordo com Lopes (2012) e Teixeira (2011). Os melhores resultados, em termos de trabalho automático, em diferentes idiomas, são os de abordagens que utilizam regras linguísticas em seu algoritmo. Contudo, a quase totalidade desses trabalhos são destinados ao inglês, alemão ou francês, sendo raro encontrar abordagens voltadas ao português.

Outra dificuldade da extração automática é a especificidade de certas abordagens que são dependentes de domínios bem pontuais, tal como o trabalho de Bui e Slot (2010), que se aplica à área do comportamento biológico de proteínas. Assim, algumas metodologias de extração tornam-se muito específicas de uma língua ou de um domínio e tendem a não poderem ser replicadas para *corpora* de diferentes áreas de conhecimento.

### 1.2 Foco deste trabalho

Este trabalho relata parte dos resultados da pesquisa *Recuperação da informação em representação do conhecimento em bases de textos científicos de Linguística e de Medicina*, que associa Processamento de Linguagem Natural (PLN), Linguística de *Corpus* (LC) e Estudos do Texto e do Discurso. Neste artigo, relatamos um estudo com o *corpus* da área de Linguística, representado pelo Curso de Linguística Geral (CLG). Na pesquisa ampla são explorados dois tipos de *corpora* de textos científicos em português brasileiro (PB): a) um *corpus* da área de Medicina, sobre Pneumopatias Ocupacionais – com textos de artigos científicos, dissertações, teses e textos de popularização para leigos; e b) um *corpus* da área de Linguística, representado pelo todo do texto em português do CLG. Esses dois *corpora*, bastante heterogêneos, foram escolhidos, deliberadamente, por representarem, respectivamente duas situações distintas:

a) um domínio e gêneros textuais e discursivos recorrentemente tratados em pesquisas[2] de PLN ou de LC e também em estudos bibliométricos, que visam traçar um perfil da produção científica em diferentes tópicos ou em temas das Ciências da Saúde ao longo de um dado período de tempo, tipo de publicação ou de um periódico específico;

b) um domínio (Linguística) e um gênero (o manual acadêmico ou livro-texto universitário) ainda pouco explorados por pesquisas de LC[3] e de PLN especialmente dedicadas ao PB.

Ambos os *corpora* são tratados linguística e computacionalmente com vistas à identificação das melhores metodologias de representação automática do seu conteúdo e à

[2] Como exemplo de um estudo bibliométrico, ver SANTIN *et al* (2015). Para trabalhos de PLN, entre inúmeros, sugere-se ver LOPES *et. al* (2009). Na LC, cabe citar Di-Felippo (2013). Hoje em dia, especialmente na área de mineração de dados e de recuperação de informação (subáreas da Ciência da Computação), há inúmeros trabalhos que se ocupam de organizar, reunir ou modelar o conteúdo disperso de trabalhos publicados em periódicos de ciências da saúde disponíveis na internet.

[3] Um trabalho a ser citado sobre a linguagem e as terminologias da Linguística, que contempla o PB, é Fromm e Yamamoto (2013).

sistematização, com apoio informatizado, de sua informação lexical, terminológica e textual. De acordo com os princípios da LC, tal como apresentada no Brasil por Berber Sardinha (2004), os *corpora* sob estudo devem ser processados e comparados com outros, na sistemática *corpus* de estudo *versus corpus* de referência, na proporção de 1 para 3 ou 5 vezes o número de palavras de cada um. Essa comparação estatística visa destacar as *keywords* ou palavras-chave, específicas de um *corpus* (*corpus* de estudo) em relação a um *corpus* genérico (*corpus* de referência). Essa metodologia de LC está detalhada no artigo didático de Kader e Richter (2013), seguindo indicações de Berber Sardinha (1999).

Pela ótica do PLN[4], para esse mesmo fim, podem ser adotados métodos estatísticos distintos e técnicas variadas. Para identificação de palavras ou expressões típicas de um dado *corpus*, podem ser feitos, por exemplo. apenas contrastes de *corpora* de uma área de conhecimento e gênero textual com outros *corpora* também da mesma área e gênero textual. E ainda sem se levar em conta apenas a relação de tamanho de dois *corpora* para contrastes, conforme observam Vecchia *et al* (2014) e Ferreira (2012). Um exemplo de comparação de *corpora* de domínios[5] pode ser acessado em http://vhflabs.com.br/nontax/uso.php.

Assim, tomando o texto em português do CLG como um *corpus* de estudo*,* este artigo relata a busca de um método de representação automática de conteúdo através de ferramentas e de técnicas de PLN e de LC que são então comparadas em seus rendimentos. Dar-se-á maior destaque para uma ferramenta de PLN, o ExATOlp - a seguir detalhada, visto que é pouco conhecida entre pesquisadores linguistas. O objetivo do experimento foi extrair termos relacionados a conceitos importantes do CLG que pudessem se enquadrar na condição de "candidatos a termos", conforme Teixeira (2011).

## 2. Metodologias do trabalho com o CLG: PLN e LC

Uma vez submetido ao *parser* PALAVRAS, um etiquetador morfossintático para a língua portuguesa, o *corpus* do CLG - devidamente escaneado, preparado e salvo em formato .txt - que apresenta 7.606 *types* e 73.586 *tokens*, passou pela ferramenta extratora de sintagmas nominais relevantes, denominada ExATOlp (LOPES, 2012). O nome ExATOlp corresponde a

[4] Para detalhes sobre a natureza e escopo do PLN, ver Dias-da-Silva et al. (2007).
[5] *Corpus de domínio* é como, no PLN, são chamados os *corpus* que reúnem textos de uma mesma área de conhecimento e, normalmente, de mesmo gênero textual. Assim, artigos acadêmicos da área de Medicina, por exemplo, pertenceriam a um mesmo *domínio*.

*Extrator Automático de Termos para Ontologias em Língua Portuguesa*, recurso que implementa diversas técnicas de PLN de base linguística e de base estatística, tendo sido desenvolvido pelo grupo de PLN da PUCRS (http://www.inf.pucrs.br/linatural/) para trabalhos de construção automática de ontologias de domínio com o PB. Essas ontologias servem, entre outras coisas, para a representação automática de conteúdo de textos ou de *corpora.*

Infelizmente, ainda não há acesso livre e *online* a esse sistema, que necessita que o *corpus* a ser examinado seja pré-processado com algum tipo de *parser* – assim como também os *corpora* que se tomem para contraste. Em seguida ao processamento pelo ExATOlp, foram geradas listas e gráficos hierarquizados de elementos lexicais do CLG, incluindo-se sintagmas nominais e verbos recorrentemente associados a esses sintagmas.

Em uma segunda etapa, o mesmo *corpus* foi submetido à ferramenta AntConc (ANTONY, 2013), uma ferramenta de acesso livre bastante empregada em trabalhos de LC e pouco utilizada em PLN. Nessa ferramenta, utilizamos as funcionalidades *Wordlist, Clusters* e *Keywords* para obtenção de palavras ou expressões mais típicas do CLG, tendo sido tomado como *corpus* de referência o Lácio-Ref (disponível em http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm). Foram também adotados os procedimentos básicos descritos em trabalhos de LC, como, por exemplo, os procedimentos apontados no trabalho de Teixeira (2011), no qual se fez uma análise do desempenho de extratores automáticos de terminologias em *corpora*, embora não se tenha tratado do AntConc ou do ExATOlp.

## 3. Estudo comparativo

Após serem convertidos do formato .pdf para .txt, os arquivos de texto do CLG foram corrigidos manualmente, para que as unidades lexicais pudessem ser corretamente processadas pelos *softwares*, tanto o AntConc, como o ExATOlp.

Utilizando o AntConc, foram extraídas listas das palavras-chave. O *corpus* utilizado para contraste foi o Lácio-Ref, um *corpus* aberto e de referência do Projeto Lácio-Web, desenvolvido pela FFLCH da USP e pelo NILC-USP, composto de textos de vários gêneros em português brasileiro, tendo como característica serem escritos respeitando a norma culta.

No experimento com a ferramenta ExATOlp, foram extraídas listas dos SNs (sintagmas nominais) mais relevantes do *corpus*, geradas a partir de um processo de extração híbrido, que utiliza heurística de base linguística, para detectar os termos, e cálculos estatísticos, para estimar

a relevância de cada termo para o domínio. Para essa estimativa de relevância, foi utilizado o mesmo *corpus* de contraste utilizado no experimento com a ferramenta AntConc.

Além disso, um aspecto diferenciado da identificação de SNs relevantes feita pelo ExATOlp está na sua representação visual, que conta com diagramas, nuvens e árvores hiperbólicas dinâmicas. É importante observar que, se a partir do AntConc, a proposta é a de extrair palavras-chave de um domínio, com o ExATOlp, a funcionalidade é um pouco diferente, pois foi projetado para extrair SNs que sejam relacionados a conceitos específicos ao domínio do *corpus* de estudo. Acreditamos, no entanto, que se possa estabelecer uma relação estreita entre esses conceitos específicos representados por SNs e as palavras-chave de um *corpus*. Nos dois casos, trata-se de elencar termos ou expressões que designem temáticas de importância maior para um determinado *corpus*.

## 4. Resultados

### 4.1. SN candidatos a termos

A Tabela 1 apresenta os 20 unigramas[6] mais relevantes obtidos através de cada um dos processos. Numa primeira análise das listas, percebe-se que o processo estatístico, representado na tabela pela ferramenta AntConc, apresenta palavras gramaticais e outras que não poderiam ser candidatas a termos, já que não designam conceitos. Essas palavras estão grifadas. Já o processo de metodologia híbrida, representado na tabela pela ferramenta ExATOlp, gera uma lista em que todas as primeiras 20 palavras são termos específicos da área de Linguística. Esse aspecto já aponta para uma confirmação do melhor desempenho, por parte do processo híbrido, de identificar automaticamente a relevância dos termos.



[6] *n- gramas* é a expressão usada para referir itens que são coletadas de textos de um *corpus*. Um n-grama de 1 item é chamado de unigrama, de 2 itens, bigrama e assim por diante.

Tabela 1. Listas dos 20 primeiros unigramas extraídos a partir de processo híbrido (ExATOlp) e a partir de processo estatístico (AntConc) com a funcionalidade *keywords*.

| | **ExATOlp** | **AntConc** |
|---|---|---|
| 1 | Signos | **não** |
| 2 | Linguística | **é** |
| 3 | Latim | **se** |
| 4 | Acento | língua |
| 5 | Plural | **são** |
| 6 | Linguistas | **que** |
| 7 | Francês | **à** |
| 8 | Significação | linguística |
| 9 | Sintagma | palavra |
| 10 | Sânscrito | **uma** |
| 11 | Grego | **etc** |
| 12 | Fonologia | sons |
| 13 | Fonema | francês |
| 14 | Desinência | **só** |
| 15 | Gramáticos | **um** |
| 16 | Surdos | latim |
| 17 | Sílaba | **também** |
| 18 | Adjetivo | **assim** |
| 19 | Genitivo | **por** |
| 20 | Ortografia | signo |

Usando a funcionalidade *stoplist*, da ferramenta AntConc, podemos eliminar palavras que sabemos, de antemão, não serem relevantes, conforme o objetivo do estudo. Assim, aplicamos essa funcionalidade para evitar palavras, como artigos, preposições, numerais, pronomes, conjunções e alguns verbos, conforme a *stoplist* para português do Brasil disponível em http://www.ranks.nl/stopwords/brazilian. Na Tabela 2, então, mostramos a nova lista de palavras do AntConc, agora gerada com uma *stoplist*, novamente em cotejo com os resultados do ExATOlp.

Tabela 2. Listas dos 20 primeiros unigramas extraídos a partir de processo híbrido (ExATOlp) e a partir de processo puramente estatístico (AntConc) com as funcionalidades *keywords* e *stoplist*.

| | **ExATOlp** | **AntConc** |
|---|---|---|
| 1 | signos | língua |
| 2 | linguística | linguística |
| 3 | latim | palavra |
| 4 | acento | sons |
| 5 | plural | francês |
| 6 | linguistas | latim |
| 7 | francês | signo |
| 8 | significação | elementos |
| 9 | sintagma | fatos |
| 10 | sânscrito | unidade |
| 11 | grego | escrita |
| 12 | fonologia | linguagem |
| 13 | fonema | dizer |
| 14 | desinência | princípio |
| 15 | gramáticos | ponto |
| 16 | surdos | analogia |
| 17 | sílaba | ideia |
| 18 | adjetivo | fenômeno |
| 19 | genitivo | relação |
| 20 | ortografia | formas |

Observamos na Tabela 2 que, com a *stoplist*, o AntConc apresenta uma lista de palavras relevantes para o CLG, sem dúvida. Contudo, nem todas específicas da obra ou da área da Linguística, como é o que observamos na lista gerada pelo ExATOlp. Além disso, a lista gerada pelo ExATOlp possui uma ordenação que reflete a relevância dos termos, ou seja, um cálculo que leva em conta tanto a especificidade, quanto a frequência dos termos. Assim, podemos dizer que a precisão do ExATOlp é maior. Além disso, o resultado gerado pelo ExATOlp é totalmente automático, enquanto que, para o AntConc, foi preciso um trabalho parcialmente manual de confeccionar a *stoplist*.

Um segundo aspecto sobre as listas é o fato de que, ainda que todas as palavras da ferramenta híbrida possam ser consideradas pertencentes à Linguística, elas não são, necessariamente, termos relacionados a conceitos importantes do CLG, isso sob o olhar de um especialista em Saussure. É necessário ressaltar que, como sabemos, as palavras têm seu sentido atribuído somente quando em uso, nas relações de oposição e combinação com as outras palavras do texto. Assim, palavras como "arbitrariedade" ou "diacronia" - fundamentais no

CLG - só assumem o seu valor como termos importantes nessa obra quando em relações de combinação.

Consequentemente, para encontrar termos específicos desse domínio, é necessário extrair também os termos compostos. Na Tabela 3 apresentamos os resultados dos principais 20 bi- e trigramas mais relevantes extraídos pelo processo híbrido (ExATOlp), o que, por si só, já é uma vantagem. Observamos que as listas em n-gramas do AntConc dizem respeito à frequência simples de *clusters* e, por isso, não serão colocados aqui em cotejo.

Tabela 3. Listas dos 20 primeiros bigramas e trigramas extraídos a partir do ExATOlp.

| | **ExATOlp** | |
|---|---|---|
| | **Bigramas** | **Trigramas** |
| 1 | imagem acústica | estado de língua |
| 2 | mudanças Fonéticas | arbitrariedade do signo |
| 3 | signos linguísticos | evolução de sons |
| 4 | cadeia falada | fato de gramática |
| 5 | fenômenos fonéticos | sistema de valores |
| 6 | fatos diacrônicos | mecanismo de língua |
| 7 | indo europeu | noção de valor |
| 8 | palavra francesa | contraparte de imagem |
| 9 | impressão acústica | estudo da linguagem |
| 10 | som laríngeo | fatos de língua |
| 11 | igual modo | interior de língua |
| 12 | linguística diacrônica | linha de conta |
| 13 | signo gráfico | objeto da Linguística |
| 14 | sistema linguístico | partida de xadrez |
| 15 | alto alemão | passagem de ar |
| 16 | imagem auditiva | sequência de sons |
| 17 | unidade linguística | valor do termo |
| 18 | alfabeto grego | vida da língua |
| 19 | aparelho vocal | arbitrário do signo |
| 20 | elo implosivo | ciência da língua |



Na listagem de bigramas do ExATOlp, já aparecem mais termos que podem ser reconhecidos como específicos da Linguística saussuriana, como "imagem acústica", "signos linguísticos", "linguística diacrônica" e "sistema linguístico". E, com exceção de "igual modo", todos os bigramas são termos específicos do domínio da Linguística. Nos trigramas, são colocados em relevo outros tantos termos importantes e específicos da teoria de Saussure, como "estado de língua", "arbitrariedade do signo" e "objeto da Linguística". Aparentemente, bi- e trigramas são menos propensos à ambiguidade do que unigramas. A combinação de termos simples em compostos concede especificidade ao conceito, caracterizando o SN como termo.

Além disso, há uma tendência, e essa é uma hipótese a ser confirmada em nossos *corpora*, em trabalhos futuros, de que termos compostos sejam mais frequentes em textos especializados.

A extração automática dos candidatos a termos do CLG proporcionou o delineamento de uma ontologia - conjunto de termos a partir dos quais são especificados, formalmente, conceitos importantes contidos nessa obra. A metodologia híbrida de extração mostrou-se como a mais produtiva, como se pode observar pelas listas, tanto nos unigramas, como nos bigramas e trigramas. Esses termos podem ser visualizados, a partir do ExATOlp, também em nuvens, como mostram as Figuras 1 e 2.

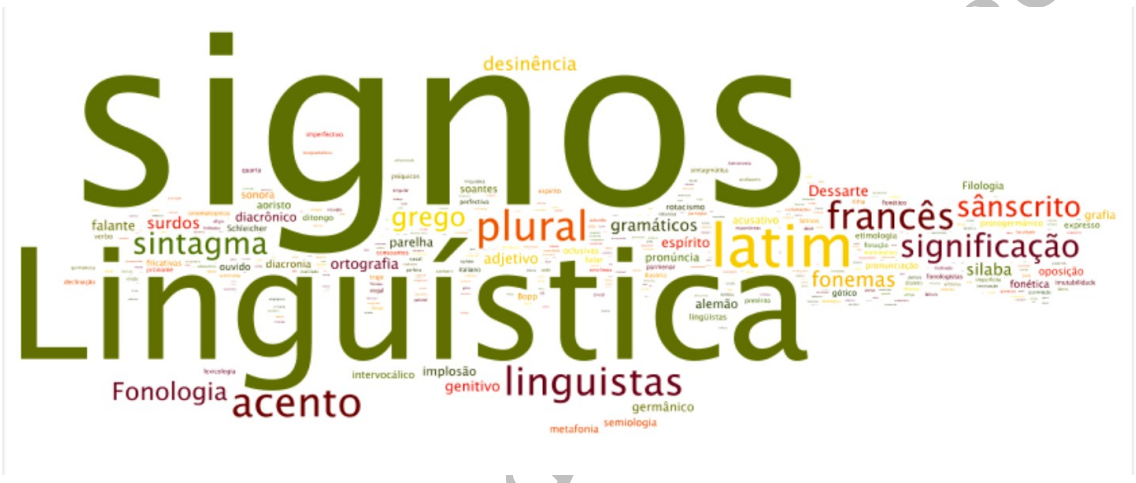


Figura 1. Nuvem de unigramas do CLG extraídos a partir do ExATOlp.

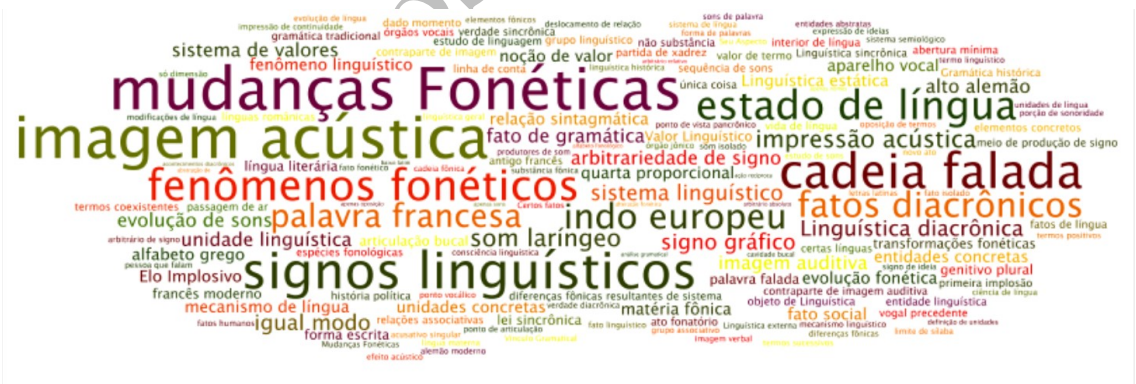


Figura 2. Nuvem de bigramas do CLG extraídos a partir do ExATOlp.

Este estudo com o CLG revela que a aplicação de métodos automáticos para a extração de informação, especialmente os híbridos, com regras linguísticas e estatísticas mais sofisticadas, como o ExATOlp, pode ser muito útil para o trabalho do terminólogo, identificando de maneira rápida itens lexicais que podem ser considerados efetivamente como

termos. Essa mesma tarefa, considerando-se o trabalho humano, demandaria muito mais tempo, especialmente numa obra complexa como o CLG.

Um aspecto que salta aos olhos, em especial nas nuvens, é a grande importância dos termos relacionados à Fonologia sugerida pelo resultado do ExATOlp. Não é surpresa que o tema da Fonologia seja parte da obra, mas mais frequentemente, na literatura, são mencionadas outras questões, como, por exemplo, os estudos sincrônicos e as relações sintagmáticas.

### 4.2.Explorando outras possibilidades do ExATOlp

Posto que a extração de bigramas e trigramas indicou uma qualidade superior do ExATOlp, fizemos a extração de quadrigramas e pentagramas do *corpus* CLG. Na Tabela 4 estão os 20 quadrigramas e 20 pentagramas mais relevantes segundo a extração do ExATOlp. Cabe salientar que a metodologia de extração para estes termos é a mesma empregada anteriormente, ou seja, identificam-se sintagmas nominais, aplicam-se heurísticas linguísticas e calcula-se o índice de relevância.

Tabela 4. Lista dos 20 quadrigramas e 20 pentagramas extraídos do *corpus* CLG com a ferramenta ExATOlp

| | **ExATOlp** | |
|---|---|---|
| | **quadrigramas** | **pentagramas** |
| 1 | contraparte de imagem auditiva | meio de produção de signo |
| 2 | interior de mesma língua | aspectos diversos de mesmo fato |
| 3 | alteração de signos linguísticos | contato com capa de água |
| 4 | aspecto paradoxal de questão | gramática tradicional de francês moderno |
| 5 | cadeia falada em sílabas | imobilidade de latim em época |
| 6 | correspondência exata de valores | leis estranhas a sua função |
| 7 | definição de unidade linguística | papel de esses mesmos órgãos |
| 8 | domínio fechado existente próprio | radical coincidente com oposição gramatical |
| 9 | gramática tradicional de francês | sistema de fonemas de russo |
| 10 | idêntico estado de coisas | vínculo entre ideia e signo |
| 11 | imagem acústica com conceito | acumulação de consoantes em alemão |
| 12 | interior de cada signo | análise de partes de sintagma |
| 13 | jogo de diferenças fônicas | atenção concedida a língua literária |
| 14 | número determinado de letras | ausência de todo suporte material |
| 15 | parte conceitual de valor | bosquejo de sistema de Fonologia |
| 16 | ponto de vista pancrônico | caráter fônico de signo linguístico |
| 17 | próprio de instituição linguística | compostos mais próximos de palavras |
| 18 | questão de aparelho vocal | condições de vida de línguas |
| 19 | sistema de elementos sonoros | condição essencial de signo linguístico |
| 20 | série de palavras análogas | condição mecânica com efeito acústico |

Ao que parece, para o CLG, os quadri- e pentagramas expressam a importância de alguns contextos de termos importantes, mas não exatamente termos relacionados a conceitos importantes.

Contudo, ressaltamos que uma facilidade adicional do ExATOlp é justamente identificar, junto com os termos extraídos, os seus contextos de utilização. Dentre outras informações, o ExATOlp disponibiliza para cada termo extraído o predicado para o qual o termo exerce a função de sujeito ou de objeto. Desta forma, é possível extrair as ocorrências de verbos mais relevantes de um *corpus* e, a exemplo do que é feito para os termos, é possível utilizar *corpora* contrastantes que permitam associar um índice de relevância de cada predicado para o *corpus* de estudo. Aplicando esta funcionalidade da ferramenta ExATOlp ao *corpus* CLG, identificamos os 30 predicados mais relevantes, conforme apresenta a Tabela 5. Note-se que, devido ao uso da técnica de *corpora* contrastante, foi possível filtrar predicados muito frequentes, mas que não são particularmente significativos para o *corpus* CLG, como é o caso dos verbos "ser", "ter", "poder" e "estar".

Tabela 5. Os 30 predicados mais relevantes do *corpus* CLG segundo extração feita pelo ExATOlp

| | **predicados relevantes** |
|---|---|
| 1 | exprimir |
| 2 | intervir |
| 3 | pronunciar |
| 4 | apagar |
| 5 | evocar |
| 6 | assinalar |
| 7 | suscitar |
| 8 | confundir |
| 9 | supor |
| 10 | transtornar |
| 11 | designar |
| 12 | unir |
| 13 | afrouxar |
| 14 | ater |
| 15 | negligenciar |
| 16 | poder empregar |
| 17 | poder ser chamar |
| 18 | recobrir |
| 19 | atribuir |
| 20 | acarretar |
| 21 | repousar |
| 22 | dever trazer |

| | |
|---|---|
| 23 | parecer favorecer |
| 24 | suceder |
| 25 | equiparar |
| 26 | reger |
| 27 | continuar ser |
| 28 | suprimir |
| 29 | distinguir |
| 30 | precipitar |

No entanto, como a extração é feita por predicados, há predicados que contém estes verbos com a função de auxiliar, como é o caso do décimo-sétimo predicado mais relevante "poder ser chamar", onde os verbos "poder" e "ser" são utilizados como auxiliares, por exemplo na expressão "pode ser chamado". De toda maneira, os predicados apresentados na Tabela 5 deixam clara a importância de verbos como "exprimir", "intervir" e "pronunciar", que são ao mesmo tempo abundantes no *corpus* CLG e consideravelmente menos frequentes nos *corpora* de contraste empregados. A Figura 3 ilustra a relevância dos verbos extraídos através de uma nuvem de predicados, em que o tamanho de cada predicado é proporcional à sua relevância.

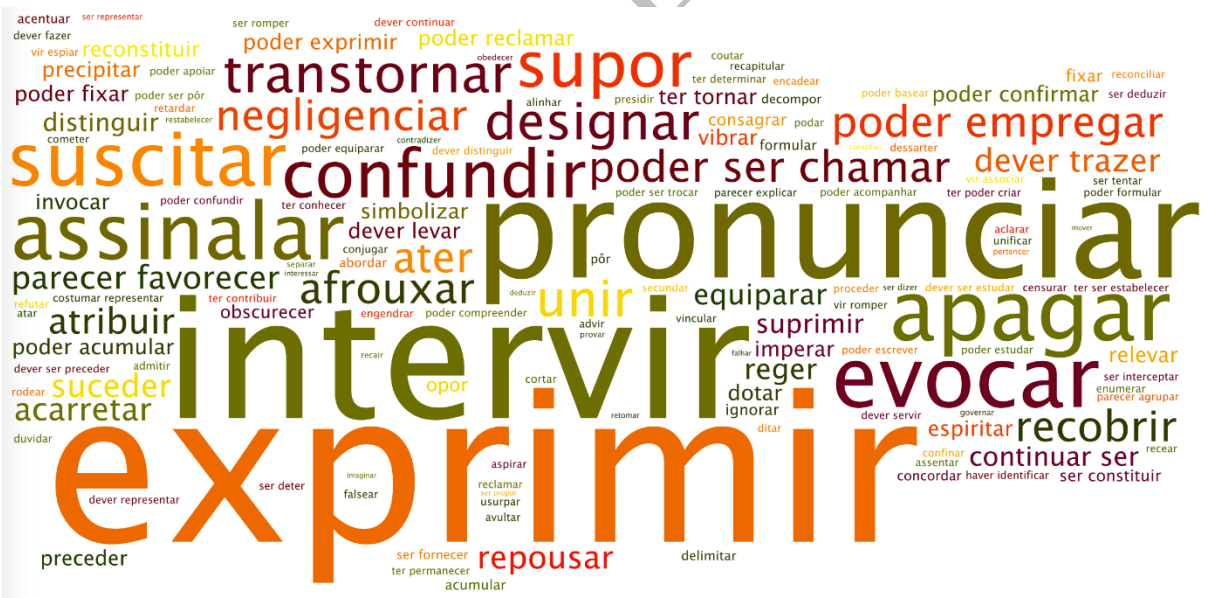


Figura 3. Predicados mais relevantes do *corpus* CLG segundo extração feita pelo ExATOlp

Os verbos, além do papel essencial na gramática das línguas, oferecem, via predicação, elementos para a representação de diferentes relações e podem corresponder a termos ou conceitos de um domínio, como podemos especular, aqui, ao visualizar as nuvens. Assim, o

estudo e a descrição dos verbos podem ser importantes também para subsidiar uma série de recursos de representação e recuperação de informação.

## 5. Considerações finais e trabalhos futuros

Na comparação do desempenho das duas ferramentas de extração automática de candidatos a termos, quais sejam, o AntConc e o ExATOlp, concluímos que ambas auxiliam no trabalho de levantamento de dados lexicais para estudos terminológicos. Contudo, o ExATOlp traz a vantagem de realizar todo o trabalho automaticamente, além de, pelo menos para o CLG, apresentar candidatos a termos que são mais específicos da teoria saussuriana. Além disso, o ExATOlp lista também automaticamente os bi- e trigramas mais relevantes do *corpus* de estudo, enquanto que o AntConc, mesmo com o auxílio de *stoplists*, lista os bi- e trigramas apenas dos *clusters* mais frequentes. E é justamente nos bi- e trigramas que encontramos os termos mais específicos dos temas tratados pelo CLG.

Observamos também a grande relevância, que pode ser facilmente visualizada nas nuvens geradas pelo ExATOlp, de termos relativos à Fonologia no CLG. Tal resultado sugere uma maior importância do tema em Saussure, que, no entanto, muitas vezes fica em segundo plano, cedendo lugar à discussão de outras ideias saussurianas, como as dicotomias e a arbitrariedade do signo. Assim, outra sugestão de estudos futuros é a de investigar mais a fundo os conceitos de Saussure sobre Fonologia, a partir dos principais candidatos a termos elencados pelo ExATOlp.

Outra pesquisa que se apresenta como sugestão, a partir dos resultados da extração automática de termos, é sobre as recategorizações, ou seja, com que termos e tipos de construções os conceitos são designados e retomados anaforicamente e quais as consequências dessas escolhas. Ainda que se trate de um trabalho mais voltado para os estudos linguísticos do texto, ele pode ser útil no sentido de aperfeiçoar sistemas de extração automática e a análise de seus resultados.

## Referências bibliográficas

ANTHONY, L. (2013) **AntConc** (Version 3.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Disponível em : http://www.antlab.sci.waseda.ac.jp/

BERBER SARDINHA, T. UsingKeyWords in text analysis: Practical aspects. **DIRECT Papers**, 42. LAEL, Catholic University of Sao Paulo, Brazil / AELSU, University of Liverpool, England, 1999. Disponível em : www.direct.f2s.com

_____.**Linguística de *Corpus***. Barueri: Manole, 2004.

BUI, Q-C.; SLOOT, P.M.A. Extracting biological events from text using simple syntactic patterns. In: BioNLP Shared Task 2011 Workshop. **Proceedings of** BioNLP Shared Task, Association for Computational Linguistics, 2011, pp. 143–146.

CIULLA, A.; FINATO, M. J. B. O CLG e sua tradução para o português brasileiro - algumas questões sobre a reconstrução da noção de signo linguístico. **Revista Traduzires**. Brasília: Editora da UnB, número especial em homenagem ao centenário de Saussure, 2013, v.2, n.1.

DIAS-DA-SILVA, B.C.; MONTILHA, G.; RINO, L.H.M.; SPECIA, L.; NUNES, M.G.V.; OLIVEIRA JR., O.N.; MARTINS, R.T.; PARDO, T.A.S. Introdução ao Processamento das Línguas Naturais e Algumas Aplicações. Série de **Relatórios do NILC**, NILC-TR-07-10. São Carlos, SP, agosto, 2007, 121p.

DI FELIPPO, A. Extração automática de termos a partir de *corpus* e sua validação para a construção de *Wordnets* terminológicas em português do Brasil. In: Stella Tagnin; Cleci Bevilacqua. (Org.). ***Corpora* na Terminologia**. 1ed. São Paulo: HUB Editorial, 2013, v. 1, p. 63-86.

FERREIRA, V. H. **Uma proposta para descoberta automática de relações não-taxonômicas a partir de *corpus* em língua portuguesa**. 86 f. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Informática, PUCRS. Porto Alegre, RS, 2012.

FROMM, G.; YAMAMOTO, M.I. Terminologia, Terminografia, Tradução e Linguística de *Corpus*: a criação de um vocabulário bilíngue sobre Linguística. In: Stella Tagnin; Cleci Bevilacqua. (Org.). ***Corpora* na Terminologia**. 1ed. São Paulo: HUB Editorial, 2013, v. 1, p. 129-152.

KADER, C.C.; RICHTER, M.G. Linguística de *corpus*: possibilidades e avanços**.** **Instrumento**: Revista de Estudo e. Pesquisa em. Educação. Juiz de Fora: Universidade de Juiz de Fora, v. 15, n. 1, jan./jun. 2013. p.13-23. Disponível em: http://instrumento.ufjf.emnuvens.com.br/revistainstrumento/article/download/2641/1903

LOPES, L. *et al.* Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, [S.l.], v. 3, n. 1, mar. 2009. ISSN 1981-6278. crossref http://dx.doi.org/10.3395/reciis.v3i1.244pt

LOPES, L. **Extração automática de conceitos a partir de textos em língua portuguesa**. 2012. 113f. Tese (Doutorado em Ciência da Computação) - Faculdade de Informática, PUCRS, Porto Alegre, RS, 2012.

SAUSSURE, F. **Curso de Linguística Geral**. São Paulo: Cultrix, 1975.

SANTIN, D. M.; NUNEZ, ZIZIL, A. G.; MOURA, A. M. M. de. Produção científica brasileira em células-tronco nos anos 2000 a 2013: características e colaboração internacional. **Revista**

**Eletrônica de Comunicação, Informação & Inovação em Saúde**, [S.l.], v. 9, n. 2, jun. 2015. ISSN 1981-6278. Disponível em: http://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/965.

TEIXEIRA, R. de B. S. e. Análise do desempenho de extratores automáticos de candidatos a termos: proposta metodológica para tratamento de filtragem dos dados. **Tradterm**, [S.l.], v. 18, p. 297-319, dez. 2011. ISSN 2317-9511. Disponível em: http://www.revistas.usp.br/tradterm/article/view/36765.

VECCHIA, A. D. ; WILKENS, R. ; BOITO, M. Z. ; PADRO, M. ; VILLAVICENCIO, A. Size does not matter. Frequency does. A study of features for measuring lexical complexity. In: 14th Ibero-American Conference on Artificial Intelligence, 2014, Santiago. **Proceedings of the 14th Ibero-American Conference on Artificial Intelligence**, 2014. v. 1.

Artigo recebido em: 30.03.2015
Artigo aprovado em: 22.06.2015