Calidoscópio
Vol. 4, n. 3 , p. 179-188, set/dez 2006
© 2006 by Unisinos

Claudia Oliveira
cmaria@de9.ime.eb.br

Maria Claudia de Freitas
claudiaf@let.puc-rio.br

Classes de palavras e etiquetagem na Lingüística Computacional

Part of speech and tagging in Computational Linguistics

Resumo - A categorização da palavra de acordo com traços que a posicionam dentro do sistema lingüístico é um elemento formal subjacente a qualquer descrição gramatical. Na Lingüística Computacional, *etiquetagem* consiste na atribuição de categorias a porções do texto. O objetivo desse artigo é discutir, no contexto da Lingüística Computacional, a procedência da informação lingüística nos conjuntos de etiquetas de POS – do inglês *part of speech*. Ao longo da discussão evidenciamos a relevância da participação do lingüista na compilação teoricamente bem fundamentada dos conjuntos de etiquetas da prática do Processamento de Linguagem Natural (PLN). Direcionamos nosso olhar, especificamente, para fenômenos relacionados à anotação por classes de palavras, mas que têm recebido um tratamento secundário por parte da lingüística - como as formas nominais do verbo, notadamente o particípio, as palavras denotativas e o aposto.

Palavras-chave: conjunto de etiquetas, particípio, aposto, palavras denotativas, lingüística computacional, PLN.

Abstract: The categorization of words according to features that determine the position they occupy in the language system is a formal requirement of any grammatical description. In Computational Linguistics, tagging is the assignment of categories to portions of a text. The objective of this paper is to discuss, in the context of Computational Linguistics, the source of linguistic information in POS tagging – part of speech, in English. As we present a critical view of this process, it becomes clear that the linguist has a very relevant part to play in the elaboration theoretically sound tagsets for Natural Language Processing. We focus, in particular, three part of speech related language phenomena that have notoriously been overlooked in linguistic studies: the participle verb form, the denotative words, and the appositive.

Key words: tagset, participle, appositive, denotative words, computational linguistics, NLP.

Introdução

A categorização da palavra de acordo com traços que a posicionam dentro do sistema lingüístico é um elemento formal subjacente a qualquer descrição gramatical. Na Lingüística Computacional, *etiquetagem* consiste na atribuição de categorias a porções do texto. Especificamente, na etiquetagem morfossintática, a palavra é a unidade textual considerada. Na prática, porém, o que isso significa? Quais são as categorias utilizadas?

O objetivo desse artigo é discutir, no contexto da Lingüística Computacional, a procedência da informação lingüística nos conjuntos de etiquetas de POS – do inglês *part of speech* – deixando à mostra a multiplicidade de critérios empregados na escolha das categorias representadas. Ao longo da discussão evidenciamos a relevância da participação do lingüista na compilação teoricamente bem fundamentada dos conjuntos de etiquetas da prática do Processamento de Linguagem Natural (PLN).

De maneira geral, a informação lingüística classificada em um sistema de PLN pode ser produzida de duas formas distintas. Uma abordagem empiricista assume que as estruturas da linguagem podem ser adquiridas a partir de modelos gerais, que seriam instanciados indutivamente por técnicas computacionais de casamento de padrões ou processamento estatístico (Manning e Schütze, 1999), por meio da utilização de grandes coleções de textos, que exemplificam a língua-alvo em uso.

Em contraste, uma abordagem racionalista privilegia a elaboração completa da gramática a partir do conhecimento do lingüista, freqüentemente usando esse conhecimento para a produção dos mecanismos

de inferência necessários à compreensão e produção da língua.

A etiquetagem de POS pode ser entendida como um processo de PLN que utiliza duas fontes de conhecimento lingüístico: (i) um léxico de palavras e de processos, categorizado pelo conjunto de etiquetas do sistema; e (ii) o conjunto de possibilidades sintagmáticas para as categorias presentes no léxico. O léxico é quase sempre construído segundo a abordagem racionalista. Por outro lado, as possibilidades de arranjo das categorias nos enunciados, que podem ser representadas por modelos simbólicos de regras ou modelos probabilísticos, vêm cada vez mais, e com maior precisão, sendo construídas empiricamente.

Tomando como exemplo a frase 1.a, o processo de etiquetagem de POS pode ser descrito da seguinte maneira: de um lado está o léxico, que contém o item "o", categorizado como pronome oblíquo ou como artigo; "canto", categorizado como verbo ou substantivo; "de", categorizado como preposição, e assim por diante.

1.a) *O canto da sala está sujo*

Devido às possíveis ambigüidades, é necessário verificar, de acordo com regras ou probabilidades, se as seqüências "artigo + substantivo", artigo + verbo", "pronome oblíquo + substantivo", "pronome oblíquo + verbo" são plausíveis na língua.

A relação entre o léxico e as seqüências de classes palavras (as seqüências de POS) é feita justamente pelo conjunto de etiquetas de POS. É fundamental ajustar esse conjunto de etiquetas para que ele permita a possibilidade de representação de seqüências plausíveis, descartando as seqüências não plausíveis.

Consideremos, por exemplo, um conjunto de etiquetas com uma única categoria para todos os verbos (etiqueta V), além das etiquetas de substantivo (S), adjetivo (ADJ), advérbio (ADV) e artigo (ART). Segundo este conjunto, a frase 1.b) seria etiquetada pela seqüência V / V / ADV. Consideremos agora a frase 1.c). Se a seqüência V / V for considerada plausível, então a frase tem grandes chances de receber as etiquetas ART / N / V / V / ART / N, contrariando o fato de que *sujo* é primariamente um adjetivo.

1. b) *Tenho lido muito.*
1. c) *A menina deixou sujo o chão*

Uma solução simples é o desmembramento da categoria verbo (V) em duas categorias, verbo (V) e verbo auxiliar (VAUX), e o refinamento das combinações sintagmáticas para registrar que assumir que a combinação V / V não é plausível, mas VAUX / V é.

Esse pequeno exemplo ilustra o uso de uma categoria tradicional da gramática, a dos verbos auxiliares, na ampliação de um conjunto de etiquetas inadequado. A inferência de conhecimentos lingüísticos em outros níveis, a partir de seqüências sintagmáticas de POS está intimamente vinculada ao projeto do conjunto de etiquetas.

O restante do artigo está organizado da seguinte maneira: na próxima seção, são discutidos os principais aspectos da delimitação das chamadas *classes de palavras*, trazendo uma apresentação geral do tema a respeito do qual os pontos de vista, embora teoricamente divergentes, terminam gerando propostas de classificação bastante semelhantes. Nas seções seguintes procuramos mostrar como o refinamento com relação a determinadas classes em um conjunto de etiquetas é de grande ajuda na delimitação de sintagmas nominais; na identificação de relações taxonômicas e na classificação semântica de nomes próprios.

Classes de palavras

A separação das unidades lexicais em classes de palavras faz parte da visão aristotélica de linguagem. A vasta maioria dos modelos lexicais utilizou essa idéia. A classificação das palavras em escaninhos rotulados, obedecendo a critérios que lhes atribuem propriedades em comum faz parte de uma tradição gramatical adotada do grego e do latim. É surpreendente, como mostra a Tabela 1, como as categorizações clássicas perduram e podem ser encontradas nas mais modernas descrições gramaticais (Priscianus Charisius, *in* Jurafsky e Martin, 2000; Biderman, 1978).

Rosa (2000) discute extensivamente diversas propostas históricas de categorização das palavras e suas motivações, observando que esses esquemas se mantiveram com poucas alterações significativas. Entretanto, nota que:

> ... a classificação das palavras deixou de basear-se em critérios semânticos e passou a ter por fundamentos critérios distribucionais, funcionais e sua categorização. A diferença de foco está, até certo ponto, refletida na nomenclatura: o uso da expressão *classe de palavra*, em lugar de *parte do discurso*, procura assinalar a ruptura com as noções que norteavam os estudos tradicionais (Rosa, 2000).

Tabela 1. Classes de palavras na tradição grega (esq.) e latina (dir.)

ὄνομα	- nome	nomina	- nome
ἀντωνυία	- pronome	pronomina	- pronome
φῆμα	- verbo	verba	- verbo
ἐπίρρημα	- advérbio	adverbia	- advérbio
μετοχή	- particípios	participia	- particípios
φῦνδεσμος	- conjunções	coniunctiones	- conjunções
πρόθεφις	- preposições	praeposiotiones	- preposições
		interiectiones	- interjeições
ἄφθφον	- artigo		

A utilização do termo "classes de palavras", com apoio em Câmara Jr (2000) e Basilio (1999), traz em conta que é possível adotar três diferenciados critérios na categorização: o critério semântico, o morfológico (pelas categorias flexionais que apresentam) e o sintatico-funcional. Lyons (1977) apresenta, em dois paralelos, as motivações principais para as classes de palavras, como mostra a Tabela 2.

Tabela 2. Motivações para as classes de palavras

aspecto semântico-ontológico	aspecto sintático	aspecto gramatical
sujeitos	entidades	substantivos
predicados	qualidades	adjetivos
predicadores	ações	verbos

O critério semântico ou nocional é o motivador mais tradicional para a classificação das palavras. Grosso modo, os *substantivos* designam as pessoas, os objetos ou as situações, os *verbos* designam processo, os *adjetivos* designam as qualidades dos substantivos e os *advérbios* designam as qualidades dos verbos ou dos adjetivos (Dubois *et al.*, 2001). As *preposições* e *conjunções* indicam relações lógicas entre outros elementos do discurso, os *artigos* determinam os substantivos e os *pronomes* os substituem.

Essa tradição da categorização nocional é duramente criticada por formalistas, que consideram a abordagem imprecisa e não preditiva, além de não se mostrar apropriada para expressar generalizações gramaticais que não se adequem aos limites das classes de palavras existentes. Por outro lado, Lyons (1977), mesmo acreditando na possibilidade de formulação de "procedimentos diagnósticos" de cunho sintático para o traçado preciso e definitivo de uma classificação vocabular, afirma que o interesse nesses procedimentos deve ser motivado pelo estabelecimento de propriedades semânticas das classes resultantes. Além disso, Lyons observa que:

> The fact that there appears to be a positive correlation in all languages between syntactically defined and semantically defined expression-classes would tend to support the traditional view that there is a high degree of interdependence between the syntactic structure of sentence-nuclei and the semantic function of their constituent expressions. Despite what has been said at times by certain linguists there is no reason to doubt that the traditional view is, to this extent at least, well-founded.

Apesar da procedência de uma das principais críticas às classificações tradicionais – a de que são inaplicáveis, em todos os seus detalhes, a línguas cujas estruturas gramaticais diferem significativamente das línguas indo-européias – existe uma quase unanimidade com respeito à distinção entre verbo e substantivo. Os verbos e os substantivos são classes tradicionalmente destacadas como as essenciais na gramática. Parecem ser classes indispensáveis na enunciação, possuindo realidade psicolingüística e tendo sido identificadas universalmente em extensos estudos trans-lingüísticos (Rosa, 2000). Sapir, um conhecedor de numerosas e variadas línguas norte-americanas, reafirma a essencialidade dos verbos e nomes na seguinte passagem (Sapir, 1921, p. 117):

> There must be something to talk about and something must be said about this subject of discourse once it is selected. [...] The subject of discourse is a noun. As the most common subject of discourse is either a person or a thing, the noun clusters about concrete concepts of that order. As the thing predicated of a subject is generally an activity in the widest sense of the word, a passage from one moment of existence to another, the form which has been set aside for the business of predicating, in other words, the verb, clusters about concepts of activity. No language wholly fails to distiguish noun and verb [...].

Mesmo que o texto de Sapir não seja formalmente preciso (Lyons [1977] observa que deve haver uma distinção entre "nome" e "expressão nominal", por exemplo), o autor consegue estabelecer o paralelo entre as classes verbal e nominal e os processos de predicação e referenciação, respectivamente.

Diversos autores estruturalistas, mais notadamente Bloomfield (1966), rejeitaram a universalidade das categorias dos nomes e dos verbos (ou de qualquer outra categoria). No entanto, contemporaneamente, o consenso a esse respeito tem sido restabelecido principalmente por estudos da psicologia cognitiva e da psicolingüística (Laudanna e Voghera, 2002).

O conjunto das classes de palavras pode ser ainda subdividido para refletir determinadas propriedades das classes. Rosa (2000) discute a distinção entre classes de palavras com *significado lexical* e com *significado gramatical*. Se por um lado é problemático admitir que algumas palavras isoladamente são portadoras de significado, e não outras, por outro lado, há variantes a essa oposição que terminam subdividindo as classes de palavras de maneira semelhante: palavra de conteúdo vs. palavra de forma, palavra lexical vs. palavra gramatical, palavra plena vs. palavra vazia, contentivo vs. functor, vocábulo forma vs. vocábulo conectivo (Câmara Jr., 2000). Um outro critério de subdivisão do conjunto das classes de palavras é quanto ao potencial de gerar vocabulário: as *classes abertas* e as *classes fechadas*.

A atribuição de classes de palavras é fundamental para muitas das aplicações do Processamento de Linguagem Natural (PLN), tais como *parsing* sintático e semântico, análise do discurso e processamento de fala. Para atender à necessidade computacional de lidar com classes menos ambíguas, ou seja, com menor interseção entre seus membros, os critérios utilizados para a categorização tendem a se basear em distribuição sintática e morfologia.

A precisão da classificação também é um requisito computacional importante, por isso recursos computacionais tais como bases de dados lexicais e corpora anotados com etiquetas de POS tendem a utilizar classes de maior granularidade. Por exemplo, o Penn Treebank (Marcus *et al.*, 1993) utiliza 45 classes, incluindo, junto com as clássicas, rótulos diferenciados para palavras estrangeiras, nomes próprios, símbolos fora do alfabeto, entre outros. Os verbos são também rotulados quanto à morfologia da ocorrência (forma básica, passado, gerúndio, particípio e terceira pessoa do singular), assim como os substantivos (singular e plural). Outros conjuntos chegam a ter centenas de classes.

Para o português, podemos mencionar quatro conjuntos de etiquetas de POS distintos: o conjunto do etiquetador AnELL[1], do etiquetador do LAEL[2], do projeto Lacio-Web[3] e do parser PALAVRAS[4]. Com relação às classes de palavras, o conjunto do etiquetador AnELL contém 12 etiquetas distintas, além de quatro etiquetas para contrações; o do etiquetador do LAEL contém 15 etiquetas; o do Lácio-Web conta com 24 etiquetas, além de sete etiquetas complementares; e o PALAVRAS utiliza 18 etiquetas para as classes de palavras.

O particípio e a delimitação do sintagma nominal

O particípio é tradicionalmente considerado uma das formas nominais do verbo. Conforme Câmara Jr. (2000), o particípio foge do padrão mórfico verbal por admitir as marcas nominais de gênero e número, o que o torna, segundo o autor, "um nome adjetivo, que semanticamente expressa em vez da qualidade de um ser, um processo que nele se passa".

O processo de formação de adjetivos deverbais V-do pode confundir-se com a formação verbal dos particípios. Pimenta-Bueno (1986) utiliza a denominação "forma deverbal [V+do]" para caracterizar a não-uniformidade do comportamento dessas formas quanto ao seu caráter verbal ou nominal. Há uma série de propriedades do particípio que os aproxima dos adjetivos e os difere de verbos, tais como: posição exclusivamente predicativa dentro do enunciado, ocorrência em expressões comparativas, existência de formas superlativas, ocorrência com modificadores de grau, possibilidade de coordenação com outros adjetivos, flexão de gênero e número.

Dada a sua natureza predicativa, adjetivos podem semanticamente corresponder-se a predicados n-argumentais. Isto se traduz sintaticamente em complementos, muito embora não se possa atribuir-lhes a mesma variedade e riqueza da complementação verbal.

Em um trabalho propondo a delimitação automática de sintagmas nominais (SN), baseada em regras extraídas por aprendizado de máquina, Freitas *et al.* (2005) apresentam uma definição formal de um modelo de SN que busca ser lingüisticamente motivado, ao mesmo tempo atendendo a requisitos computacionais e empíricos. As duas principais motivações para o modelo são: i) estender o conceito de *chunk* de SN (Abney, 1991) para abranger um conjunto mais significativo de SNs no português e ii) restringir esse mesmo conjunto para incluir apenas SNs lexicais – aqueles cujo núcleo é uma palavra lexical. O trabalho situa-se na área de Terminologia e Lexicografia Computacional, por isso a importância dos SNs lexicais, em detrimento dos "mais gerais" – cujo núcleo pode ser pronominal ou elíptico.

O modelo de Freitas *et al.* está estruturado como *núcleo (+complementos) (+especificadores)*, onde o núcleo é um substantivo, de acordo com as seguintes restrições:

(1) o núcleo é um único substantivo, não elíptico;
(2) os complementos são sintagmas adjetivais e preposicionais; orações relativas e apostos são descartados;
(3) os especificadores são artigos, pronomes demonstrativos, possessivos e indefinidos, ou quantificadores, como numerais;
(4) a continuidade da seqüência de palavras dentro do SN é exigida;
(5) o particípio é sempre considerado como um adjetivo exceto quando ocorrer com um verbo auxiliar ("o barco abandonado" vs "o barco foi abandonado").

O item 5 tem inspiração no trabalho de Pimenta-Bueno (1986), que propõe a categorização dos particípios de acordo com as seguintes hipóteses: nos contextos posteriores aos verbos "ter" e "haver" eles são verbos; em todos os outros contextos são adjetivos, exceto nos contextos V N ("Zé foi nomeado síndico") e V Adj ("Zé foi declarado incompetente"). Nesses casos, a autora classifica a forma V+do como *particípio passivo (PP)*. Tendo em vista essa análise, Freitas *et al*. (2005) consideram o particípio como forma verbal ou PP sempre que estiver em contextos posteriores a verbos auxiliares ou de ligação, como em 3.a. Nos outros casos, o particípio é um adjetivo e pode ser complemento, como em 3.b e 3.c.

(3.a) dois soldados israelenses foram atingidos por tiros
(3.b) as pequenas agremiações formadas a partir da divisão do PLD
(3.c) o diálogo contrasta com o apoio dado pelo governo

[1] http://acdc.linguateca.pt/AnELL, acessado em 25/10/2006.
[2] http://lael.pucsp.br/corpora/etiquetagem, acessado em 25/10/2006.
[3] http://nilc.icmc.sc.usp.br/lacioweb/, acessado em 25/10/2006.
[4] http://visl.sdu.dk/visl/pt, acessado em 25/10/2006.

O aprendizado automático de regras que levam à identificação dos SNs do modelo foi feito sobre o corpus Mac-Morpho (Marchi, 2003), etiquetado morfossintaticamente com o conjunto de etiquetas do projeto Lacio-Web. A etiqueta utilizada para qualquer forma V+do é PCP, indistintamente, para formas com comportamento adjetivo, verbal ou em casos ambíguos. Portanto, o processo de aprendizado automático não poderia gerar regras que resolvessem esse impasse.

No trabalho de Oliveira *et al.* (2006), os resultados de Freitas *et al*. (2005) são analisados de forma detalhada, tanto em termos da eficácia do sistema na classificação do que é SN, quanto com relação à complexidade informacional dos SNs extraídos. Observando tais SNs, como por exemplo 3.d, 3.e e 3.f, percebemos a existência de uma estrutura verbal embarcada em SNs que contêm particípios.

(3.d) as conversações patrocinadas pela ONU no Zâmbia
(3.e) os números tirados do bolso do colete sobre o peso dos encargos no Brasil
(3.f) as pessoas postadas diante de caixas enviadas pelo Banco Central

Dessa forma, a lógica do modelo de desconsiderar orações encaixadas como complementos do SN fica comprometida, pois os exemplos acima podem ser interpretados como contendo orações reduzidas de particípio. Concluímos, então, que a distinção entre verbo e adjetivo no conjunto dos particípios não foi bem definida, como conseqüência das limitações do conjunto de etiquetas morfossintáticas, o que aponta para um aprofundamento do estudo de classes de palavras, principalmente na fronteira entre verbos e adjetivos.

A palavra denotativa e a extração de relações taxonômicas

Palavras denotativas constituem uma classe "marginal" no contexto de classes de palavras. Chamadas por Kury (1960) de "palavras de difícil designação", ocupam, nas gramáticas de Cunha e Cintra (1985) e Bechara (1999), uma pequena seção dentro da classe dos advérbios. Uma descrição mais detalhada das denotativas aparece apenas em Oiticica (1940 apud Pereira, 1995), que propôs a denominação "palavras denotativas" e elencou 17 tipos. A Nomenclatura Gramatical Brasileira considera *denotadores* um grupo à parte, heterogêneo, que coincide parcialmente com a proposta de Oiticica. Tais *denotadores* podem ser de inclusão, exclusão, explicação, situação, retificação, designação, realce, etc. Pereira (1995) já aponta para a polêmica suscitada pela classe das denotativas, que ora são colocadas à parte, ora incluídas entre os advérbios, e ora não são sequer mencionadas.

Para Oiticica, as denotativas são acidentes do discurso, que se prendem a um elemento qualquer do enunciado, e "indicam certos movimentos ou operações subjetivas e indispensáveis à compreensão do pensamento". Porém, como assinala Pereira em sua reflexão sobre as palavras denotativas, a proposta de Oiticica, embora interessante e prática, é demasiado extensa, com uma nomenclatura tão variada que dificulta a memorização dos diferentes tipos de palavras denotativas. Além disso, alerta a autora, levando-se em consideração aspectos como significação, variações da língua em determinadas situações, lógica e contexto, por exemplo, a lista de tipos de palavras denotativas pode ser enorme, correndo-se o sério risco de resvalar para o "achismo" na classificação. Pereira sugere, então, que, entre (i) reconhecer tais palavras simplesmente como palavras denotativas e (ii) incluí-las em classes já existentes, como advérbios, conjunções, preposições etc, a primeira opção parece mais coerente, uma vez que há, na língua, diversas palavras cuja classificação pode variar conforme o emprego. Palavras denotativas são um recurso que a língua oferece, e por isso devem ter status próprio, sendo desnecessário o estabelecimento de uma classificação granular do tipo "*denotativa de...*". Uma vez assumida autonomia da classe das denotativas, o passo seguinte é o estabelecimento de critérios específicos para sua definição.

Concordamos com Pereira quanto à necessidade de classificação à parte das denotativas. Neste ponto, alguns trabalhos em Lingüística Computacional vêm corroborar a importância desta classe. De fato, a LC – principalmente a baseada em corpus – por tratar do processamento da linguagem em situações "naturais", não pode se dar ao luxo de lidar apenas com fenômenos de "primeira classe": a pouca importância dada às palavras denotativas se reflete na inconsistência de etiquetas correspondentes nos etiquetadores automáticos. Considerando os conjuntos de etiquetas para o português, vemos que o conjunto do projeto Lácio-Web é o único que oferece a etiqueta PDEN para a categoria das palavras denotativas.

Porém, como exemplificaremos adiante, etiquetas PDEN não são necessárias apenas por representarem de maneira mais fiel determinados fenômenos lingüísticos, mas também por possibilitarem – no caso específico da palavra "como", que trataremos a seguir – uma alta precisão na identificação automática de relações taxonômicas.

A identificação de relações semânticas entre termos é de grande utilidade no PLN. A aquisição de informação lexical é de grande valia para tarefas de extração automática de informação, construção automática de ontologias e elaboração de dicionários e léxicos semânticos (Riloff e Shepherd, 1997), por exemplo. Dentre as relações semânticas usualmente tratadas, destaca-se a relação de hiperonímia/hiponímia, pois é responsável por organizar os termos em uma taxonomia – forma relativamente simples de compreensão da informação.

Um termo X é considerado hipônimo de um outro termo Y se o significado de X inclui o significado de Y e,

de modo inverso, Y é considerado hiperônimo de X. Por exemplo, *cachorro* é um hipônimo de *animal*, e *animal* é hiperônimo de *cachorro*.

O trabalho de Hearst (1998) é pioneiro na tentativa de identificação automática de relação de hiperonímia em textos. A autora propõe uma série de padrões léxico-sintáticos, listados na Tabela 3, que expressariam tais relações.

Tabela 3. Padrões léxico-sintáticos de Hearst

(i)	NP_0 such as NP_1 {, NP_2 ... , (and | or) NP_i}
(ii)	such NP_0 as {NP ,}* {(and | or)} NP
(iii)	NP {, NP}* {,} or other NP_0
(iv)	NP {, NP}* {,} and other NP_0
(v)	NP_0 {,} including { NP ,}* {or | and} NP
(vi)	NP_0 {,} especially { NP ,}* {or | and} NP

O padrão (i) – "such as" – pode ser literalmente traduzido para "tais como". Porém, na língua portuguesa, freqüentemente apenas o "como" é utilizado neste tipo de construção, como ilustram os exemplos 4.a e 4.b.

(4.a) A tentativa posterior de clonar outros mamíferos tais como camundongos, porcos, ...

(4.b) A tentativa posterior de clonar outros mamíferos como camundongos, porcos , ...

Ou seja, para que o padrão revele uma quantidade significativa de relações de hiperonímia no português, é preciso considerar a variante "como". Porém, se há um ganho do ponto de vista da abrangência, uma vez que mais relações podem ser identificadas, do ponto de vista da precisão essa inclusão é um complicador: "como" é uma palavra que se enquadra em diferentes classes gramaticais, dificultando o trabalho dos etiquetadores automáticos e, conseqüentemente, acarretando problemas na identificação do padrão desejado.

Pela gramática tradicional, "como" pode ser advérbio, preposição acidental, pronome relativo ou conjunção. Quando conjunção, pode ser subordinativa – adverbial ou integrante – ou coordenativa. A Tabela 4 ilustra cada um dos casos, porém, o *como* responsável pela expressão da relação de hiperonímia não se encontra em nenhum dos casos exemplificados. Aliás, ele quase não aparece nas gramáticas. Não por acaso, ele também não recebe nenhuma etiqueta especial pelos etiquetadores automáticos.

Na frase 4.c, o "como" foi etiquetado como preposição pelos etiquetadores Brill e TreeTagger (que utilizam o conjunto de etiquetas do Lácio-Web – o único que conta com uma etiqueta específica para a classe das denotativas), como um advérbio pelo etiquetador do LAEL e também pelo PALAVRAS (neste, especificamente, como um advérbio que introduz uma oração elíptica – "*outros mamíferos como [o são] camundongos*...") e como conjunção, interjeição ou verbo pelo etiquetador AnELL.

(4.c) Com a entrada de instrumentos como flauta, bandolim e cavaquinho, estava completa a gestação do chorinho.

Porém, nesse caso, o "como" que interessa na identificação da relação de hiperonímia, pode ser utilizado no lugar (ou acrescido de) "por exemplo":

(4.d) Com a entrada de instrumentos como por exemplo flauta, bandolim...

Em 4.d, trata-se de um "como" que pode ser classificado como uma "palavra denotativa" do mesmo modo que seria a expressão "por exemplo". Ou seja, o "como" palavra denotativa, semelhante a "tais como" e equivalente a "por exemplo", tem chances mínimas (senão nulas) de receber uma etiqueta PDEN.

Conseqüentemente, uma busca pelo padrão "SN como SN" que considera a etiqueta PDEN de "como" provavelmente leva a um alto índice de precisão e, do mesmo modo, a desconsideração da etiqueta leva a inúmeros erros. Uma pista já utilizada por Hearst (1998) para a identificação do "tais como" (no caso do inglês, *such as*) é a presença de coordenação (lista de SNs) após o "tais como". Porém, embora a coordenação seja pista eficaz e prática, pois elimina a dependência de um etiquetador altamente preciso, ela não é suficiente. No exemplo 4.e, te-

Tabela 4. Classes de palavras do "como"

Frase	classe gramatical
Os meninos não sabiam como se proteger da doença	Conj. Sub. Integr.
Como é bastante difícil comprovar os efeitos das campanhas..	Conj. Sub. Adv.
A expectativa de vida tanto em países desenvolvidos como em países em desenvolvimento...	Conj. Coord.
... a doença periodontal têm tido como conseqüência o edentulismo...	Advérbio
... de qualquer modo, cabe aqui uma outra frase como resumo do pensamento de...	Preposição acidental
... verdade no modo como ele interpreta aquela dualidade...	Pronome Relativo

mos uma seqüência de "SN como {lista de SN}" que não corresponde ao padrão desejado:

(4.e) [...] o homem ainda vê a mulher livre como instigante, sedutora, ele deseja...

Além disso, embora pouco freqüentes, as estruturas em que "como" é palavra denotativa, mas vem seguida por um único SN – e não por uma lista – também deixam de ser identificadas quando se considera exclusivamente a pista da coordenação, como no exemplo 4.f.

(4.f) A falta de minerais como o ferro pode causar uma anemia.

A inclusão do padrão "como_PDEN" nos deixa com um problema: por um lado, é altamente confiável como expressão de relação de hiperonímia e muito mais freqüente na língua do que o padrão "tais como" (em um córpus da área de saúde, com aproximadamente 1.800.000 palavras, há cerca de 2700 ocorrências de "como_PDEN" contra apenas 232 ocorrências de "tais como"); por outro lado, o sucesso de sua identificação depende de um fator externo – depende de um etiquetador capaz de reconhecer o "como_PDEN" – o que ainda não existe, a etiquetagem deve ser feita manualmente. Concluímos que há a necessidade da formalização lingüística da classe das palavras denotativas.

O aposto e a classificação semântica de nomes próprios

O aposto é um elemento não classificado pela sintaxe da língua. Talvez por isso – por não ter função sintática específica, repetindo a função do termo a que está relacionado – seja um tema "nebuloso", sobre o qual há poucas pesquisas e muitas opiniões dissonantes. Embora o aposto não seja uma categoria gramatical, mas uma função sintática, o que o elimina do conjunto tradicional de etiquetas de POS, o projeto Lácio-Web disponibiliza a etiqueta do aposto como etiqueta complementar. Além disso, incluímos o aposto neste trabalho por ser mais um exemplo de fenômeno lingüístico ao qual normalmente se atribui pouca importância, mas cuja relevância no PLN é alta.

Em geral, o aposto é definido como um termo de caráter nominal que se junta a outro a título de explicação. Porém, é apenas aí em que há concordância sobre o tema. Após a imprecisa definição, compêndios gramaticais apresentam uma lista exaustiva de tipos de aposto, cuja classificação varia de gramático para gramático.

Cunha e Cintra (1985), por exemplo, descrevem o aposto de especificação, aposto enumerativo, aposto representado por uma oração, aposto que se refere a uma oração inteira e aposto predicativo. Kury (1999) trata do aposto explicativo e enumerativo, do aposto de oração e do aposto de especificação. Bechara (1999) lista os seguintes tipos: explicativo, enumerativo, distributivo, circunstancial, especificativo e aposto em referência a uma oração inteira. À classificação de Bechara (1999), Henriques (2003) acrescenta o aposto resumitivo ou recapitulativo e o aposto distributivo. Said Ali (1964, p. 127) escreve as seguintes linhas sobre o aposto:

> É o termo acessório que se pospõe ao sujeito ou objeto como explicação ou a título de equivalência. Pode ser um simples substantivo ou uma frase de certa extensão:
> Carlos I, rei da Inglaterra, foi decapitado em 1699.
> Renato, amigo nosso, não nos abandonará.
> Matamos a onça, terror das nossas matas.

Ou seja, o único aposto que Said Ali considera é o que outros chamam de explicativo.

Perini (1996) é quem apresenta uma visão diferente sobre o assunto. Para ele, o aposto integra os chamados elementos parentéticos: "elementos que podem posicionar-se livremente entre os constituintes oracionais e que na escrita são sempre separados por vírgula" (1996, p. 120). São exemplo de parentéticos:

(5.a) Creio eu, Dorival dispensou o sócio;
(5.b) Os deputados dão, oferecem de graça, empregos na Assembléia;
(5.c) Simone, irmã de Carlinhos, ganhou um carro novo;
(5.d) Irritado, Dorival dispensou a secretária.

Porém, "irritado", da frase 5.d) pode receber diferentes tratamentos, conforme a gramática utilizada. Assim, em Perini (1996) ele é parentético. Já para Cunha e Cintra (1985), trata-se de aposto predicativo. Para Henriques (2003) é predicativo do sujeito.

Nos interessa aqui o aposto explicativo. Consideramos, especificamente, o que Quirk *et al.* (1985) chamam de aposto "completo, estrito e não-restritivo" ("*full, strict and nonrestrictive apposition*"), isto é, estruturas de aposto que (i) podem ser separadamente omitidas, sem afetar a oração resultante (completo); (ii) pertencem à mesma classe sintática (estrito); (iii) estão em unidades de informação distintas (não-restritivo).

No Processamento de Linguagem Natural (PLN), a identificação de aposto explicativo contribui para a elaboração de léxicos semânticos (Phillips e Riloff, 2002; Caraballo, 1999) resolução de co-referência de sintagma nominal (Cardie e Wagstaff, 1999; Soon *et al.*, 2001; Ng e Cardie, 2002) e extração de informação a partir de textos. Além disso, freqüentemente um dos elementos do par envolvido em uma relação de aposto explicativo é um nome próprio, o que torna a identificação automática de apostos uma tarefa de grande relevância para a área de Reconhecimento de Entidades Mencionadas (REM).

Nomes Próprios costumam ser considerados, pela teoria lingüística, um fenômeno periférico, por não ofere-

cerem contribuições relevantes sobre o funcionamento da estrutura da(s) língua(s). Talvez em conseqüência dessa desvalorização, imagina-se que sua identificação e classificação semântica automática seja uma tarefa simples, o que não corresponde à realidade.

Por outro lado, a compreensão de nome próprio é crucial na análise de textos. Unidades lingüísticas que aparecem com freqüência bastante significativa na língua, a tipologia semântica de nomes próprios varia segundo o gênero de texto em que aparecem. Em artigos científicos da área da Biologia, predominam nomes de espécies, gêneros, proteínas, por exemplo; já em textos sobre História, nomes de pessoas, lugares, povos.

Alguns trabalhos sobre identificação e classificação automática de nomes próprios fazem uso de listas de antropônimos e topônimos, ou de outras bases de conhecimento. Porém, tais listas costumam apresentar limitações, como a custosa elaboração manual, que acarreta em dificuldades de atualização e extensão e, freqüentemente, uma quantidade sempre insuficiente de nomes próprios. O fato de tais nomes constituírem uma classe ainda mais "aberta" do que a dos substantivos comuns, uma vez que novos nomes são criados a todo momento, deixa ainda mais evidente a necessidade de atualização constante e, conseqüentemente, de metodologias capazes acrescentar nomes – e suas classes semânticas – automaticamente.

Nesse contexto, a identificação automática do aposto explicativo pode ser de grande utilidade. Porém, acreditamos que o potencial informativo veiculado pelo aposto ainda não foi totalmente explorado na LC, e uma evidência desse desinteresse é pequena quantidade de trabalhos sobre o assunto (Phillips e Riloff, 2002; Cardie e Wagstaff, 1999; Soon *et al*., 2001; Ng e Cardie, 2002; Freitas *et al.* 2006). Na identificação automático de apostos em português, só sabemos da existência do parser PALAVRAS (Bick, 2000) e da recente tentativa de Freitas *et al.* (2006) utilizando técnicas de aprendizado de máquina, mas os resultados desses dois trabalhos ainda deixam a desejar. Freitas *et al.* comparam os resultados obtidos na identificação de apostos utilizando regras elaboradas manualmente por lingüistas e diferentes técnicas de aprendizado de máquina. Os resultados de precisão obtidos com o aprendizado de máquina foram apenas um pouco (cerca de 4%) superiores aos obtidos com as regras lingüísticas, sendo a recuperação baseada em regras ligeiramente superior (cerca de 3%) às demais técnicas. Tais resultados sugerem que, talvez para o aposto, a elaboração de regras lingüísticas seja uma forma eficaz de resolução do problema, mas ainda assim mais pesquisas são necessárias.

Considerações Finais

A lingüística computacional é uma área naturalmente interdisciplinar. Da perspectiva "computacional", o conhecimento lingüístico normalmente é baseado em compêndios gramaticais – que por sua vez são fortemente influenciados por uma visão da linguagem que vem desde a Antigüidade clássica. Porém, se, por um lado, este conhecimento já está totalmente estabelecido, por outro, ainda se ressente de um tratamento mais sistemático de determinados fenômenos relacionados às classes de palavras – o que, por sua vez, indica que lingüistas têm muito com o que contribuir nessa parceria.

Este trabalho evidenciou a relação cada vez mais próxima entre conhecimentos lingüísticos "tradicionais" e a formalização da língua para o PLN. Discutimos, por meio de estudos de problemas encontrados na prática do PLN, como o trabalho do lingüista é de alta relevância para a construção de aplicações sofisticadas e robustas, enfatizando fenômenos que têm recebido um tratamento secundário por parte da lingüística, como as palavras denotativas, o aposto e as formas nominais do verbo, notadamente o particípio. Especificamente, direcionamos nosso olhar para a questão das classes de palavras, cuja aplicação mais direta na Lingüística Computacional está na tarefa de etiquetagem morfossintática.

Com relação à produção de conhecimento lingüístico, observa-se que os métodos empíricos vêm ganhando um espaço extraordinário. Durante a última década, o Aprendizado de Máquina (AM) tem demonstrado ser uma ferramenta bastante eficaz na viabilização de tarefas lingüísticas, por meio de aquisição de informações de língua a partir de corpus, que de outro modo seria impossibilitada devido à enorme quantidade de tempo e mão-de-obra necessários. As técnicas de AM supervisionado exigem a construção de uma grande massa de exemplos corretamente identificados do fenômeno lingüístico a ser aprendido. De acordo com os experimentos descritos em (Ngai e Yarowsky, 2000), na identificação de SNs em textos em inglês, por exemplo, é mais vantajoso utilizar recursos humanos para fazer a anotação do corpus e utilizá-lo para treinar um identificador de SNs do que utilizar recursos humanos para criar manualmente regras de transformações para uma gramática de identificação. Dentre as vantagens listadas no trabalho por Ngai e Yarowsky, que podem ser generalizadas para outros tipos de tarefas de PLN, destacam-se:

(1) a aquisição distribuída de conhecimento, pois com a utilização de AM fica mais fácil a combinação de esforços de um grupo de pessoas. Corpora de treino criados por pessoas diferentes podem ser combinados facilmente para formarem um corpus maior. Em contraste, é muito difícil, ou quase impraticável, a combinação de listas de regras criadas manualmente por pessoas diferentes;

(2) a robustez do conhecimento baseado em observações empíricas, pois o desempenho de sistemas que utilizam regras codificadas manualmente tende a apresentar uma maior varia-

ção, enquanto que os resultados de sistemas treinados com corpora anotados são mais uniformes;

(3) independência dos mecanismos de inferência, pois, uma vez construído um corpus de treino, o aprendizado pode ser realizado por diversas técnicas, e subseqüentes progressos nos algoritmos de treinamento podem trazer melhorias nos resultados sem a necessidade de alterações no corpus. Em contraste, o desempenho obtido por um conjunto de regras codificado manualmente é definitivo, a não ser que haja uma revisão humana das regras.

Podemos adicionar à lista de Ngai e Yarowsky o fato, observado durante nossas pesquisas, de que a atividade de anotação de corpus de treino para a viabilização dos métodos simbólicos e estocásticos de aprendizado supervisionado permite o surgimento de valiosos *insights* sobre a sistematização lingüística e, como conseqüência, sobre o fenômeno lingüístico como um todo.

Referências

ABNEY, S. 1991. Parsing By Chunks. *In*: BERWICK, R.; ABNEY, S.; e TENNY, C. (eds.), *Principle-Based Parsing.* Dordrecht, Kluwer Academic Publishers.

BASILIO, M. 1999. *Teoria Lexical*. São Paulo, Ática, 94 p.

BECHARA, E. 1999. *Moderna Gramática Portuguesa*. 37ª ed., Rio de Janeiro, Lucerna, 671 p.

BICK, E. 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus, Dinamarca. PhD Thesis. Aarhus University, 150 p.

BIDERMAN, M.T. 1978. *Teoria Lingüística : Lingüística quantitativa e computacional*. Rio de Janeiro, LTC, 277 p.

BLOOMFIELD, L. 1966. A set of postulates for the science of language. *In*: M. JOOS (ed.), *Readings in Linguistics I.* Chicago, University of Chicago Press, p. 26-31.

CÂMARA JR., J.M. 2000. *Estrutura da Língua Portuguesa*. 32ª ed., Petrópolis, Editora Vozes, 114 p.

CARABALLO, S. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. *In*: THE 37TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON COMPUTATIONAL LINGUISTICS, Maryland, 1999. *Anais*... Maryland, College Park, p. 120-126.

CARDIE, C. e WAGSTAFF, K. 1999. Noun phrase coreference as clustering. *In*: THE JOINT SIGDAT CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND VERY LARGE CORPORA, Maryland, 1999. *Anais*… Maryland, University of Maryland, p. 82-89.

CUNHA, C. e CINTRA, L. 1985. *Nova gramática do português contemporâneo*. Rio de Janeiro, Nova Fronteira, 719 p.

DUBOIS, M.; GIACOMO, M.; GUESPIN, L.; MARCELLESI, C.; MARCELLESI, J. e MEVEL, P. 2001, *Dicionário de Lingüística*. 8ª ed., São Paulo, Cultrix, 653 p.

FREITAS, M.C; GARRÃO, M.; OLIVEIRA, C.; SANTOS, C. N. e SILVEIRA, M.C. 2005. A anotação de um corpus para o aprendizado supervisionado de um modelo de SN. *In*: XXV CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, São Leopoldo, 2005. *Anais*... Porto Alegre, Sociedade Brasileira de Computação, meio digital.

FREITAS, M.C.; DUARTE, J.C.; SANTOS, C.N.; MILIDIÚ, R.L.; RENTERÍA, R.P e QUENTAL, V. 2006. A Machine Learning Approach to the Identification of Appositives. Advances in Artificial Intelligence. *In*: X IBERO-AMERCICAN ARTIFICIAL INTELLIGENCE CONFERENCE – XVIII BRAZILIAN ARTIFICIAL INTELLIGENCE SYMPOSIUM (IBERAMIA-SBIA 2006), Ribeirão Preto, 2006. *Anais*... Heidelberg, Springer, p. 309-318.

HEARST, M. 1998. Automated Discovery of WordNet Relations. *In*: C. FELLBAUM (ed.), *WordNet: An Electronic Lexical Database*. Cambridge, MIT Press, p. 131-153.

HENRIQUES, C. 2003. *Sintaxe Portuguesa*. 3ª ed., Rio de Janeiro, Oficina do Autor, 140 p.

JURAFSKY, D. e MARTIN, J. 2000. *Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition*. EUA, Prentice Hall, 934 p.

KURY, A.G. 1960. *Português Básico*. Rio de Janeiro, Agir, 275 p.

KURY, A.G. 1999. *Novas lições de análise sintática.* 8ª ed., São Paulo, Ática, 207 p.

LAUDANNA, A. e VOGHERA, M. 2002. Nouns and Verbs as Grammatical Classes in the Lexicon. *Rivista di Linguistica*, **14**(1):9-26.

LYONS, J. 1977. *Semantics*. Cambridge, Cambridge University Press, 460 p.

MANNING, C. e SCHÜTZE, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MIT Press, 680 p.

MARCHI, A.R. 2003. Projeto Lacio-Web: Desafios na Construção de um Corpus de 1,1 Milhão de Palavras de Textos Jornalísticos em Português do Brasil. *In*: SEMINÁRIO DO GRUPO DE ESTUDOS LINGÜÍSTICOS DO ESTADO DE SÃO PAULO, 51, SÃO PAULO, 2003. *Anais*... São Paulo, Grupo de Estudos Lingüísticos de São Paulo, meio digital.

MARCUS, M. P.; MARCINKIEWICZ, M. A. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics,* **19**(2)313-330.

NG, V. e CARDIE, C. 2002. Improving machine learning approaches to coreference resolution. *In*: THE 40TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, Philadelphia, 2002. *Anais*... Philadelphia, Association for Computational Linguistics, p. 104-111.

OLIVEIRA, C.; FREITAS, M.C.; QUENTAL, V.; SANTOS, C.N.; LEME, R.P. e SOUZA, L. 2006. A Set of NP-Extraction Rules for Portuguese: Defining, Learning and Pruning. *In*: R. VIEIRA; P. QUARESMA; M. NUNES; N. MAMEDE; C. OLIVEIRA e M. DIAS (eds.), *Computational Processing of the Portuguese Language. Heidelberg,* Springer, p. 150-159.

PEREIRA, M.T. 1995. Palavras denotativas: temas e problemas. *In*: FLORES VERBAIS, UMA HOMENAGEM LINGÜÍSTICA E LITERÁRIA PARA ENEIDA DO REGO MONTEIRO BOMFIM NO SEU 70° ANIVERSÁRIO. HEYE, J. (org). Rio de Janeiro: 34 Editora, p. 15-21.

PERINI, M. 1996. *Gramática Descritiva do Português*. 2ª ed., São Paulo, Ática, 380 p.

PIMENTA-BUENO, M. 1986. As formas V+do do português: um estudo de classes de palavras. *DELTA,* **2**(2)207-229.

PHILLIPS, W. e RILOFF, E. 2002. Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. *In*: THE 2002 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP 2002), Morristown, NJ, 2002. *Anais*… Filadélfia, The Association for Computational Linguistics, **2**:125-132.

QUIRK, R.; GREENBAUM, S.; LEECH, G. e SVARTVICK, J. 1985. *A Comprehensive Grammar of the English Language*. London, Longman, 1779 p.

RILOFF, E. e SHEPHERD, J. 1997. A corpus-based approach for building semantic lexicons. *In*: THE SECOND CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, Providence, Rhode Island, 1997. *Anais*… Fi-

ladélfia, The Association for Computational Linguistics, p. 197-132.
ROSA, M.C. 2000. *Introdução à Morfologia.* São Paulo, Editora Contexto,156 p.
SAID ALI, M. 1964. *Gramática Secundária da Língua Portuguesa*. São Paulo, Ed. Melhoramentos, 325 p.
SAPIR, E. 1921. *Language: an Introduction to the Study of Speech.* New York, Harcourt/Brace and Company, 258 p.
SOON, W.; NG, H. e LIM, D. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics - Special issue on computational anaphora resolution,* **27**(4):521-544.

Su*bmetido em: 10/2006*
Aceito em: 11/2006

Claudia Oliveira
Instituto Militar de Engenharia

Maria Claudia de Freitas
PUC-RJ