Descrição de numerais segundo modelo *Universal Dependencies* e sua anotação no português

Magali Sanches Duran, Lucelene Lopes, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos, Brasil

magali.duran@gmail.com, lucelene@gmail.com, taspardo@icmc.usp.br

Abstract. *This paper describes the instantiation of Universal Dependencies (UD) guidelines for the annotation of numerals in Portuguese. We show the importance of knowing the topic to be instantiated in the context of several languages, since UD annotation aims to promote, as much as possible, parallelism between languages. We then explore the description of numerals in grammars, seeking subsidies to elaborate instructions for assignment of the UD part of speech tag NUM. The results of combining the UD guidelines with the characteristics of numerals in Portuguese are presented in detail, with examples, along with arguments supporting each decision made.*

Resumo. *Este artigo descreve a instanciação das diretrizes do modelo Universal Dependencies (UD) para a anotação de Numerais em português. Mostramos a importância de conhecer o tópico a ser instanciado no contexto de várias línguas, uma vez que a anotação UD tem por objetivo promover, o máximo possível, o paralelismo entre as línguas. Exploramos então a descrição dos numerais nas gramáticas, buscando subsídios para elaborar instruções para atribuir a etiqueta NUM da UD. Os resultados da combinação das diretrizes da UD com as características dos numerais em português são apresentados em detalhes, com exemplos, juntamente com os argumentos que amparam cada decisão tomada.*

1. Introdução

Universal Dependencies (UD) é uma iniciativa multinacional de anotação sintática de córpus que tem por objetivo estabelecer um maior paralelismo entre as línguas. A ideia subjacente é a de que, usando um mesmo esquema de anotação, facilita-se o estudo e a construção de aplicações multilíngues de processamento de línguas naturais. Por "esquema de anotação", entende-se tanto conjuntos comuns de etiquetas (morfológicas, morfossintáticas e de relações de dependência sintática) quanto diretrizes comuns para atribuição dessas etiquetas. O projeto UD é uma abordagem lexicalista (as palavras são as unidades mínimas às quais se atribuem as etiquetas) e fundamentada em gramáticas de dependências (Nivre et al., 2020). Atualmente o site da UD[1] disponibiliza cerca de 200 córpus anotados em mais de 100 línguas, o que demonstra a grande adoção do

[1] https://universaldependencies.org/

modelo. Das línguas do Brasil, no momento da escrita deste artigo, há três córpus de português disponíveis (córpus PUD, GSD e Bosque-UD) e seis córpus de línguas indígenas (apurina, guajajara, kaapor, makurap, mundukuru e tupinambá).

Um desafio que se apresenta para todas as línguas antes de fazer a anotação é adaptar as diretrizes da UD para as especificidades da língua. Essa fase de um projeto de anotação de córpus é chamada por Hovy e Lavid (2010) de "instanciação da teoria". A UD recomenda que, uma vez que o trabalho de instanciação da teoria tenha sido concluído, as diretrizes específicas da língua sejam publicadas no próprio site da UD, de forma que as demais iniciativas de anotação possam tomá-las como base. Até onde é de nosso conhecimento, ainda não foram publicadas diretrizes específicas para atribuição das etiquetas da UD para o português. É muito desejável que essa lacuna seja sanada para que alcancemos anotações consistentes em UD, não apenas dentro de um mesmo córpus, mas entre os córpus de português. Material teórico de suporte à anotação específico da língua é um recurso essencial para que eventuais anotadores de córpus tomem decisões menos subjetivas (na medida do possível) quando se deparam com situações não previstas nas diretrizes gerais da UD.

Imbuídos do propósito de prover os anotadores de UD em português com diretrizes de anotação, decidimos divulgá-las sob a forma de manuais de anotação. Nesses, cada uma das 17 *PoS tags* (*Part-of-Speech tags* ou etiquetas morfossintáticas) e das 37 *deprel* (*dependence relations* ou etiquetas de relações de dependência) da UD serão objeto de instruções detalhadas, ricamente ilustradas com exemplos. Neste artigo, descrevemos parte do esforço de elaborar esse material. Abordamos a questão da anotação morfossintática dos numerais em português de acordo com as diretrizes da UD e os estudos e argumentos que embasaram nossas decisões.

O artigo está organizado em 5 seções, incluindo esta introdução. Na Seção 2, apresentamos reflexões acerca dos numerais e sua expressão linguística. Em seguida, na Seção 3, expomos as divergências que encontramos entre gramáticos no que concerne à classificação dos numerais no português. Na Seção 4, descrevemos as diretrizes que produzimos para a anotação de numerais em português seguindo a abordagem da UD e, por fim, na Seção 5, tecemos nossas considerações finais.

2. Os numerais como parte integrante das línguas naturais

Os sistemas numéricos (ou sistemas de contagem) têm uma quantidade finita de algarismos únicos. Por exemplo, os algarismos arábicos constituem um conjunto com 10 integrantes (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) e os algarismos romanos constituem um conjunto com 7 integrantes (as letras I, V, X, L, C, D e M). A partir desses algarismos únicos, os sistemas de contagem utilizam estratégias de posicionamento, de soma, de multiplicação e de subtração para expressar outras quantidades. Assim, um "1" na segunda casa decimal representa uma dezena e, somado a um "3" na primeira casa, resulta no numeral 13. Nos algarismos romanos, além de multiplicação e adição, utiliza-se também a operação de subtração, motivo pelo qual "quarenta" é expresso por XL (L menos X, ou seja, cinquenta menos dez).

Na escrita, um token de numeral pode conter, além de vários algarismos, sinais de pontuação, como vírgula, ponto e barra: 1,07; 127.234; 4/7. No entanto, quando um numeral é expresso por extenso, nem sempre corresponde a um único token: 8 é "oito" (um *token*), mas 349 é "trezentos e quarenta e nove" (cinco *tokens*). Há também casos em que um número constituído de mais de um algarismo corresponde uma única palavra, como 15 (dois algarismos), que é expresso como "quinze" (um *token*). Em suma, há falta de simetria entre a quantidade de algarismos de um número e a quantidade de palavras usadas para expressá-lo por extenso.

Além disso, outra assimetria é observada: línguas que compartilham um mesmo sistema numérico possuem quantidades diferentes de itens lexicais para expressar os números. Apesar dessas diferenças, Hurford (2007) defende que há regras universais atuando na forma como as línguas usam um conjunto finito de palavras para expressar uma série infinita de números. Para ele, as estratégias usadas para compor os números a partir de uma quantidade finita de algarismos são as mesmas usadas na expressão linguística desses mesmos números. As principais estratégias são a junção (soma) e a multiplicação. No Português, a junção de algarismos é comumente indicada pela coordenação de palavras, usando uma conjunção aditiva, enquanto a multiplicação é indicada pela justaposição de palavras. Assim, 37 (30 + 7) é "trinta *e* sete" e 8.000 (8 x 1000) é "oito mil". Outras línguas, contudo, utilizam a justaposição em ambas as situações. Há, inclusive, línguas que concatenam as palavras em outra ordem, como em alemão, na qual 37 é siebenunddreißig (sieben=7 und=e dreißig=30).

O fato de cada língua possuir uma quantidade diferente de palavras para expressar os números prejudica um pouco o paralelismo entre elas. No francês, por exemplo, setenta, oitenta e noventa não são números expressos por palavras únicas: 70 é soixante-dix (sessenta mais dez), 80 é quatre-vingt (quatro vezes vinte) e 90 é quatre-vingt-dix (quatro vezes vinte, mais 10). Da mesma forma, no inglês e no francês, ao contrário do português, não há palavras para designar as centenas e, por isso, são usadas combinações da palavra "cem" com a quantidade de centenas, ficando subentendida a multiplicação. Assim, 200 é *two hundred* (dois cem) em inglês e *deux cent* (dois cem) em francês . Essas poucas comparações são suficientes para se perceber que é um desafio, para a UD, manter o máximo de paralelismo entre as línguas no que diz respeito à anotação de numerais.

3. Os numerais nas gramáticas do português

Em 1959, o Ministério de Educação e Cultura publicou, com força de lei, a Nomenclatura Gramatical Brasileira (NGB). O objetivo era criar uma ontologia dos tópicos de gramática que deveriam ser ensinados no Brasil, padronizando os termos usados para designar cada tópico. A NGB estabeleceu que há 10 classes de palavras no português, uma das quais é a dos numerais. Segundo o documento, a classe dos numerais se subdivide em quatro subclasses: cardinais (um, dois, três), ordinais (primeiro, segundo, terceiro), multiplicativos (dobro, triplo) e fracionários (meio, terço).

A NGB[2] está em vigor até hoje e orientou a confecção de diversas gramáticas escolares desde 1959. Todos os gramáticos que almejavam ter suas obras referenciadas para uso no ensino e em concursos tiveram que seguir a norma estabelecida pela NGB. Curiosamente, quando Portugal publicou a Nomenclatura Gramatical Portuguesa, em 1967, apresentou as mesmas 10 classes de palavras e, na classe dos numerais, incluiu a subclasse dos "numerais coletivos", ou seja, palavras que trazem embutida a ideia de uma quantidade (quinzena, bimestre, biênio, dúzia, dezena, par, quinteto, etc.).

Desde o início da vigência da NGB, contudo, não faltaram críticas a vários de seus tópicos. Uma das críticas é justamente com relação à classe dos numerais. Câmara Jr. (1986), por exemplo, diz que a NGB misturou critérios de forma, função e sentido das palavras ao propor 10 classes gramaticais (para ele só existem três classes de vocábulos: Nomes, Pronomes e Verbos). O autor considera os numerais como pertencentes à classe dos Nomes, a qual abriga as funções de substantivo e adjetivo. Monteiro (1991) também considera os numerais como integrantes da classe dos Nomes e exemplifica: em "três é ímpar", a palavra "três" exerce a função de substantivo e, em "três ímpares", a palavra "três" exerce a função de adjetivo.

Para Azeredo (2000), numeral é uma função semântica e, por isso, não deveria haver uma classe chamada "numerais". Segundo o autor, "O numeral é sempre constituinte de um sintagma nominal, ora ocupando a posição de núcleo - numerais fracionários e multiplicativos, ora ocupando a posição de termo adjacente - numerais cardinais e ordinais." (Azeredo, 2000, versão digital).

Uma visão intermediária entre a NGB e os autores que descartam a pertinência de uma classe de numerais é apresentada por Macambira (1987). Segundo esse autor, apenas os numerais cardinais pertencem à classe dos numerais propriamente ditos. Os numerais ordinais, segundo ele, comportam-se mais como os adjetivos, ao passo que os fracionários e multiplicativos comportam-se como substantivos. Além disso, apenas numerais que modificam diretamente um substantivo devem ser considerados, o que implica a exclusão de palavras quantitativas que se ligam ao substantivo por intermédio da preposição "de". Assim, em "mil canetas", "mil" é numeral, enquanto em "um milhão de canetas" "um" é numeral e "milhão" é substantivo. Macambira (1987) destaca também que apenas o número "um" e o feminino "uma" são singulares, pois os demais números, por expressarem mais de um, trazem implicitamente a ideia de plural. Com isso, o autor conclui que nenhum numeral admite a flexão de número, pois não há sentido em fazer o plural de uma palavra que já é plural. Esses critérios de Macambira para julgar se uma palavra pertence ou não à classe dos numerais serão, como veremos na Seção 4.4, muito relevantes dentro da abordagem de anotação UD.

É importante esclarecer que, embora nenhum numeral admita flexão de número, na língua portuguesa há onze numerais que admitem flexão de gênero:

---

[2] Os bastidores da criação da NGB são discutidos por Henriques (2009), o qual traz anexas a Nomenclatura Gramatical Brasileira e a Nomenclatura Gramatical Portuguesa.

meio/meia, um/uma, dois/duas, duzentos(as), trezentos(as), quatrocentos(as), quinhentos(as), seiscentos(as), setecentos(as), oitocentos(as) e novecentos(as).

4. A anotação de numerais segundo as diretrizes da UD

As diretrizes da UD apresentam muitas diferenças em relação ao que está previsto para os numerais na NGB e na NGP. Há uma *PoS tag* específica para numeral na UD (NUM), porém ela é reservada unicamente aos numerais cardinais, seja em sua forma como algarismos (arábicos ou romanos), seja por extenso. Os numerais ordinais, por outro lado, são anotados na UD como adjetivo (ADJ). Assim, **9** é NUM e **9°** é ADJ. As demais subclasses de numerais previstas na NGB e na NGP (multiplicativos, fracionários e coletivos) não são mencionadas pela UD.

Para a UD, a classe dos numerais (representada pela *PoS tag* NUM) é uma classe que abriga palavras de uma classe fechada (os algarismos e suas respectivas formas por extenso), que ora exercem função de modificador nominal, ora função de pronome, e ora função de substantivo:

> Ele teve **três** chances de acertar. As **três** foram desperdiçadas. E achava que o **três** era seu número de sorte!

Na primeira sentença, "três" atua como modificador nominal; na segunda, como pronome e na terceira, como substantivo. Em todas as sentenças, contudo, "três" é anotado como NUM. Assim, os tokens em negrito a seguir são anotados com NUM:

> Nós **dois** somos brasileiros.
> Há **2** anos que nos conhecemos.
> São **dois** os motivos pelos quais desistimos.
> **Dois** de nós são brasileiros e **dois** são estrangeiros.
> **Dois** mais **dois** são quatro.
> O Canal **2** é a TV Cultura.

Apresentamos, em seis subseções, as principais questões que suscitam dúvidas durante a anotação de numerais.

4.1 Frações

As frações não são sinônimo de numerais fracionários. Os numerais fracionários da NGB correspondem aos denominadores das frações, quando expressos por uma única palavra. Na UD, o numerador de uma fração é sempre expresso por um numeral cardinal e o denominador é expresso por um substantivo (de "meio" a "décimo") ou por um cardinal seguido do substantivo "avos": 3/15 = três quinze avos (a partir do denominador 11: onze avos). Embora os denominadores de "quarto" a "décimo" tenham formas iguais aos numerais ordinais (anotados como ADJ), eles devem ser anotados como NOUN nessa função.

> ⅔ = dois terços (NUM, NOUN)
> ¾ = três quartos (NUM, NOUN)
> 4/11 = quatro onze avos (NUM, NUM, NOUN)

A palavra "meio" (½) é a única que denota uma fração completa e, por isso, é anotada como NUM quando se liga diretamente a um substantivo, quantificando-o.

½ = meio, meia (NUM): Comeu meia pizza.

No entanto, em contexto matemático usa-se dizer “três meios” (correspondendo à fração imprópria 3/2) e, nessa situação, “meio” é NOUN, como os demais denominadores de frações, pois flexiona no plural. Cabe aqui mencionar que há outro NUM representado pela palavra “meia”, que é quando o algarismo 6 é expresso por extenso como “meia”, alternativamente a “seis”.

4.2 Algarismos romanos

A UD, em suas diretrizes, trata os algarismos romanos da mesma forma que os algarismos arábicos, ou seja, considera ambos numerais cardinais e, consequentemente, atribui-lhes a etiqueta NUM. Contudo, é importante ressaltar que os algarismos romanos podem apresentar uma leitura de numeral ordinal em algumas situações no português e, mesmo assim, devem ser anotados como NUM e não como ADJ como os numerais ordinais. Isso porque, enquanto os numerais ordinais têm uma marca gráfica indicando sua condição, os algarismos romanos não a têm em português. É possível, inclusive, em alguns casos, admitir as duas leituras, como é o caso de “capítulo XII” (capítulo doze ou capítulo décimo segundo). Assim, nos dois exemplos abaixo, o numeral VIII é anotado como NUM, embora no segundo tenha leitura de um numeral ordinal.

No século **VIII** não se sabia que a Terra era redonda ainda. (VIII = oito)
O rei Henrique **VIII** casou-se sete vezes. (VIII = oitavo)

4.3 Palavras que expressam ideias numéricas

O fato de a UD não mencionar outras formas de numerais, além dos cardinais (NUM) e ordinais (ADJ), não é estranho. Como vimos, os críticos da NGB diziam que os numerais pertencem à classe dos nomes, a qual abriga as funções de substantivo e adjetivo. Por exemplo, entre os multiplicativos estão incluídos os substantivos “dobro” e “triplo” e os adjetivos “duplo” e “tríplice”; entre os coletivos, estão incluídos os substantivos “bimestre” e “biênio” e os adjetivos “bimestral” e “bianual”. Além disso, há os prefixos que permitem formar novos substantivos: “bi”, “di”, “tri”, etc. Algumas palavras podem, inclusive, ora ser ADJ, ora ser NOUN, dependendo do contexto:

O DNA é uma cadeia de **dupla** hélice. (dupla = ADJ)
A **dupla** sertaneja foi formada nos anos 90. (dupla = NOUN)

4.4. O caso de “cento”, “milhão”, “bilhão”

Há algumas palavras que, por participarem da expressão de um número, podem causar estranheza por serem anotadas como substantivos (NOUN). É o caso de “cento”, “milhão”, “bilhão”, “trilhão”, etc.

101 = cento e um
2.221.000 = dois milhões e duzentos e vinte e um mil

Contudo, tais palavras nunca se ligam a um substantivo sem a intervenção da preposição “de”, ao contrário dos numerais, que se ligam diretamente como modificadores nominais. É agramatical dizer:

***Cento** pessoas compareceram.

*__Milhões__ pessoas compareceram.

"Cento" é uma palavra que substitui "cem" nas expressões de números de 101 a 199, ou seja, num contexto bem limitado. Há contextos, porém, em que o comportamento de "cento" mostra claramente tratar-se de um substantivo, sendo passível de flexão de número e ligado a outro substantivo por meio da preposição "de":

Dois **centos** de docinhos

No caso de "milhão" (assim como "bilhão", "trilhão", etc.), percebe-se que é palavra flexionável em número e exige a preposição "de" para se ligar a um substantivo. Quando precedida de um verdadeiro numeral, "milhão" denota quantidade precisa:

Dois **milhões** de pessoas compareceram.

Quando utilizada no plural e não precedida de um verdadeiro numeral, "milhão" denota uma quantidade imprecisa, genérica:

Milhões de pessoas compareceram.

Portanto, os critérios de Macambira (1993), apresentados na Seção 3, mostram-se relevantes para distinguir um numeral (NUM) de um substantivo (NOUN) na UD e ratificam a interpretação de que "cento", "milhão", "bilhão", etc. devem ser anotados como NOUN.

4.5 A ambiguidade das formas "um/uma"

O caso que apresenta maior dificuldade na anotação de numerais está associado à ambiguidade das palavras "um/uma". Nem sempre é simples decidir entre as etiquetas DET (determinante, que é a categoria da UD que abriga os artigos indefinidos), PRON (pronome) e NUM. Para auxiliar os anotadores nessa decisão, fixamos algumas regras, parte das quais coincide com pistas de desambiguação mencionadas por Macambira (1987). Trata-se de NUM se:

- "um" ou "uma" responder à pergunta "Quantos?". Ex: Ganhou uma medalha de ouro. Quantas medalhas ganhou? Resposta: uma.
- "um" ou "uma" estiver em oposição a outras quantidades na mesma sentença. Ex: Comprou um abacate e três maçãs.
- "um" ou "uma" estiver seguido da preposição "de", indicando a seleção de um entre vários, mas que poderia ser "dois", "três", etc.. Ex: **Um** dos alunos foi eleito representante.[3]

Trata-se de DET se:

- o substantivo que sucede "um" e "uma" for incontável. Ex: Tenho **uma** grande esperança de que isso funcione.
- "um" ou "uma" puder ser suprimido. Ex: Tive **uma** impressão errada. Tive impressão errada.

[3] Observa-se que a sentença não muda de sentido com a inversão dos constituintes: "Dos alunos, **um** foi eleito representante". Além disso, "um" poderia ser substituído por "dois" ou outro numeral: "dois dos alunos foram eleitos representantes".

- “um” ou “uma” não puder ser substituído por outros números. Ex: Saiu para dar **um** passeio. *Saiu para dar **dois** passeios (não é agramatical, mas é improvável)

Trata-se de PRON se:

- “um” ou “uma” estiver em oposição a outros pronomes indefinidos, como “outro” e “outra”. Ex: **Um** ou **outro** vai vencer.
- “um” ou “uma” estiver precedido do pronome indefinido “cada” (DET na UD). Ex: **Cada um** deu o melhor de si. **Cada um** de nós deu o melhor de si.[4]

4.6 Casos especiais de tokens que incluem numerais

Na UD, problemas de tokenização podem alterar a forma de anotação. Se um numeral cardinal fizer parte de um token que contém um substantivo ou um nome próprio, por exemplo, ele não será anotado como NUM, mas sim com a etiqueta da outra classe:

> A banda **U2** fez muito sucesso no Brasil.
> (U2 é anotado com a etiqueta de nome próprio PROPN).
>
> Hoje é **24/jul/2021**
> (como contém o nome do mês, esse token é anotado na UD como substantivo, ou seja, um NOUN).
>
> Já andamos **24km** nesta semana.
> (24km é anotado como substantivo, pois contém uma unidade de medida e unidades de medida, abreviadas ou por extenso, são NOUN, segundo a UD).

6. Considerações finais

Apresentamos o desafio de instanciar as diretrizes da UD na língua portuguesa no que se refere à anotação morfossintática dos numerais. Esperamos, com isso, auxiliar outros estudos linguísticos e projetos de anotação de córpus que se filiem ao mesmo modelo, pois poderão antecipar questões relevantes. No nosso caso, embora tenhamos elaborado um manual prévio para orientação dos anotadores, tivemos que complementá-lo para contemplar situações que geraram dúvidas durante a anotação. Também produzimos instruções para as demais etiquetas utilizadas pela UD e, em breve, esperamos divulgar o Manual para Anotação de *PoS tags* da UD na Língua Portuguesa, assim como produzir outros artigos com descrições e discussões linguísticas relevantes.

O trabalho reportado faz parte de um projeto maior - o POeTiSA[5], cujo propósito é avançar as pesquisas em sintaxe e *parsing* para o português brasileiro, construindo um grande treebank multi-gênero e produzindo *taggers* e *parsers* do estado da arte.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da USP (C4AI - http://c4ai.inova.usp.br/), com o apoio da IBM e da FAPESP (#2019/07665-4).

[4] Aqui a regra do “de” é sobreposta pela regra do “cada”. Na expressão “cada um de”, não há opção de inverter a ordem dos constituintes: *Dos alunos, cada **um** deu o melhor de si.
[5] https://sites.google.com/icmc.usp.br/poetisa

Referências bibliográficas

Azeredo, José Carlos de. Fundamentos de Gramática do Português. Editora Zahar, 2000. Versão eletrônica.

Câmara Jr., Joaquim Mattoso. Dicionário de Linguística e Gramática: referente à língua portuguesa. Petrópolis: Vozes, 1986, 13ª ed.

Henriques, Claudio Cezar. Nomenclatura Gramatical Brasileira: cinquenta anos depois. São Paulo: Parábola, 2009.

Hovy, E. and Lavid, J. (2010). Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. International Journal of Translation, Vol. 22, N. 1, pp. 13-36.

Hurford, James R. (2007). A performed practice explains a linguistic universal: Counting gives the Packing Strategy. Lingua,Volume 117, Issue 5, p. 773-783.

Macambira, José Rebouças. A Estrutura Morfo-Sintática do Português. São Paulo: Livraria Pioneira Editora, 1987.

Monteiro, José Lemos. Morfologia Portuguesa. Campinas: Pontes, 1991.

Nivre, J.; Marneffe, M-C.; Ginter, F.; Hajič, J.; Manning, C.D.; Pyysalo, S.; Schuster, S.; Tyers, F.; Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In the Proceedings of the 12nd International Conference on Language Resources and Evaluation (LREC), pp. 4034-4043.