Linha D'Água: São Paulo, v. 34, n. 02, p. 103-116, mai.-ago. 2021
http://dx.doi.org/10.11606/issn.2236-4242.v34i2p103-116

Artigo / Article

A deixis: uma proposta de anotação em XML no âmbito do texto

Deixis: A Proposal for XML Annotation within the Text

Miguel Magalhães*
migmaglit@fcsh.unl.pt
https://orcid.org/0000-0003-0055-8971

Matilde Gonçalves**
matilde.goncalves@fcsh.unl.pt
https://orcid.org/0000-0003-0039-4401

Resumo

Neste trabalho, propomos uma metodologia para a anotação dos deíticos e da deixis em XML. A utilização da linguagem XML para a anotação de *corpora* tem conhecido um desenvolvimento nos últimos anos, com a publicação de diversas metodologias ou refinamento de outras já existentes. Mas a anotação da deixis tem colocado problemas uma vez que esta opera a vários níveis. De facto, a análise da deixis depende de um conjunto de elementos linguísticos, mais ou menos expressos que não podem ser analisados individualmente, mas sim na relação que estabelecem entre eles e o contexto de produção e de circulação dos textos. Esta contingência conduz a problemas de sobreposição de níveis de análise. A metodologia, aqui apresentada, não só é sensível às relações com o contexto de produção e circulação dos textos, como também permite analisar essas mesmas relações sob diversas perspectivas.

Palavras-chave: Deixis; *Corpora*; Anotação; XML; Texto.

Abstract

In this work, we propose a methodology for annotating deictics and deixis in XML. The use of XML for corpora annotation has increased in recent years with the publication of several methodologies or the refinement of existing ones. However, noting deixis has posed problems since it operates on several levels. In fact, the analysis of deixis depends on a set of linguistic elements, more or less expressed, that cannot be analysed individually, but only within the relationship they establish between them and the context of production and

* Faculdade de Ciências Sociais Humanas da Universidade Nova de Lisboa; Centro de Linguística da Universidade Nova de Lisboa, Lisboa, Portugal. Bolsista da Fundação para a Ciência e Tecnologia (PD/BD/142789/2018).

** Faculdade de Ciências Sociais e Humanas da Universidade NOVA de Lisboa; Investigadora do Centro de Linguística da Universidade Nova de Lisboa, Lisboa, Portugal.

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

circulation of texts. This contingency leads to problems with overlapping layers of analysis. This methodology is not only sensitive to these relationships with the context of production and circulation of texts, but also allows us to analyse these relationships from different perspectives.

Keywords: *Deixis; Corpora; Annotation; XML; Text.*

Introdução

O presente artigo[1] insere-se numa tese de doutoramento em curso, em linguística do texto e do discurso, e tem como objetivo apresentar uma proposta de anotação de *corpus* em XML que permita não só quantificar os elementos deíticos nos textos, mas também a visualização da relação que se estabelece entre estes elementos, nomeadamente temporais, espaciais e pessoais, na construção da deixis.

A anotação de *corpora* para o tratamento automático de línguas não é nova e existem algumas normas que têm sido publicadas e desenvolvidas (TEI, Eagles, a título de exemplo) mas estas normas, como aponta (HARDIE, 2014), são exaustivas e orientadas para projetos de anotação volumosos. Ou seja, não são necessariamente pertinentes para *corpus* de tamanho mais reduzido, nem podem ser executados por investigadores a um nível individual.

Atualmente, os trabalhos em linguística tendem a depender da análise de *corpus* construídos propositadamente, com anotações específicas e orientadas para o objeto de estudo em questão. Em consequência, assistimos, na última década, ao esforço de desenvolver metodologias e ferramentas que permitam analisar e observar de forma automática fenómenos semânticos em *corpora*.

No seguimento do trabalho que nos propomos realizar, no âmbito do doutoramento, a anotação dos deíticos é uma das ferramentas que nos permite perceber se existem padrões linguísticos associados à deixis e se esses padrões nos fornecem evidências para o estudo de géneros textuais. Deste modo, partindo das noções de parâmetros de género, mecanismos de realização textual e marcadores de género (MIRANDA, 2010), tal como foram definidos por (COUTINHO; MIRANDA, 2009), noções centrais na linguística do texto e do discurso, na linha do Interacionismo Sociodiscursivo (ISD) (BRONCKART, [1997] 1999), procuramos mostrar a relevância deste processo na aplicação prática nas áreas de ciência de dados. Concomitantemente, visamos perceber de que modo a ciência de dados pode funcionar como um instrumento auxiliar à análise textual de *corpora*.

Embora a deixis seja um dos aspetos a analisar, é também um dos mais complexos em termos técnicos pela própria natureza do objeto, como podemos observar:

[1] Trabalho financiado por fundos nacionais portugueses- Fundação para a Ciência e Tecnologia, como parte do projeto do Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020.

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

> [...] but now we are into deep theoretical water, because now the language of thought has indexicals, and in order to interpret THEM *we would need all the apparatus we employed to map contexts into propositions that we need in linguistics but now reproduced in the lingua mentalis, with a little homunculus doing all the metalinguistic work.* (LEVINSON, 2006, p. 100, sublinhado nosso).

Do que fica dito nesta citação de Levinson, a interpretação da deixis não está dependente apenas de uma descrição objetiva da linguagem mas do mapeamento dos contextos e de um conjunto de ferramentas que permitam descrever esse mapeamento. Assim, a análise da deixis depende de um conjunto de elementos linguísticos, mais ou menos expressos (como os tempos verbais, sujeitos expressos ou nulos, localizadores temporais e espaciais, entre outros) que não podem ser analisados individualmente, mas sim na relação que estabelecem entre eles e com o contexto de produção e de circulação dos textos.

Para além da introdução e da conclusão, este artigo estrutura-se em três partes. A primeira incide no tratamento automático de textos, a segunda sobre a deixis e a terceira sobre estrutura e anotação em XML.

1 *Corpora* e tratamento automático de textos

O uso de *corpus* para a análise linguística tem assistido, nos últimos anos, a um crescimento exponencial, ligado ao desenvolvimento de ferramentas e algoritmos, que permitem analisar, cada vez, maiores volumes de texto com ferramentas automáticas. A democratização do acesso a ferramentas informáticas tem permitido a constituição e uso de *corpus* anotados para trabalhos de investigação linguística mais específicos, permitindo construir *corpus ad-hoc*. A publicação de vários documentos normativos, como a *Text Encoding Iniciative* (TEI) ou *Expert Advisory Group on Language Engineering Standards* (Eagles) permitiu a estandarização da anotação de grandes *corpora*. No entanto, como referido em trabalhos anteriores (HARDIE, 2014), o formato TEI foi criado e desenvolvido num momento em que a criação de *corpora* era uma tarefa desenvolvida em projetos de investigação com equipas numerosas e, por isso, a anotação foi desenvolvida para ser profunda e para que tivesse o maior número de usos possíveis. Um dos exemplos desta prática é o British National *Corpus* (BNC) que foi desenvolvido sob a alçada de um consórcio composto por instituições públicas e privadas, e cujas equipas eram compostas por linguistas de várias áreas, bem como programadores informáticos. O BNC tornou-se, por isso, um *corpus* de referência para a investigação em linguística do inglês (britânico).

Atualmente, os *corpora* são construídos e anotados para projetos de investigação específicos, com equipas pequenas ou por investigadores individuais que não têm meios nem tempo para seguir as normas publicadas ou, porque não têm interesse de tornar público esse trabalho de anotação. Deste modo, têm surgido trabalhos que, não abandonando completamente os documentos normativos, propõem formas mais simplificadas de anotação (HARDIE, 2014). Por outro lado, o uso de *corpus* anotados que até recentemente se restringia à análise de

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

fenómenos lexicais, morfológicos ou sintáticos, tem-se alargado à análise semântica, como consequência do interesse no desenvolvimento de ferramentas para o processamento automático das línguas, em áreas como *data mining*, *text mining*, *marketing*, *machine learning*, entre outros.

A análise semântica dos textos, através de ferramentas automatizadas, coloca vários desafios, tanto da perspectiva linguística como da perspectiva técnica:

- A análise semântica tem uma componente subjetiva porque é uma avaliação que se faz sobre o enunciado e/ou o texto.
- Depende de fatores intra e extra linguísticos (contexto de comunicação, ancoragem espácio-temporal, operações de recuperação e ancoragem de informação) que ainda não são totalmente conhecidos e compreendidos;
- Não sendo possível estabelecer com clareza as operações de recuperação e ancoragem de informação, porque são de natureza abstrata, é difícil definir algoritmos que permitam analisar os textos de forma automática.

No entanto, algum trabalho tem sido desenvolvido nesta área e a análise semântica do texto só pode ser executada até um determinado ponto: extração de entidades e relações, desambiguação e análise de sentimentos, por exemplo. Mas o processamento das línguas naturais (NLP, sigla inglesa) a um nível mais profundo requer um conhecimento alargado do contexto de comunicação, e que é designado frequentemente por "senso comum". O senso comum refere-se ao contexto de produção e reconhecimento do uso do documento. É este contexto que permite a desambiguação do processo de comunicação:

> Reasoning about time is one of the most important aspects of commonsense reasoning. Linking a formal theory for time with an annotation scheme aimed at extracting rich temporal information from natural language text is significant for at least two reasons. It will allow us to use the multitude of temporal facts expressed in text as the ground propositions in a system for reasoning about temporal relations. It will also constitute a forcing function for developing the coverage of a temporal reasoning system, as we encounter phenomena not normally covered by such systems, such as complex descriptions of temporal aggregates. (HOBBS; PUSTEJOVSKY, 2003, p.74)

No que concerne aos estudos linguísticos relativos à análise do contexto, destacamos os trabalhos realizados no âmbito da Teoria Formal Enunciativa de A. Culioli (1999), sublinhando a compatibilidade epistemológica deste quadro com o do ISD. Sendo um modelo descritivo e explicativo sustentado em operações predicativas e enunciativas, possibilita dar conta dos diversos ajustamentos intersubjetivos e discursivo-textuais realizados pelos produtores textuais em função dos constrangimentos de origem contextual e sócio-cultural (CULIOLI 1981, p.53-54).

Esta teoria introduz um modelo descritivo e explicativo do modo como a significação é construída através e pela enunciação, evidenciando a importância do contexto na interpretação:

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

> [...] the context-dependence principle is considered as the decisive factor underlying any strict form of the principle of compositionality. According to this understanding, the study of natural languages is regarded as a study of interpretation. (VALENTIM, 2015, p.297)

De acordo com o exposto, fica claro que o utilizador não pode projetar exclusivamente as unidades linguísticas e as suas combinatórias no enunciado, tem, sim, de observar sistematicamente o contexto em que ocorrem (VALENTIM, 2015, p. 297). A vantagem desta abordagem é que (i) permite descrever a língua de uma forma mais complexa do que composicional e (ii) pressupõe que exista um conjunto de operações, comuns a um grupo, com as quais podemos reconstruir e interpretar os enunciados produzidos por outros. Qualquer tipo de anotação que pretenda descrever os aspetos semânticos de um texto tem, necessariamente, de relevar estes dois pontos: descrever o valor semântico dos elementos e tornar visível as operações de interpretação que lhe estão subjacentes, através da interrelação que se estabelece entre os elementos textuais.

Uma das marcas visíveis das operações cognitivas e linguísticas, elaboradas pelos enunciadores e coenunciadores e altamente dependentes do contexto, são os deíticos, que iremos abordar na seção seguinte.

2 Deixis

A relação entre discurso e contexto tem sido largamente debatida ao longo do tempo e não está no âmbito deste trabalho explanar as implicações teórico-filosóficas dessa relação. Vamos, por isso, partir da reflexão feita por Fernanda Fonseca e assumir que:

> A relação dinâmica que se estabelece entre o discurso e o contexto tem o seu fulcro nas operações de referenciação deítica, isto é, as operações que pressupõem a existência de um contexto referencial e viabilizam a sua representação conceptual sob a forma de um mundo. (FONSECA, 1992, p.138)

Deste modo, a construção dos referidos "mundos" está dependente do estabelecimento de coordenadas espácio-temporais e da rede de relações que estas estabelecem no ato enunciativo que as institui. Assim, os deíticos e localizadores surgem como a face visível desse mundo e a análise do seu funcionamento constitui uma via de acesso às operações enunciativas que permitem a construção do referente pela língua.

Destacamos igualmente o contributo de Lyons (1977) na definição da deixis:

> By deixis is meant the location and identification of persons, objects, events, processes and activities being talked about, or referred to, in relation to the spatiotemporal context created and sustained by the act of utterance and the participation in it, typically, of a single speaker and at least one addressee. (LYONS, 1977, p.637)

Face ao exposto, assume-se que a deixis é o processo através do qual é feita a construção de valores linguísticos, que resultam na representação da referência, relativamente aos sujeitos

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

e ao espaço-tempo no enunciado e no texto. Esse processo é materializado linguisticamente através dos deíticos, enquanto "gestos verbais cuja função primária é estabelecer a ligação entre o explícito e o implícito na comunicação verbal". (FONSECA, 1992, p. 70)

Deste modo, a mostração ou ancoragem situacional que a deixis revela não pode ser observada e analisada apenas num campo físico, sensorialmente observável, mas num contexto que é partilhado: "a memória e os contextos de vária ordem a que podem implicitamente recorrer". (FONSECA, 1992, p.71)

Do que fica dito anteriormente, a deixis não se realiza apenas através das marcas que o locutor-enunciador lhe imprime, mas através de operações:

> Assim, as propriedades das diferentes formas e os valores das construções, tendo uma incidência na gramática de cada língua, constituem-se como marcadoras de operações abstratas, observáveis, nos textos através das diferentes "determinações" que desencadeiam (CORREIA; PEREIRA, 2015, p.51)

São estas marcas linguísticas com valor deítico que nos propomos a anotar e quantificar por forma a visualizar a relação que se estabelece entre estes elementos nos textos com o contexto.

2.1 Espaço, tempo e pessoa: a ancoragem deítica

Como referimos anteriormente, os deíticos permitem fazer a ancoragem do enunciado e interpretar quais são os sujeitos enunciadores e compreender as relações espácio-temporais que se estabelecem entre factos e referentes do ponto de vista dos interlocutores. Os deíticos são de três tipos: pessoais, temporais e espaciais.

Relativamente aos sujeitos da enunciação, as formas pessoais que adquirem valor deítico são as formas das primeiras e segundas pessoas, tanto no singular como no plural:

> First and second person are the only forms among all the pronominal forms mentioned above that have deictic values, since both function as subjective linguistic indices. In enunciative terms, the first person results from the process of identification that may occur between the enunciator subject/speaker and the enunciation subject (which coincides with the syntactic subject). In turn, the second person arises from the differentiation between the enunciator subject/speaker and the enunciation subject. However, in this case the enunciation subject (or syntactic subject) identifies itself with the co-enunciator subject/interlocutor. The third person does not assume any deictic value. (VALENTIM, 2015, p.300)

Deste modo, as marcas de pessoa, expressas pelos pronomes pessoais e marcas de flexão verbal (no caso do português europeu), pronomes/determinantes possessivos e pronomes de tratamento constituem-se como os elementos necessários ao nosso trabalho, enquanto marcas externas das operações cognitivas e linguísticas da referenciação.

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

Os deíticos espaciais são formas que indicam o espaço e que permitem a interpretação em relação ao espaço da enunciação. Assim, a deixis espacial inclui não só os advérbios e expressões adverbiais de lugar, mas também os pronomes e determinantes demonstrativos que indicam proximidade ou afastamento do locutor e/ou recetor, e também os verbos que demonstram movimento / localização de e para o espaço do emissor, como podemos observar no exemplo seguinte:

(1)

Chegámos à CES 2015, Consumer Electronics Show de Las Vegas, com alguma expectativa sobre os televisores quantum dot LED. O entusiasmo esmoreceu ao confirmarmos que o dito novo tipo de iluminação de ecrã já tinha sido usado em alguns modelos da série Triluminos da Sony, em 2013. Tínhamos testado alguns e confrontámos a teoria com os resultados obtidos na altura, em laboratório. (ID718)

No exemplo (1) podemos observar que as marcas morfológicas de 1ª pessoa do plural são um deítico pessoal e o verbo "chegar" com a preposição "à" indica um movimento a partir do espaço do interlocutor, funcionando, em interrelação com o evento (CES), como um localizador espacial. É nesta localização espacial que se irá desenvolver o enunciado. Este exemplo mostra também que a deixis não funciona exclusivamente com formas e construções fechadas e fixas, mas numa rede de interdependências, quer linguísticas, quer extralinguísticas, como a seguir evidenciamos.

No exemplo (1), podemos também observar a deixis temporal. O primeiro elemento a considerar é o tempo verbal pretérito perfeito simples (PPS) (*chegámos* e *confrontámos*) ao qual é atribuído um valor de anterioridade em relação ao momento da enunciação (T0). Concomitantemente, a presença de dois localizadores temporais autónomos (neste caso, a data *2015* e *2013*) irão determinar o ponto de partida das duas sequências temporais, a partir das quais se vai desdobrar a temporalidade semiotizada neste segmento textual: uma relativa ao eixo temporal de 2015 construída com o tempo verbal PPS e outra relativa ao eixo temporal de 2013 edificada com o mais-que-perfeito composto (*tinha sido* e *tínhamos testado*).

Este exemplo mostra como as formas linguísticas e os valores semânticos que estas assumem estão interdependentes e não são estáticas. O exemplo em análise mostra como a anotação morfosintática não é suficiente para dar conta das relações de interdependência que constroem as referências temporais e que são atualizadas através de operações de referenciação, a saber uma referência temporal relativa ao momento de enunciação (T0), outra relativa ao localizador 2015 e uma terceira a 2013, retomada anaforicamente pela expressão temporal "na altura". E que o valor dos verbos (valores de anterioridade, posterioridade e simultaneidade) é atualizado em função destes localizadores e da relação que se estabelece entre eles.

Qualquer anotação que queira dar conta destas operações tem de ter em conta três aspetos: i) a ancoragem com o momento de enunciação/produção do texto; ii) a coordenação entre os elementos de construção de referenciação temporal e iii) o valor que eles adquirem nessas operações.

Na seção seguinte, iremos mostrar que, através de uma linguagem de marcação flexível como o XML, é possível anotar e relevar estas operações.

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

3 Modelo(s) de anotação: abordagens e limites

Nos últimos anos têm-se desenvolvido diversos trabalhos de caráter experimental e exploratório para a anotação de elementos semânticos nos textos. A anotação da deixis ainda não foi objeto de trabalhos específicos de anotação. Seja pela sua natureza abstrata, seja pela necessidade de mapeamento das operações abstratas que lhe estão subjacentes, como já referimos, a anotação da deixis coloca dois problemas ao nível técnico: (i) a necessidade de ligar elementos com formas e funções distintas (a ancoragem temporal relativamente ao momento da enunciação, por exemplo) e (ii) o mapeamento dessas mesmas operações de um modo que possa ser observável. Neste sentido procurámos trabalhos que oferecessem propostas para a resolução destes desafios técnicos e, por isso, analisámos algumas propostas de anotação que, não incidindo especificamente sobre a deixis, oferecessem soluções aproximadas. De entre os vários trabalhos analisados (GOECKE, LIINGEN, METZING, STIIHRENHERG, 2010), encontrámos alguns estudos para a anotação de coreferências e relações anafóricas, como Recasens, Martí; Taulé (2007) e Recasens, Martí; Taulé (2007b), este último abordando o conceito de *bridging* e oferecendo uma proposta de anotação em vários níveis (*multilevel*). Embora sejam propostas orientadas para a resolução de anáforas, oferecem soluções flexíveis que podem ser adaptadas à deixis. De acordo com os objetivos do trabalho, os autores propõem uma metodologia de anotação que dê conta de dois tipos de ligações correferenciais: a deixis do discurso (*discourse deixis*) e as relações anafóricas. É importante esclarecer que a deixis do discurso é entendida e definida pelos autores como "as reference to a discourse segment, that is, to a non-nominal antecedente" (Recasens et al., 2007, p. 205), uma anáfora cuja referência é não-nominal, reservando as que têm como antecedente um NP ou VP para a anáfora propriamente dita:

> Our approach classifies bridging (or associative anaphors) those definite or demonstrative NPs that are interpreted on the grounds of a metonymic relationship with a previous NP or VP. (RECASENS et al, 2007, p. 205)

O uso do termo *deixis* afasta-se, portanto, da nossa definição. No entanto, este trabalho apresenta algumas vantagens, em relação a outras propostas, e que consideramos úteis e adaptáveis para o nosso objetivo. A primeira vantagem é que é orientado para a anotação da língua espanhola que, devido à proximidade com o português, lida com algumas idiossincrasias da língua (como as três formas dos pronomes demonstrativos) que não existem em inglês. A segunda vantagem, é o facto de os autores abordarem o texto enquanto uma unidade de sentido: "the text taken as a scene in the sense that it builds up both a textual and a contextual framework as the result of an interaction between the discourse and the global context". (RECASENS, et al, 2007, p. 206). Esta abordagem "global" ao texto aproxima-se da ideia de texto enquanto objeto comunicativo (BRONCKART, 1997) e objeto complexo (COUTINHO, 2003). No entanto, existe um outro ponto no qual o nosso trabalho diverge de Recasens et al. (2007) e que concerne à metodologia de análise. Se em Recasens et al (2007), a análise é metodologicamente ascendente, ou seja, parte das unidades linguísticas específicas para níveis mais amplos, a nossa metodologia análise sustenta-se numa abordagem descendente, na esteira dos trabalhos de

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

Bakhtine; Volochinov (1929/1977) e de Bronckart (1997/1999). Assim, a análise parte em primeiro das atividades de linguagem, passa pelos géneros de texto, enquanto formatos textuais formados pelas gerações anteriores, em seguida pelos textos enquanto materialização linguística dos géneros e objetos comunicativos complexos, para finalmente alcançar as unidades linguísticas que enformam os textos (BRONCKART, 1997, RASTIER, 2001, COUTINHO, 2003).

> Les textes sont des produits de l'activité humaine, et à ce titre (...) ils sont articulés aux besoins, aux intérêts et aux conditions de fonctionnement des formations sociales au sein desquelles ils sont produits. (BRONCKART, 1997, p.74).

Esta noção de texto implica que qualquer tipo de anotação, sobretudo da deixis, terá de ser sensível e, de certo modo, revelar esta ligação entre o contexto em que é produzido o texto e as unidades linguísticas presentes que edificam e revelam esta ligação.

O modelo de anotação que propomos, de acordo com a metodologia descendente, possibilita uma análise linguística que tem em conta a influência dos fatores físico-culturais que atuam na criação dos valores semânticos e discursivos adquiridos e estabilizados no texto.

Na seção seguinte, iremos fazer uma breve descrição do *corpus* utilizado neste trabalho e iremos mostrar alguns exemplos práticos.

3.2 O *corpus*

Como *corpus* de análise, foram utilizados textos selecionados, recolhidos no âmbito das atividades do grupo Gramática e Texto integradas no CoRus - Projeto Estratégico 2015-2020, desenvolvido no Centro de Linguística da Universidade Nova de Lisboa (CLUNL). Os textos dizem respeito a comentários e foram selecionados de acordo com a canonicidade, a representatividade e a atividade da linguagem em que se inseriram. Por canonicidade entendemos o estatuto de prestígio social, cultural, económico que uma determinada prática textual goza durante um certo período e que reflete uma legitimação consensual e normativa. A representatividade reflete-se nas várias formas de texto e não na representatividade estatística do mesmo. As atividades da linguagem em que se inserem os textos foram selecionadas pela representatividade que esta prática textual tem nestas áreas, nomeadamente, a jornalística, a académica e a jurídica. Os textos foram, depois, analisados pelo anotador categorial LX-Tagger e etiquetados em formato XML.

3.3 Estrutura e anotação

A anotação de um *corpus* permite tornar visível informação sobre o texto, seja informação de natureza implícita ou explícita (como autoria, data, fonte, entre outras) ou informação que está codificada na disposição física do texto (fronteiras de palavras, parágrafos ou frases, por exemplo), e para realizar estes processos são necessários tanto um processo conceptual como um processo técnico. De forma breve, consideramos o primeiro como o

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

modelo que parte de um conceito teórico (a sintaxe, a morfologia, a semântica, entre outras) e o segundo compreende a execução técnica da anotação, dependente do sistema de anotação escolhido para o processo. Assim, e para simplificar a terminologia usada neste trabalho, vamos designar o modelo conceptual de nível e a realização técnica de camada (GOECKE, LIINGEN, METZING; STIIHRENHERG, 2010).

A anotação foi organizada em várias camadas, cada uma englobando funções diferentes:

- Estrutural: engloba as anotações que definem a estrutura do documento (parágrafos, frases, corpo de texto, peritexto).
- *Inline annotation*: engloba anotações linguísticas que descrevem um elemento estrutural único (POS tag, lemmas, etc.)
- *Span annotation*: engloba anotações linguísticas que descrevem um ou mais elementos estruturais (entidades, expressões multipalavras, dependências, etc). Esta anotação está integrada num nível mais abrangente (como a frase ou parágrafo).

É nesta última camada que foram inseridas as marcações dos deíticos uma vez que a deixis é uma função da linguagem que se insere em diversas categorias gramaticais e opera ao nível (con)textual. Assim, a deixis foi incluída na *span annotation* <sa> como um *element* específico. Dentro deste *element* são incluídos os NP, VP ou SP, frase ou frases que tenham uma função deítica, como nos exemplos seguinte:

(1a)
No verão de 2015

(1b)
<sa function="deitic" type="time"> <w ort="EM"> <lex cat="PREP"></lex> No </w> <w ort="VERÃO"> <lex cat="CN"> </lex><msd gender="m" number="s"></msd> verão </w> <w ort="DE"> <lex cat="PREP"></lex> de </w> <w> <lex cat="DGT"> </lex> 2015 </w> </sa>

No exemplo (1b) podemos observar como através da função *element* podemos inserir os deíticos temporais da frase, neste caso introduzido pela preposição de tempo *Em*. A informação é de caráter descritivo com dois atributos: a função (*function*) e o tipo (*type*). Para a função definiram-se duas funções específicas, a deítica e a anafórica (embora não se enquadre neste trabalho, fica a referência para futuras anotações). Relativamente ao tipo de deítico, definiram-se três tipos de acordo com os deíticos possíveis: pessoais (*personal*), espaciais (*space*) e temporais (*time*). De notar que a proposta aqui apresentada é um exemplo específico que não contempla todos os elementos com função deítica, como os morfemas de tempo e de pessoa. Essa ancoragem só é possível se for feita dentro da descrição morfossintática do verbo. Nesse sentido, tornou-se necessário estabelecer uma ligação entre o elemento (morfológico ou outro) ao qual é atribuído o valor deítico e a anotação que descreve o valor deítico atribuído, e é nesta ligação que se estabelece a relação entre o nível e a camada. Definiu-se, por isso, dois atributos para o <sa>: o primeiro será *ref* e irá incluir um caminho *XPath* para o verbo em questão e o segundo *value* irá explicitar o valor que o deítico tem no texto em relação ao tempo verbal (posteridade, anterioridade, simultaneidade):

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

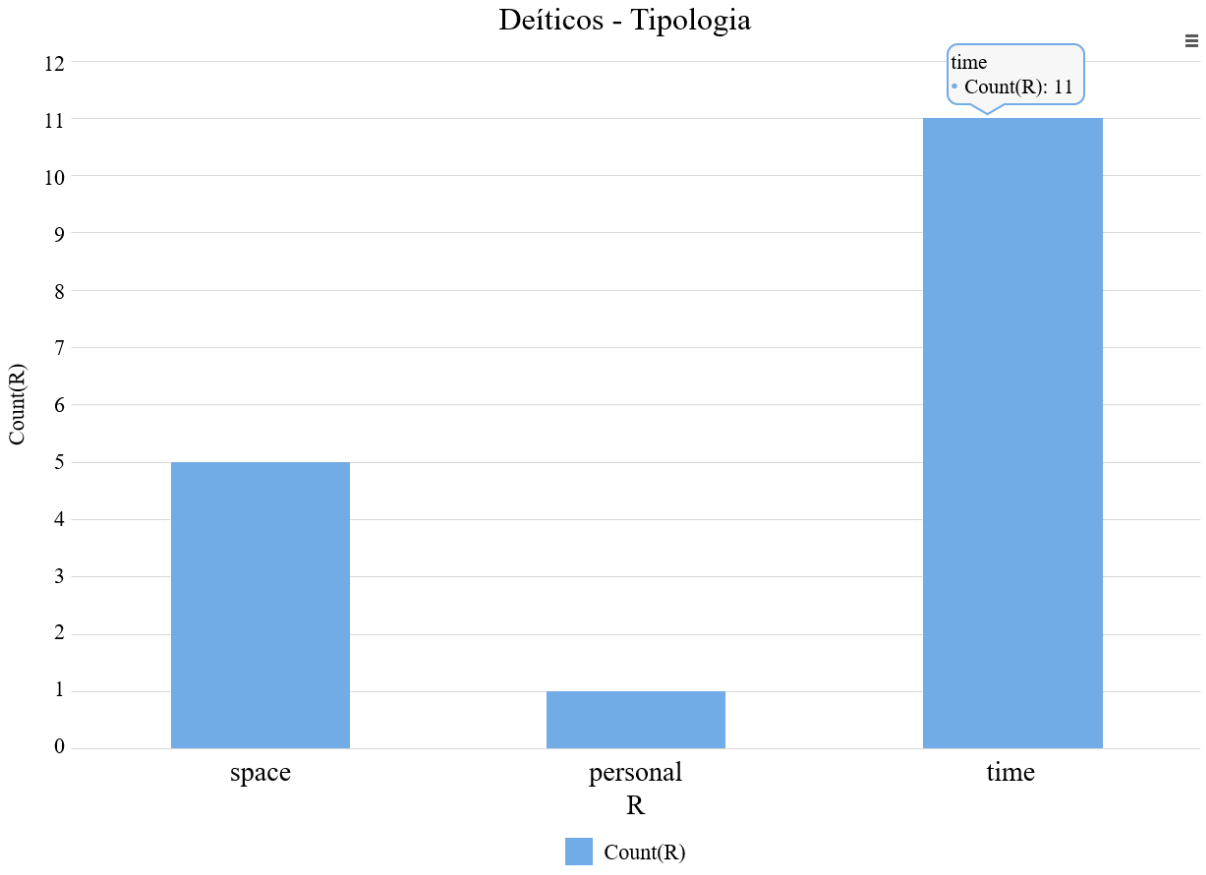
(2a)
Chegámos à CES 2015

(2b)
<sa function="deitic" type="space"> <w ort="CHEGAR"> <lex cat="V"></lex> <msd number="p" person="1" mood="ind" tense="pp"> </msd> Chegámos </w> <w ort="A"> <lex cat="PREP"></lex> à </w> <w ort="CES"> <lex cat="PNM"> </lex> CES </w> </sa>
<sa function="deitic" type="personal" ref="/Comentários[1]/comentário[1]/body[1]/p[1]/sent[1]/sa[1]/w[1]/msd[1]"></sa>
<sa function="deitic" type="time" ref="/Comentários[1]/comentário[1]/body[1]/p[1]/sent[1]/w[1]" value="ant"> <w> <lex cat="DGT"> </lex> 2015 </w></sa>

Do mesmo modo, os atributos *function* e *type* podem ser utilizados para os deíticos espaciais ("Chegámos à CES") e pessoais. Neste último, o objeto sobre o qual incidiu a <sa> foi a descrição morfosintática <msd> do verbo. Embora não seja visível, o XPath permite, neste exemplo, recuperar o atributo da <msd> que contém a informação sobre o tempo ou a pessoa.

Esta anotação que aqui se propõe permite recuperar os dados anotados, filtrá-los e analisá-los de várias formas:

Figura1: Deíticos temporais

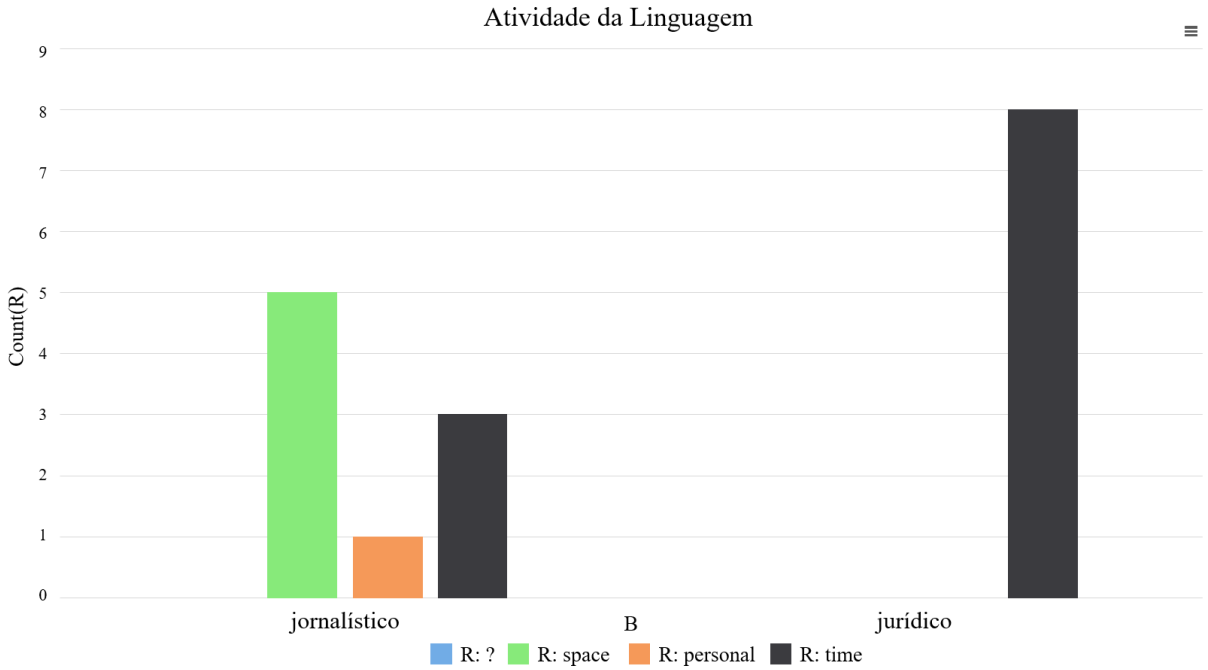


Fonte: elaborado pelos autores.

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

Na figura 1 podemos observar, por exemplo, que os deíticos temporais sobressaem nesta amostra. Mas, através de ferramentas de *text mining*, podemos também co-relacioná-los com a atividade da linguagem ou qualquer outra variável que estiver presente na anotação:

Figura 2: Relação entre deíticos e atividade de linguagem



Fonte: elaborado pelos autores.

Na figura 2, podemos observar a relação entre os diversos tipos de deíticos e a atividade da linguagem em que se inserem os textos (nesta amostra, a atividade jornalística e a atividade jurídica).

Conclusão

De caráter exploratório e experimental, o presente artigo visou apresentar uma proposta de anotação de *corpus* em XML possibilitando quantificar os deíticos (pessoais, temporais e espaciais) nos textos, bem como a evidenciação e a visualização das relações entre os diversos elementos da deixis e desses com as atividades de linguagem nas quais os textos circulam.

A deixis, sendo a face visível de um conjunto de operações abstratas, necessita de uma camada de anotação que opere em diferentes níveis de análise - morfológicos, sintáticos, semânticos e textuais. Esta característica da deixis evidencia como a categorização e etiquetagem de elementos semânticos e textuais coloca vários desafios técnicos que requerem um modelo flexível, capaz de atuar em diversos patamares (gramaticais e (con)textuais). Para tal, a anotação foi organizada em diversas camadas, com funções diferentes, agrupando nelas

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

diferentes funções e níveis de anotação, e que permitem recolher e visualizar informação sob vários ângulos e variáveis. No que toca à deixis, a marcação dos elementos deíticos realiza-se na última camada designada por *span annotation*. Esta camada engloba tipos de anotação (textual) que abrangem um ou mais elementos estruturais. A própria camada está embutida num nível estrutural de escopo mais amplo do que a palavra ou a frase, se necessário.

De destacar que a proposta de anotação, aqui apresentada, possibilita a visualização da sistematicidade dos valores que os deíticos adquirem no seio de um texto e de os analisar sob diversos ângulos. Estamos em crer que apresenta um grande potencial para colmatar a necessidade de criar ferramentas flexíveis atuando a níveis textuais meso e macro e em *corpus* menos extensos mas escaláveis. Sendo um trabalho exploratório, iremos continuar a desenvolver e a testar a metodologia, aumentando o número de textos e desenvolvendo novas soluções.

Referências

BAKHTINE, M., VOLOCHINOV, V. N. *Le marxisme et la philosophie du langage:* essai d'application de la méthode sociologique en linguistique. Préface de Roman Jakobson. Traduit du russe preésenté par Marina Yaguelo. Paris, Minuit, 1977.

BRONCKART, J.-P. Activité langagière, textes et discours. Pour un interactionisme socio-discursif. Paris: Delachaux et Niestlé, 1997.

CORREIA, C. N. ; PEREIRA, S. Formas e construções linguísticas no português europeu: Ferramentas referenciais e género textual. Cadernos de Linguagem e Sociedade, v. 16, n. 1, p. 48–60, 2015.

COUTINHO, M. A.; MIRANDA, F. To describe genres: Problems and strategies. In: BAZERMAN, C.; BONINI, A.; FIGUEIREDO, D. (Ed.). Genre in a Changing World. Fort Collins: The WAC Clearinghouse, 2009, p. 35-55. Available in: http://wac.colostate.edu/books/genre/ chapter3.pdf. Acesso em: 13 de nov. 2020.

COUTINHO, M. A. Texto(s) e competência textual. Lisboa: Fundação Calouste Gulbenkian-Fundação para a Ciência e a Tecnologia, 2003.

CULIOLI, A. Sur le concept de notion. Bulletin de Linguistique Appliqué e Générale, n. 8, p. 62-79, 1981.

CULIOLI, A. Pour une linguistique de l'énonciation. Formalisation et opérations de repérage (t2). Paris: Ophrys, 1999.

FONSECA, F. I. Deixis, tempo e narração. Porto: Fundação Eng. António de Almeida, 1992.

GOECKE, D.; LIINGEN, H.; METZING, D., STIIHRENHERG, M. Different Views on Mark up Distinguishing Levels and Layers. In: WITT, A; METZING, D. (Eds.), Linguistic Modeling of Information and Markup Languages. Contributions to Language Technology. Dordrecht: Springer, 2010, p. 1-22.

HARDIE, A. Modest XML for Corpora: Not a standard, but a suggestion. ICAME Journal, v. 38, n. 1, 2014, p. 73–103. Available in: https://doi.org/10.2478/icame-2014-0004

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.

HOBBS, J. R.; PUSTEJOVSKY, J. Annotating and Reasoning about Time and Events. Working Papers of the 2003 {AAAI} Spring Symposium on Logical Formalization of Commonsense Reasoning, p. 74–82, 2003.

LEVINSON, S. C. The Handbook of pragmatics (L. R. Horn; G. Ward, Eds.). Choice Reviews Online, v. 41, Blackwell Publishing Ltd, 2006. Available in: https://doi.org/10.5860/choice.41-6349

LYONS, J. Deixis, Space and Time. In: STEINBERG, D. D.; JAKOBOVITS, D. D. (Eds.). Semantics. Cambridge: CUP, 1977, p. 636-724.

MIRANDA, F. Textos e géneros em diálogo: uma abordagem linguística da intertextualização. Lisboa: FCG/FCT, 2010.

RASTIER, F. Arts et Sciences du Texte. Paris: P.U.F, 2001.

RECASENS, M., MARTÍ, M. A.; TAULÉ, M. Where anaphora and coreference meet. Annotation in the Spanish CESS-ECE corpus. International Conference Recent Advances in Natural Language Processing, RANLP, p. 504-509, 2007.

RECASENS, M., MARTÍ, M.; TAULÉ, M. Text as scene: discourse deixis and bridging relations. Procesamiento del Lenguaje Natural. Sociedad Española para el Procesamiento del Lenguaje Natural Jaén, n. 39, p. 205-212, 2007b.

VALENTIM, H. T. Deixis in European Portuguese: Representation and Reference Construction. In: JUNGBLUTH, K.; DA MILANO, F. (Ed.). Manual of Deixis in Romance Languages. Berlin/Bos: Mouton De Gruyter, 2015, p. 247-314.

Recebido: 28/02/2021.

Aprovado: 10/06/2021.

Todo conteúdo da *Linha D'Água* está sob Licença Creative Commons CC -By 4.0.