http://dx.doi.org/10.1590/1678-460X2020360203

D E L T A

Artigos

Desambiguação das preposições do português no livro terceiro da Clavis Prophetarum[1]

Disambiguation of Portuguese prepositions in Clavis Prophetarum's third book

Carlos Assunção[2]
José Paulo Tavares[3]
Gonçalo Fernandes[4]

RESUMO

O livro terceiro da Clavis Prophetarum (Clavis III PT), da autoria do padre António Vieira, na sua tradução portuguesa do ano 2000, constitui o corpus deste trabalho. Para se poder trabalhar este corpus, precisamos de recursos linguísticos eletrónicos formalizados de forma a obter a maior cobertura possível e passíveis de serem utilizados em sistemas adequados. Para o Português dispomos já de recursos dignos de confiança, desenvolvidos desde a década de 90 pelo LabEL[5] – Laboratório

1. Este trabalho é financiado por fundos nacionais através da Fundação para a Ciência e a Tecnologia (FCT), no âmbito do Centro de Estudos em Letras, com a referência n.º UIDB/00707/2020.
2. Centro de Estudos em Letras. Universidade de Trás-os-Montes e Alto Douro – Portugal. http://orcid.org/0000-0002-5739-0754. E-mail: cassunca@utad.pt.
3. Centro de Estudos em Letras. Universidade de Trás-os-Montes e Alto Douro – Portugal. http://orcid.org/0000-0001-5674-2271. E-mail: josepantufatavares@gmail.com.
4. Centro de Estudos em Letras. Universidade de Trás-os-Montes e Alto Douro – Portugal. http://orcid.org/0000-0001-5312-6385. E-mail: gf@utad.pt.
5. Veja-se http://label.ist.utl.pt/pt/apresentacao.php.

de Engenharia da Linguagem. Este artigo tem como objetivo principal a elaboração de regras de desambiguação automática das preposições e respetivos sintagmas bem como a avaliação da eficácia da sua aplicação, de forma a permitir posteriores abordagens fiáveis no estudo desta categoria no corpus usando técnicas automáticas. A metodologia utilizada descreve pari passu as anotações, as ocorrências e a criação dos recursos eletrónicos considerados essenciais ao processo de desambiguação.

Palavras-chave: *Clavis Prophetarum; linguística de corpus; ambiguidade; desambiguação.*

ABSTRACT

The third book of Clavis Prophetarum (Clavis III PT), written by Father António Vieira, in its Portuguese translation published in 2000, is the corpus of this work. To be able to analyse this corpus, we need electronic language resources that are formalised in order to obtain the greatest possible coverage and to be used in appropriate systems. For the Portuguese language, we already have reliable resources, developed since the 90s by LabEL - Laboratory of Language Engineering. The main objective of this article is to develop rules of automatic disambiguation of prepositions and their syntagmata, as well as to assess the effectiveness of their application, in order to allow subsequent reliable approaches in the study of this category in the corpus using automatic techniques. The methodology used describes pari passu annotations, occurrences and the creation of electronic resources considered essential to the disambiguation process.

Keywords: *Clavis Prophetarum; corpus linguistics; ambiguation; disambiguation.*

Introdução

Graeme Kennedy, em *An Introduction to Corpus Linguistics*, reconhece o contributo das análises manuais de textos ao longo dos séculos, especialmente para a lexicografia, mas fala claramente na vantagem do uso do computador e da fiabilidade dos resultados obtidos neste segundo método sobre o primeiro. "The analysis of huge bodies of text 'by hand' can be prone to error and is not always exhaustive or easily replicable […]. The Corpus Linguistics is thus now inextricably linked to the computer, which has introduced incredible speed, total accountability, accurate replicability, statistical reliability and the ability to handle huge amounts of data" (Kennedy 1998: 5).

DELTA
36.2
2020

A pesquisa sobre corpus implica mais métodos indutivos do que métodos hipotético-dedutivos, por conseguinte as análises conduzidas pelos dados (data-driven) são preferidas em detrimento das análises conduzidas por regras (rule-driven).

A linguística de corpus já foi alvo de diversos estudos teóricos e teórico-práticos, no decurso do tempo. Investigadores como Stubbs (1993), Sinclair (2004), Leech (1992), Kennedy (1998), Tognini-Bonelli (2001), McEnery & Wilson (1996), Halliday (1967, 2006), Chomsky (1956), Teubert (2005), Bowker & Pearson (2002), McEnery *et al.* (2006), Beber Sardinha (2004), Gries (2010), Fadanelli & Monzón (2017), Silberztein (2004, 2015), entre muitos outros, levaram a cabo estudos no âmbito da Linguística de corpus e sobre as possibilidades da sua aplicação às línguas.

A ambiguidade é um dos maiores desafios que se coloca ao desenvolvimento de sistemas de processamento de linguagem natural e, consequentemente, à exploração de corpora, especialmente quando etiquetados. Gazdar e Mellish (1989: 7-8) distinguem entre **ambiguidade global**, quando uma frase pode ter mais do que uma estrutura, e **ambiguidade local**, quando uma parte do conjunto pode ter diferentes leituras, enquanto Small, Cottrell e Tanenhaus (1988:4) diferenciam ambiguidade **lexical**, quando uma palavra pode ter mais do que uma interpretação, de ambiguidade **estrutural**. Hutchins e Somers (1992: 85) estabelecem três tipos de ambiguidade lexical: (i) ***category ambiguity***, (ii) provocada por **homonímia** ou **polissemia** e (iii) ***transfer or translational ambiguities***.

Por se tratar de um assunto complexo no âmbito da linguística de corpus é nossa intenção construir recursos eletrónicos que descomplexifiquem a ambiguidade e favoreçam o processo de desambiguação no tocante às preposições de língua portuguesa da *Clavis Prophetarum*.

2. Ambiguidade

Numa abordagem orientada por um sistema de análise automática multinível, Bick (2000: 99) classifica os tipos de ambiguidade segundo os níveis **morfológico**, **sintáctico** e **semântico**, aventando ainda a possibilidade de um nível **pragmático**.

Silberztein (2006: 60-61), a respeito da construção de dicionários para uso no *NooJ*, refere a existência de **ambiguidade lexical** (quando uma palavra se associa a diferentes propriedades, por exemplo sintácticas ou distribucionais), o que implica uma duplicação das entradas, e de **ambiguidade morfológica** (quando uma palavra se associa a mais do que uma análise morfológica.

A resolução de ambiguidades, restringimo-nos à ambiguidade de escopo lexical, tem como objectivo eliminar rápida e eficazmente o maior número possível de análises incorretas que resultam da etiquetagem lexical, e pode ser levada a cabo de diversas formas. Tal como em outros aspetos do processamento da linguagem natural, a desambiguação pode basear-se numa abordagem puramente probabilística ou num sistema baseado em regras, havendo ainda a possibilidade de combinar ambas as técnicas. O modelo probabilístico necessita de um corpus de treino ou aprendizagem e faz uso dos *HMM* para atribuir a cada item a etiqueta mais provável, descartando as restantes possíveis.

A necessidade de um dicionário torna-se premente a partir de um nível sintático de análise, em que há sequências que funcionam ou como equivalentes a uma palavra (o caso das locuções e das formas verbais compostas), ou constituem unidades sintáctico-semânticas bem definidas, como é o caso das fraseologias, dos idiomatismos e de outras unidades como a colocação, termo introduzido na metalinguagem linguística por Firth (1957: 190-215), mostrando que o aspecto relevante do significado de uma palavra é o conjunto de todas as outras palavras que com ela se combinam, definindo-o como caracterização de uma palavra de acordo com outras palavras que tipicamente ocorrem com ela: "You shall know a word by the company it keeps!" (Firth 1968: 179). Como colocação entendemos "the habitual meaningful co-occurrence of two or more words (a node word and its collocate or collocates) in the close proximity to each other" (Halliday *et al.* 2004: 168). Posto isto, parece um dado adquirido que a etiquetagem lexical correcta é um subproduto da análise sintáctica, o que, em termos de processamento automático de grandes quantidades de texto, é um objectivo ainda distante. Uma resolução parcial, ou redução, das ambiguidades lexicais, não necessitando de uma análise sintáctica completa e sendo menos ambiciosa, é no entanto mais exequível e realista.

3. Metodologia e resultados

Qualquer que seja o método de desambiguação utilizado, é necessário ter sempre presente que o objectivo principal é o de eliminar a maior parte das análises incorretas preferencialmente todas, mas sem eliminar no processo as análises corretas.

O excesso de etiquetas é referido como taxa de ruído, enquanto que a eliminação de análises corretas corresponde à taxa de silêncio. Um sistema de desambiguação ótimo será aquele que mantém ambas no valor zero.

Utilizamos para esta tarefa o programa *NooJ*[6], criado por Max Silberztein (2002-2015). Para podermos analisar o corpus tivemos que criar uma cópia em formato *Unicode,* pois o *NooJ,* ainda que consiga trabalhar com outros formatos, nomeadamente com o formato *Ansi*, fá-lo em melhores condições em *Unicode.* Neste formato, ao contrário do que acontecia em *Ansi,* com o que o *NooJ* não reconhecia caracteres acentuados, por exemplo, todos os caracteres são reconhecidos pelo programa.

A tarefa a que nos propusemos foi a desambiguação das preposições, adoptando uma estratégia que começasse pela identificação das ambiguidades realmente presentes no corpus, pela construção e aplicação de gramáticas de resolução de ambiguidades que possam ser futuramente aplicadas a outros corpora e, finalmente, pela resolução manual das ambiguidades remanescentes. Fizemos uma primeira anotação do corpus que nos deu logo que há palavras que podem ser preposições ou pertencer a outras classes:

visto ocorre 19 vezes, doze das quais integrado na locução conjuncional *visto que*, seis como particípio do verbo ver e uma apenas como preposição, no contexto seguinte: “para que, tornando-se de novo cristão e crendo de novo em Cristo, se diga que não só há-de ser convertido mas restituído por Elias, ***visto*** ser devolvido e restituído ao seu verdadeiro e legítimo Senhor”.

6. Veja-se: http://www.nooj-association.org/

trás ocorre três vezes, sempre inserida na expressão *para trás*, naquilo que Bechara (2002: 301) designa por "acúmulo de preposições". Segundo este autor, este tipo de expressões não constitui uma locução prepositiva porque cada unidade vale por si só, juntando-se apenas "para dar maior efeito expressivo às idéias". Como nenhuma ocorrência da forma corresponde ao uso da interjeição, podem eliminar-se simplesmente todas as etiquetas que a identificam como tal.

As 82 ocorrências de *sobre* correspondem todas ao uso da preposição, pelo que, relativamente ao corpus em análise, se procederá a uma desambiguação grosseira que elimine todas as outras etiquetas.

Das 475 vezes em que surge a forma *para*, 69 são relativas à locução conjuncional *para que*, nove no acúmulo *para com* e uma ao *para entre*. As restantes ocorrências dizem respeito à preposição simples. Poderíamos providenciar que a análise lexical parasse na utilização do dicionário de unidades multipalavra, aplicado com alta prioridade, de forma a evitar que o *para* de *para que* receba outras etiquetas além de *CONJ*, ou aplicar uma regra de desambiguação de permita a manutenção apenas da etiqueta *CONJ* para a sequência *para que*.

Todas as seis ocorrências de *mediante* correspondem à preposição, pelo que serão simplesmente eliminadas todas as outras etiquetas.

Treze das 106 ocorrências de *entre* são relativas ao acúmulo *de entre*, uma ao já referido *para entre*, e as restantes à preposição. Basta eliminar todas as etiquetas que identifiquem a palavra como uma forma verbal. As 45 ocorrências de *contra* correspondem todas ao uso da preposição.

Doze das 101 ocorrências de *até* dão-se no seio da locução conjuncional *até que*, uma na *até porque*, duas no acúmulo *até a*, duas ao acúmulo *até em*, uma ao *até para*, outra ao *até de*, 17 ao uso do advérbio e as restantes ao uso da preposição.

A forma *a* ocorre 2113 vezes. Dada a ingente tarefa que seria analisar todas as 2113 ocorrências de *a*, no sentido de apurar exatamente quais delas correspondem ao uso da preposição, do pronome ou do artigo (assumindo, a partir do conhecimento prévio do corpus,

que nenhuma ocorrência corresponde ao uso do *a* enquanto nome), preferimos basear-nos em trabalhos prévios de elaboração de regras de desambiguação.

Assim, Costa (2001: 222-224) propôs cinco regras de desambiguação para *a*, considerando a distinção entre preposição, artigo e pronome.

Na primeira dessas regras estabelece que, se o *a* ocorrer precedendo um nome no feminino do singular, então o *a* deve ser etiquetado como artigo. Apesar de, por si só, esta regra identificar erroneamente o *a* da frase *O João foi a Lisboa* como artigo, a verdade é que se revela muito produtiva, podendo a taxa de silêncio ser reduzida se acrescentarmos algumas restrições, como por exemplo: *a* seguido de um nome no feminino singular é um artigo, a não ser que seja precedido por um verbo de movimento como *ir* ou *chegar*. Claro que, mesmo assim, podem surgir contextos que não permitem uma correta desambiguação, nomeadamente por causa de construções elípticas do tipo *Sabes onde foi o João? A Lisboa.* Neste caso, o *A* que precede *Lisboa* será erradamente etiquetado, mercê da dependência implícita do sintagma *A Lisboa* relativamente ao verbo *ir*; ou, em algumas circunstâncias, ao facto de, mesmo precedido de uma forma do verbo chegar (como em "não tinha chegado a pregação do Evangelho"), o *a* corresponder efetivamente a um artigo. Seríamos talvez tentados a concluir que, se o verbo *chegar* se encontrasse no particípio passado, o *a* seguinte seria um artigo, mas basta o exemplo seguinte para nos dissuadir de acrescentar essa restrição à regra: *Chegado a Lisboa, o João descansou*. No corpus *Clavis III PT*, das oito ocorrências em que o *a* é precedido por uma forma do verbo *chegar*, três dizem respeito a um *a* artigo, todos eles em contextos de utilização do verbo *chegar* não como um verbo de movimento para, mas como o resultado desse mesmo movimento, enquanto as restantes cinco dizem efetivamente respeito ao uso da preposição, em casos de utilização do verbo como verbo de movimento. Posto isto, parece lícito concluir que, para uma efetiva e eficaz desambiguação automática do *a* ocorrendo a seguir ao verbo *chegar*, seria necessário em primeiro lugar distinguir os dois valores semânticos do verbo, de modo a o *a* ser corretamente etiquetado como preposição apenas quando *chegar* transmite a ideia de movimento.

Outro problema desta regra é a ambiguidade existente entre algumas formas do verbo *ir* e do verbo *ser*: na frase "E tal foi a egressão ou saída, por assim dizer, efetiva, dos Apóstolos para ministrarem aquilo por que foram enviados", se não tiver sido previamente resolvida a ambiguidade de *foi* e este estiver etiquetado como pertencendo ao verbo *ir*, o *a* será erradamente etiquetado como preposição. Das doze vezes em que o *a* surge precedido por uma forma identificada como pertencente ao verbo *ir*, apenas duas correspondem à utilização da preposição. Na verdade, apenas nesses dois contextos a palavra precedente pertence ao verbo *ir*, sendo que nove dizem respeito a formas ambíguas, pertencentes realmente ao verbo *ser*, e no outro é usado o homónimo adjetival *vão*. Para a restrição acima mencionada, relativa ao verbo *ir,* funcionar corretamente seria portanto necessário, em primeiro lugar, resolver as ambiguidades relativas aos seus homónimos.

Dada esta análise, não nos parece produtivo, nestas circunstâncias, acrescentar as referidas restrições à primeira regra proposta por Costa (2001: 222), até porque, das 1170 linhas de concordância produzidas a partir do corpus *Clavis III PT* não desambiguado pedindo todas as ocorrências em que o *a* seja seguido por um nome no feminino singular, a grande maioria diz efetivamente respeito à utilização do *a* como artigo, exceptuando na relativa à frase "Por tal motivo é extremamente vão e do domínio da fábula que o homem tente, voando com as asas de Ícaro, chegar a essa altitude inacessível aos próprios anjos", em que outra ambiguidade vizinha, a da palavra *essa* (que tanto pode ser nome como determinante ou pronome) provoca a inserção da expressão na concordância, apesar de, neste contexto, *essa* ser um determinante. Além disso, claro que deve ser referido o facto de o *a* ser uma preposição (precedida por *chegar*), o que seria corretamente conseguido em termos de etiquetagem através da regra acima enunciada. Assim, considerando que à forma *a* são atribuídas inicialmente 5 etiquetas diferentes, a aplicação desta simples regra de desambiguação permite a eliminação de 4680 etiquetas mantendo a taxa de silêncio praticamente no zero.

A segunda regra estabelece que um *a* precedido de um gerúndio é um artigo, o que, sendo verdade para expressões como *facilitando a leitura*, não o é quando esse gerúndio pertence a um verbo de movimento como *ir* ou *chegar*, como comprova a frase "Indo a todo o

DELTA
36.2
2020

mundo, pregai o evangelho a toda a criatura", ou quando ao gerúndio se segue um complemento circunstancial introduzido pela preposição, como nos exemplos (também do corpus *Clavis III PT*) "alegando a esse propósito", "caminhando a passo tranquilo" e "filosofando a partir de causas meramente naturais".

Se descontarmos estes quatro exemplos às 43 ocorrências de um gerúndio seguido de um *a*, ficamos com 39 casos em que a segunda regra funciona de forma correta, o que justifica a sua aplicação, sobretudo se considerarmos que a aplicação das regras é recursiva, e que podemos (e devemos) elaborar regras que prevejam que, por exemplo, não seja atribuída a etiqueta de artigo a um *a* que preceda determinantes masculinos (*todo* e *esse*), nomes masculinos (*passo*) ou verbos (*partir*). Na verdade, podemos até evitar a aplicação de qualquer etiqueta à forma isolada *a* na expressão *a partir de*, se a considerarmos como uma unidade multilexémica e previrmos a paragem da análise lexical neste ponto.

A terceira regra, que estabelece que um *a* entre um particípio passado à esquerda e um pronome possessivo à direita é um artigo, aplica-se com total exatidão ao corpus *Clavis III PT*, uma vez que apenas se encontra uma ocorrência com esta configuração em que, efetivamente, o *a* é um artigo.

A regra quatro prevê que a um *a* antes de um infinitivo verbal se aplique a etiqueta de preposição, e pode aplicar-se também com total confiança, uma vez que, no corpus em estudo, as 204 ocorrências relativas a esta configuração (incluindo o infinitivo pessoal) correspondem efetivamente ao uso da preposição: 1016 etiquetas erradas eliminadas, mantendo a taxa de silêncio em zero.

O mesmo se pode dizer relativamente à quinta regra, que determina que o *a* antes de um artigo indefinido corresponde a uma preposição, o que é rigorosamente verdade nas 23 ocorrências de *a* seguido por artigo indefinido no *Clavis III PT*.

Partindo das regras elaboradas por Costa (2001), foi construída uma gramática de resolução de ambiguidades para a forma *a* representada pelo grafo seguinte:

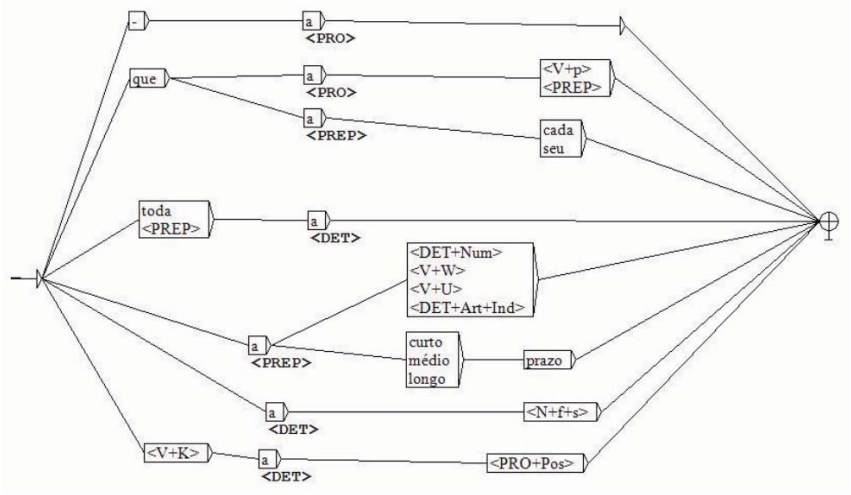

Figura 1 – FST de desambiguação de *a*

Passaremos agora a descrever as regras definidas nesta gramática, procedendo à descrição do efeito de cada caminho do grafo e dos seus efeitos na etiquetagem, começando de cima para baixo.

A primeira regra define que ao *a* imediatamente antecedido por um hífen serão atribuídas as etiquetas que o identificam como pronome (em posição enclítica). Esta regra não prevê a desambiguação entre pronome pessoal e demonstrativo, pelo que serão mantidas duas etiquetas para cada uma das 13 ocorrências do *a* nestas condições. Desambigua por exemplo: *recebendo-a*, *resume-a*.

A segunda regra diz que o *a* é um pronome se for antecedido da palavra *que* e a seguir estiver um verbo no plural (“daqueles que a sofreram”) ou uma preposição, e resolve (parcialmente, uma vez que continua a persistir a dualidade pessoal/demonstrativo) a ambiguidade de quatro ocorrências no corpus. Especificámos aqui o plural no verbo a seguir ao *a* porque, embora no corpus o *a* pronome ocorra também entre um *que* e um verbo no singular, são muitos os casos de ambiguidade com homónimos substantivos ou infinitivos, em que o *a* corresponde ora a artigo ora a preposição.

O terceiro caminho estabelece que um *a* entre um *que* e a forma *cada* ou *seu* é uma preposição. No corpus em estudo, são resolvidas apenas duas ocorrências com esta regra: "O restante, que a seu tempo deve ser revelado" e "o que é que a cada momento nos convém dizer aos outros".

A quarta regra diz que o *a* é um determinante se estiver antecedida pela palavra *toda* ou por uma preposição. Resolve 370 casos de ambiguidade no corpus, entre as quais: "com a autoridade da escritura" e "toda a matéria deste livro".

A quinta regra define que *a* é uma preposição sempre que seguida de um numeral, de um infinitivo (incluindo o flexionado) ou de um artigo indefinido. Resolve sem taxa de silêncio 282 ocorrências ambíguas, entre elas: "começar a abrir", "Ouçamos a dupla diferença" e "suceder a um homem".

a é ainda preposição quando inserida nas expressões *a curto/médio/longo prazo* (sexta regra). Nenhuma destas expressões ocorre no corpus, mas achámos pertinente manter esta regra, para eventuais trabalhos futuros.

A sétima regra corresponde à primeira definida por Costa (2001), com os efeitos acima referidos. O mesmo se pode dizer da oitava (terceira de Costa 2001).

Esta gramática resolve 1879 ocorrências ambíguas, num total de 2113.

Como vimos acima, as formas *trás*, *sobre*, *mediante* e *entre*, embora por natureza ambíguas, ocorrem sempre como preposições no corpus. Neste caso, poderíamos evitar a etiquetagem ambígua retirando do dicionário aplicado as entradas que identificam as referidas palavras como pertencentes a outras classes que não a preposição, de modo a receberem apenas as etiquetas adequadas no momento da análise lexical.

No entanto, esta operação só deve ser levada a cabo se houver um conhecimento prévio aprofundado do corpus, de modo a não eliminarmos do dicionário etiquetas correspondentes a efetivas realizações no corpus.

Além disso, a tarefa de identificação no dicionário das entradas a serem eliminadas e de compilação de novas versões (ou mesmo de novos dicionários) é complexa e demorada, pelo que é preferível construir regras de desambiguação que permitam a eliminação *a posteriori* das etiquetas erradas. Sobretudo porque nestes casos não é necessário inserir nas gramáticas de desambiguação qualquer tipo de restrição lexical e/ou contextual.

Assim, para as quatro formas acima referidas nestas condições, foi construído um grafo que permite apenas a manutenção, para cada uma das ocorrências, das etiquetas que as identificam como preposição e desambigua 197 ocorrências:

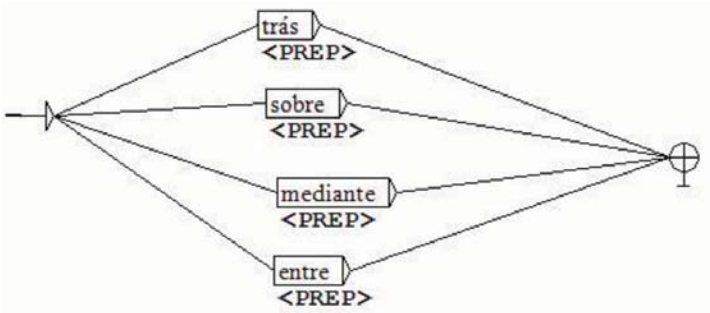

Figura 2 – FST de desambiguação de *trás, sobre, mediante* e *entre*

Importa realçar que este grafo é válido apenas para o corpus em estudo, não obstante poderem ser constituídos outros corpora em que se venha a verificar a sua validade.

Para *visto* fez-se a seguinte gramática:

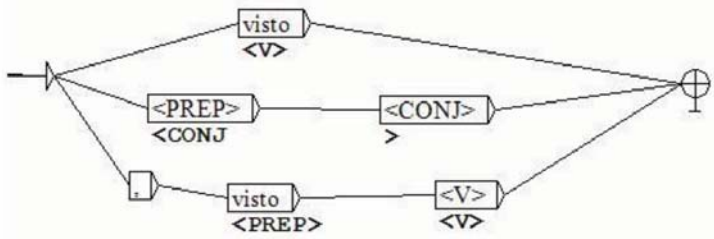

Figura 3 – FST de desambiguação de *visto*

DELTA
36.2
2020

Esta define que: estando *visto* antes de uma conjunção, ambas devem receber a etiqueta CONJ: "visto que o próprio mundo"; *visto* entre uma vírgula e um verbo é uma preposição: ", visto ser devolvido…"; noutras circunstâncias, *visto* é um verbo.

Para resolver a ambiguidade relativa à sequência *para que* elaborou-se a seguinte gramática:

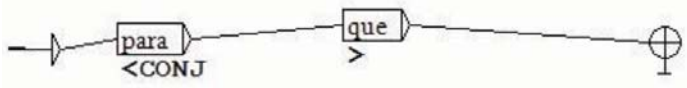

Figura 4 – FST de desambiguação da locução *para que*

Fez-se de seguida o levantamento das preposições que ocorrem contraídas com outras palavras, nomeadamente com advérbios (*PREPXADV*), determinantes (*PREPXDET*) e pronomes (*PREPXPRO*), tendo-se chegado à conclusão de que, no corpus desambiguado, se encontram 4671 formas etiquetadas como contrações cujo primeiro elemento é uma preposição, correspondendo 38 delas à categoria *PREPXADV*, a 4499 são atribuídas etiquetas que as identificam como *PREPXDET* e a 4619 etiquetas referentes a *PREPXPRO*. O facto de a soma destas categorias, contabilizadas separadamente, ser muito superior ao da consulta feita em conjunto deve-se ao elevado número de ambiguidades entre as categorias *DET* e *PRO*, que praticamente duplicam o número de etiquetas.

De início, considerando que, no momento, o interesse deste estudo se centrava na preposição, pensei não envidar esforços no sentido da resolução deste tipo de ambiguidades, cingindo-me à resolução daquelas em que a uma forma tanto podem ser aplicadas etiquetas relativas à contração de preposições com outros elementos como outras etiquetas, como por exemplo no caso de *nos* que, além de etiquetas relativas a *PREPXDET* e *PREPXPRO*, recebe mais cinco que o identificam como pronome.

No entanto, considerando que estas formas, no seu conjunto, ocorrem 3315 vezes no corpus, tentaremos elaborar algumas regras de desambiguação, ainda que não sejam passíveis de aplicação a outros corpora.

São 68 as formas diferentes concernentes à contração da preposição com outros elementos, a saber: *à, ao, aonde, aos, àquela, àquelas, àquele, àqueles, àquilo, às, connosco, consigo, da, daí, daquela, daquelas, daquele, daqueles, daqui, daquilo, das, dela, delas, dele, deles, dessa, dessas, desse, desses, desta, destas, deste, destes, disso, disto, do, donde, dos, na, naquela, naquelas, naquele, naqueles, naquilo, nas, nela, nelas, nele, neles, nessa, nessas, nesse, nesses, nesta, nestas, neste, nestes, nisso, no, nos, noutra, noutro, num, numa, pela, pelas, pelo* e *pelos*.

Destas, as formas não ambíguas são 16:

àquilo (PREPXPRO)
connosco (PREPXPRO)
daí (PREPXADV)
daqui (PREPXADV)
daquilo (PREPXPRO)
dela (PREPXPRO)
delas (PREPXPRO)
dele (PREPXPRO)
deles (PREPXPRO)
disso (PREPXPRO)
naquilo (PREPXPRO)
nela (PREPXPRO)
nelas (PREPXPRO)
nisso (PREPXPRO)
num (PREPXDET)
numa (PREPXDET)

São também várias (33) as formas cuja única ambiguidade é estarem entre *PREPXDET* e *PREPXPRO*: *à, ao, aos, àquela, àquelas, àquele, àqueles, às, da, daquela, daquelas, daquele, daqueles, das, desta, destas, do, dos, naquela, naquelas, naquele, naqueles, nessa, nessas, nesse, nesses, nesta, nestas, neste, nestes, noutra, noutro*, e *pelos*.

A forma *donde* tanto pode ser *PREPXADV* como *PREPXDET* ou *PREPXPRO*.

Para resolver estas ambiguidades, foi construída a seguinte gramática *ad hoc*, a partir da análise dos contextos das ocorrências no corpus, feita apenas para desambiguação deste:

36.2
2020

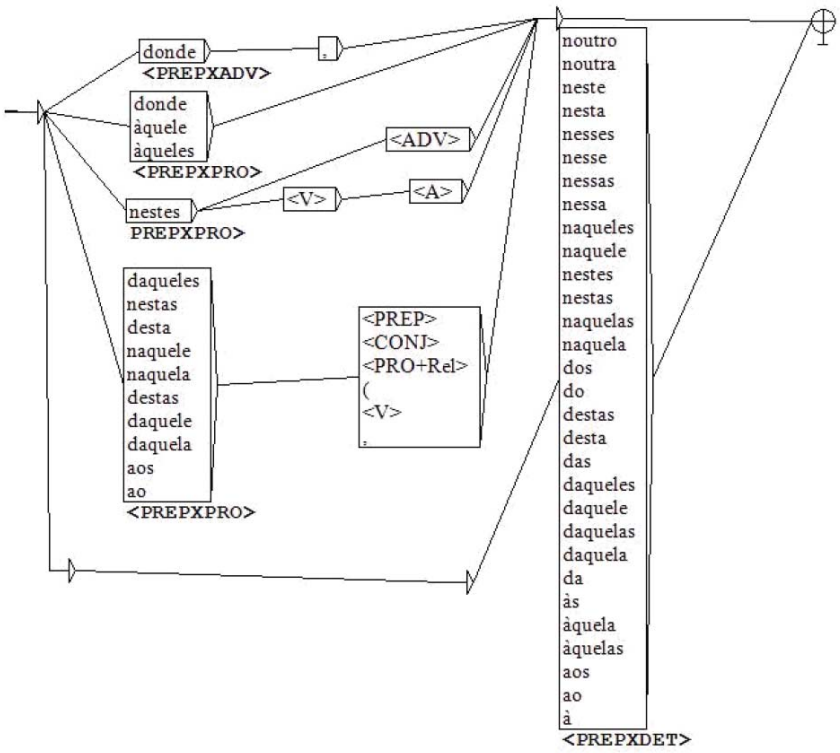

Figura 5 – FST de desambiguação de contrações de preposições com outros elementos - 1

Segundo esta gramática, donde é contração de preposição e advérbio quando seguido de vírgula ("donde, parecia incrível que…"), sendo PREPXPRO noutras circunstâncias ("as donde devem ser trazidas?"); àquele e àqueles são sempre, neste corpus, contrações de preposição e pronome: "Àqueles a quem os apóstolos…", "àquele que sobe"; nestes é PREPXPRO se seguido de advérbio ou de forma verbal e adjetivo ("Nestes não se procura a salvação das almas", "Nestes estão figurados…". Noutras circunstâncias, é PREPXDET: "nestes termos", "Nestes passos"; as formas daqueles, nestas, desta, naquele, naquela, destas, daquele, daquela, aos e ao são PREPXPRO se seguidas de preposição, conjunção, pronome relativo, verbo, vírgula ou abertura de parênteses rectos. Noutras circunstâncias, são PREPXDET, bem como todas as outras formas inseridas no nó mais à direita que, não estando presentes noutros nós do grafo, ocorrem no corpus apenas como PREPXDET.

São portanto 18 as formas que, além de etiquetas relativas à contração da preposição com outros elementos, recebem outras concernentes a outras classes:

aonde (ADV e PREPXADV)
consigo (PREPXPRO e forma do verbo conseguir)
dessa (PREPXPRO, PREPXDET e 4 formas do verbo dessar)
dessas (PREPXPRO, PREPXDET e 1 forma do verbo dessar)
desse (PREPXPRO, PREPXDET, 4 formas de dessar e 3 de dar)
desses (PREPXPRO, PREPXDET, uma forma de dar e duas de dessar)
deste (PREPXPRO, PREPXDET e 2 formas do verbo dar)
destes (PREPXPRO, PREPXDET e 1 forma de dar)
disto (PREPXPRO e forma do verbo distar)
na (PREPXPRO, PREPXDET e 2 formas do pronome pessoal)
nas (PREPXPRO, PREPXDET e 2 formas do pronome pessoal)
nele (PREPXPRO e nome)
neles (PREPPRO e nome)
no (PREPXPRO, PREPXDET e 4 formas do pronome pessoal)
nos (PREPXPRO, PREPXDET e 5 formas do pronome pessoal)
pela (PREPXPRO, PREPXDET e 4 formas do verbo pelar)
pelas (PREPXPRO, PREPXDET e 1 forma do verbo pelar)
pelo (PREPXPRO, PREPXDET e 1 forma do verbo pelar)

No corpus, *pelo*, que ocorre 101 vezes, nunca é forma do verbo *pelar*, sendo que se podem eliminar à partida estas etiquetas. O mesmo se pode dizer da forma *pelas*, que surge 24 vezes, e de *pela*, 124 vezes.

Das 123 ocorrências de *nos* 62 são relativas ao pronome pessoal (quer em posição enclítica, quer proclítica, quer mesmo uma forma que é uma citação latina), e as restantes 61 dizem respeito à contração da preposição com outros elementos. Serão construídas regras de desambiguação contextual para estes casos.

No que diz respeito a *no*, 7 das 340 ocorrências são do enclítico (precedido de hífen) e o resto contrações.

Quanto às formas *nele* e *neles*, nenhuma das 22 ocorrências pertence à categoria nome, pelo que se podem eliminar estas etiquetas.

Apenas um *na* e 5 *nas* dizem respeito às formas enclíticas do pronome (precedidas de hífen), sendo as restantes 329 ocorrências relativas à contração.

As 3 ocorrências de *disto* pertencem à categoria *PREPXPRO*, podendo eliminar--se as etiquetas que o relacionam com o verbo *distar*.

Nenhuma das 68 ocorrências de *deste* ou *destes* pertence ao verbo *dar*, pelo que serão simplesmente eliminadas as etiquetas relativas. O mesmo se fará, pelas mesmas razões, às 29 vezes em que surgem as formas *desse*, *desses*, *dessa* e *dessas*.

Das seis vezes em que *consigo* é utilizado, em nenhuma pertence ao verbo *conseguir*, pelo que estas etiquetas também serão eliminadas.

A única vez que *aonde* é utilizado é como *PREPXADV*, pelo que será eliminada a etiqueta que o identifica como *ADV*.

Relativamente às formas *aonde, consigo, disto, nele* e *nelas*, uma vez que a sua ambiguidade contempla apenas uma etiqueta relacionada com a classe das preposições, é possível construir um grafo semelhante ao elaborado para as formas *trás, sobre, mediante* e *entre*, naturalmente com as mesmas reservas de utilização. Já no caso das formas que, além de poderem pertencer a outras categorias, recebem duas etiquetas diferentes que as identificam como *PREPXDET* e *PREPXPRO*, as coisas não são tão simples: o ideal seria poder criar regras que impedissem a atribuição aos itens de determinada etiqueta, mantendo todas as outras. Como as condições atuais não permitem fazer isso, tentaremos definir regras que permitam especificar a classe de cada palavra, observando os contextos em que ocorre.

Analisados todos os contextos das formas em questão, concluiu-se que sete das nove formas (*dessas, pelo, pelos, pela, deste* e *destes*) ocorrem apenas como *EPXDET*, havendo apenas duas (*desse* e *desses*) com ocorrências quer como *PREPXPRO* quer como *PREPXDET*.

Foi pois construído o grafo de desambiguação seguinte, que **é válido apenas para este corpus**, desambiguando todas as ocorrências das palavras em análise:

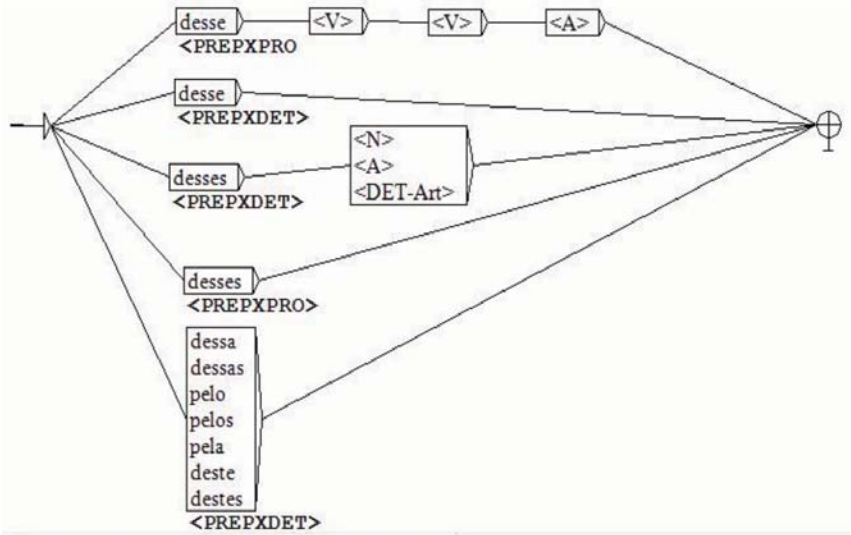

Figura 6 – FST de desambiguação de contrações de preposições com outros elementos - 2

Para a desambiguação da forma *nos* foi construída a seguinte gramática:

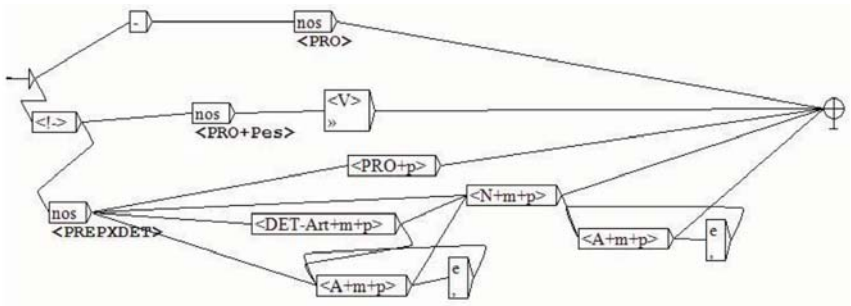

Figura 7 – FST de desambiguação de contrações de preposições com outros elementos – 3

A regra definida no caminho superior é relativa ao clítico, determinando que toda a ocorrência de *nos* imediatamente a seguir a um hífen deve ser etiquetada como *PRO*. Desambigua 31 ocorrências.

nos é etiquetado como pronome pessoal quando, não antecedido por hífen, é seguido por um verbo ou por ». A primeira condição, de carácter geral, desambigua 35 ocorrências no corpus, embora, em três delas, a palavra seguinte seja homógrafa de verbo (*uma, volumes* e *limites*) e, portanto, ela própria ambígua, sendo necessário proceder em primeiro lugar à sua desambiguação. A segunda condição foi inserida especificamente para desambiguar uma ocorrência neste corpus.

A forma é etiquetada como *PREPXDET* quando seguida por um pronome no plural (desambigua 5 ocorrências, relativas à expressão *nos quais*); seguida por um nome masculino plural (48 desambiguações); seguida por um determinante que não seja artigo e por um nome masculino plural (14 desambiguações); seguida por um determinante que não seja artigo, um ou mais adjetivos no masculino plural (separados pela copulativa ou pela vírgula) e por um nome masculino plural (desambigua apenas a expressão "nos seus divinos decretos", mas decidimos manter a possibilidade de reconhecimento de mais adjetivos em posição pré-nominal para aplicações futuras); a gramática prevê também a desambiguação de sequências formadas por *nos*, determinante artigo, nome no plural e um ou mais adjetivos em posição pós-nominal, e ainda a utilização de adjetivos em posição pré e pós-nominal, simultaneamente, ou ainda a ocorrências em que *nos* seja seguido de 1 ou mais adjetivos (optativos), um nome e 1 ou mais adjetivos (optativos), por esta ordem.

Foram adaptadas versões deste grafo para as formas *no, nas* e *na* que, reunidas na gramática *ContrPrepEM.nog*, resolvem um total de 701 ocorrências ambíguas no corpus *Clavis III PT*, das 797 em que surgem as 4 formas, no seu conjunto.

4. Conclusões

Pretendíamos com este texto apresentar uma proposta de desambiguação das preposições portuguesas em a *Clavis Prophetarum*, de António Vieira.

Para a consecução dos nossos propósitos, utilizamos as ferramentas da linguística computacional ou de corpus, programas de análise automática de textos, e métodos estatísticos, concretamente gramáticas e

dicionários específicos que aplicamos ao *NooJ* e outros procedimentos informáticos necessários à nossa pesquisa.

Numa orientação mais voltada para a exploração de um corpus concreto, aplicámos os recursos criados ao livro III da *Clavis Prophetarum* do padre António Vieira, na suas versão portuguesa, tendo (i) avaliado da pertinência e taxa de cobertura dos instrumentos, (ii) efectuado um estudo de carácter estatístico-lexical que tenta ultrapassar as meras contagens de formas e (iii) elaborado e aplicado um conjunto de regras de desambiguação relativas ao Português – algumas com maior poder de generalização do que outras –, centradas na classe da preposição, que permitam análises mais fiáveis desta categoria.

Atualmente, a utilização de programas de análise automática de textos não é, ainda, uma prática corrente no ensino do Português, limitando-se a círculos restritos de investigação. No entanto, esta é uma área cuja crescente importância e potencialidades para o ensino das línguas justificam plenamente todo o esforço de divulgação, de forma a que mais pessoas se interessem por investir e fazer dela uma mais-valia na prática educativa.

Por ser este um estudo essencialmente estatístico, também ao nível das ferramentas de linguística computacional e de corpus, utilizadas, elaboradas, dadas a conhecer e demonstradas, não ficaram aqui esgotadas as suas possibilidades. Muito pelo contrário, pois as possibilidades de pesquisa continuam, agora mais do que antes, e amanhã mais do que hoje, dada a rapidez com que se desenvolvem as ferramentas da linguística computacional e de corpus, imensas e inesgotáveis. A possibilidade de utilização e criação de novos recursos para análise e pesquisa são, em si mesmo, uma porta aberta ao nível da investigação.

Referências

BECHARA, Evanildo. 2002. *Moderna Gramática Portuguesa.* Rio de Janeiro: Editora Y. H. Lucerna Ltda.

BERBER SARDINHA, T. 2004. *Lingüística de Corpus*. São Paulo: Manole.

BICK, Eckhard. 2000. *The parsing system "palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus:Aarhus University Press.

DELTA
36.2
2020

BOWKER, L., PEARSON, P. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

CHOMSKY, N. 1956. Three models for the description of language. *IRE Transactions on Information Theory:* IT-2, 113-124.

COSTA, M. R. V. 2001. *Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas*. Dissertação de Doutoramento. Lisboa: FCSH.

FADANELLI, S. B.; MONZÓN, A. J. 2017. Gêneros textuais datasheet e artigo científico em aulas de ESP: levantamentos léxico-estatísticos para fins educacionais. *Domínios de Lingu@gem*: 11 (2), 351-378.

FIRTH, John. R.1957. Studies in Linguistic Analysis. In: *Philological Society*, chapter 1. London: Blackwell, 1-32.

FIRTH, John. R. 1968. A Synopsis of Linguistic Theory, 1930-55. In: F. R. Palmer (Ed.), *Selected Papers of J. R. Firth* (1952-59). London: Longmans. p. 168-205.

GAZDAR, G.; MELLISH, C. 1989. *Natural Language Processing in Prolog*. Addison-Wesley.

GRIES, S. 2010. Corpus linguistics and theoretical linguistics A love–hate relationship? Not necessarily. *International Journal of Corpus Linguistics*, 15(3), 327–343.

HALLIDAY, M. A. K. 2006. *Computational and Quantitative Studies*. London: Continuum.

HALLIDAY, M., *et al.* 2004. *Lexicology and Corpus Linguistics*. London: Continuum.

HALLIDAY, M.A.K. 1967. The Linguistic Study of Literary Texts. In S. Chatman & S. R. Levin (org.). *Essays on the Language of Literature*: Houghton-Miffin. p. 217- 223.

HUNSTON, S. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

HUTCHINS, W.J. & SOMERS, H.L. 1992. *An Introduction to Machine Translation*: London: Academic Press Limited.

KENNEDY, G. 1998. *An introduction to corpus linguistics*. London: Longman.

LEECH, G. 1992. Corpora and theories of linguistic performance. In: J.SVARTVIK (Ed.). *Directions in Corpus Linguistics*. Proceedings of Nobel Symposium. Berlin: Mouton de Gruyter. p. 105-122.

MCENERY, T.; WILSON, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

MCENERY, T.; XIAO, R. & TONO, Y. 2006. *Corpus-based language studies: an advanced resource book*. London: Routledge.

SILBERZTEIN, M. 2004. NooJ: A Cooperative, Object-Oriented Architecture for NLP. In: *INTEX pour la Linguistique et le traitement automatique des langues*. Cahiers de la MSH Ledoux. Paris: Presses Universitaires de Franche-Comté.

SILBERZTEIN, M. 2006. *NooJ*. Disponível em http://www.nooj-association.org/media/k2/attachments/app/NooJManual.pdf

SILBERZTEIN, M. 2015. *La formalisation des langues : l'approche de NooJ*. Paris: ISTE.

SINCLAIR, J. 2004. *Trust the Text Language*, *corpus and discourse*. London: Routledge.

SMALL, S.; COTTRELL, G. & TANENHAUS, M. (eds). 1988. *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology and Artificial Intelligence*. Palo Alto: Morgan Kaufmann Publishers.

STUBBS, M. 1993. British traditions in text analysis: From Firth to Sinclair. In: M. BAKER, F. FRANCIS & E. TOGNINI-BONELLI (Eds.), *Text and Technology: In Honour of John Sinclair:* John Benjamins. p. 1-46.

TEUBERT, W. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10 (1), 1–13.

TOGNINI-BONELLI, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

VIEIRA, A.. 2000. *Clavis Prophetarum/ Chave dos Profetas*, tradução para o português de Arnaldo do Espírito Santo. Lisboa: Biblioteca Nacional.

Recebido em: 30/04/2019
Aprovado em: 30/03/2020