



Ambiguidade e desambiguação automática das preposições latinas no livro terceiro da *Clavis Prophetarum**

Carlos Assunção (Universidade de Trás-os-Montes e Alto Douro)

 <http://orcid.org/0000-0002-5739-0754>

José Paulo Tavares (Universidade de Trás-os-Montes e Alto Douro)

 <http://orcid.org/0000-0001-5674-2271>

Gonçalo Fernandes (Universidade de Trás-os-Montes e Alto Douro)

 <http://orcid.org/0000-0001-5312-6385>

Résumé: Le troisième livre de *Clavis Prophetarum* [*La clé des prophètes*], écrit par le père António VIEIRA, S. J. (1608-1697), dans sa version latine, provenant de la bibliothèque nationale portugaise en 2000 – édition critique de Arnaldo do ESPÍRITO SANTO – constitue le corpus de cette étude. Pour que l'on puisse travailler un corpus d'une langue donnée, nous avons besoin de ressources linguistiques électroniques formalisées afin d'obtenir la couverture la plus large possible et pouvant être utilisées dans des systèmes appropriés. Si pour le portugais nous disposons déjà de ressources abondantes, depuis les années 1990 par le *LabEL* (Laboratoire d'Ingénierie Linguistique), en ce qui concerne le latin nous ne pouvons pas en dire autant. En effet, l'utilisation de programmes d'analyse automatique de texte n'est pas encore une pratique courante, car elle se limite à des cercles de recherche restreints. Cependant, il s'agit d'un domaine dont l'importance et le potentiel croissants pour la recherche des langues justifient pleinement tout l'effort de diffusion, afin que davantage de personnes soient intéressées à investir et à en faire une valeur ajoutée dans la pratique éducative. Cet article est une petite contribution à ce but et ses principaux objectifs sont d'aider à l'élaboration de règles pour la désambiguïsation automatique des prépositions et de leur syntaxe dans la version latine de *Clavis Prophetarum*, ainsi que l'évaluation de l'efficacité de leur application, afin de permettre des approches ultérieures fiables dans l'étude de cette catégorie dans le corpus à l'aide de techniques automatiques.

Palavras-chaves: P.e António VIEIRA, S. J. (1608-1697), *Clavis Prophetarum*, Linguística, *Corpus*, Ambiguidade, Recursos eletrónicos.

* Este trabalho foi 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.

0. Introdução

A ambiguidade é um dos maiores desafios que se coloca ao desenvolvimento de sistemas de processamento de linguagem natural e, conseqüentemente, à exploração de *corpora*, especialmente quando etiquetados.

GAZDAR/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/TANENHAUS 1988: 4 diferenciam ambiguidade lexical (quando uma palavra pode ter mais do que uma interpretação) de ambiguidade estrutural. HUTCHINS/SOMERS 1992: 85 estabelecem três tipos de ambiguidade lexical: (i) *category ambiguity*, (ii) provocada por homonímia ou polissemia e (iii) *transfer or translational ambiguities*. Numa abordagem orientada por um sistema de análise automática multinível, BICK 2000: 99 classifica os tipos de ambiguidade segundo os níveis morfológico, sintático e semântico, aventando ainda a possibilidade de um nível pragmático.

SILBERZTEIN 2018 [2003-]: 82-83, no entanto, 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 sintáticas 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 objetivo 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* (*Hidden Markov Model*) – trata-se de um tipo de máquina de estados finitos em que todos os símbolos aí representados podem ser gerados em qualquer estado, embora com diferentes probabilidades – para atribuir a cada item a etiqueta mais provável, descartando as restantes possíveis.

Em sistemas de desenvolvimento linguístico como o *NooJ*, parte das ambiguidades resultantes da homografia pode ser resolvida pela hierarquização dos recursos linguísticos, nomeadamente através da atribuição de graus de prioridade aos diversos

1 O *NooJ*, desenvolvido por Max SILBERZTEIN, é um ambiente de desenvolvimento linguístico que inclui grandes dicionários e gramáticas de cobertura e analisa *corpora* em tempo real. Inclui ferramentas para criar e manter recursos lexicais de grande cobertura, além de gramáticas morfológicas e sintáticas. Dicionários e gramáticas são aplicados aos textos para localizar padrões morfológicos, lexicais e sintáticos e marcar palavras simples e compostas. O *NooJ* pode construir concordâncias complexas, com relação a todos os tipos de padrões de estado finito e livres de contexto. Os usuários do *NooJ* podem facilmente desenvolver extratores para identificar unidades semânticas em textos grandes, como nomes de pessoas, locais, datas, expressões técnicas de finanças etc. (SILBERZTEIN 2018 [2003-]).

recursos ou da inserção da chave «+UNAMB» – recurso do *Noof* que permite estabelecer uma transferência padrão eliminando todas as outras possibilidades resultantes da homografia, neste caso –, em determinadas entradas, o que provoca a paragem da análise pelo sistema usando outros recursos disponíveis, evitando assim a atribuição de etiquetas desadequadas à partida.

A necessidade de um dicionário deste tipo 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 sintático-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, mostrando que o aspeto 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). Com efeito, parece um dado adquirido que a etiquetagem lexical correta é um subproduto da análise sintática, o que, em termos de processamento automático de grandes quantidades de texto, é um objetivo ainda distante. Uma resolução parcial (ou redução) das ambiguidades lexicais, não necessitando de uma análise sintática completa e sendo menos ambiciosa, é no entanto mais exequível e realista.

Qualquer que seja o método de desambiguação utilizado, é necessário ter sempre presente que o objetivo 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 a eliminação de análises corretas corresponde à taxa de silêncio. Um sistema de desambiguação ótimo será aquele que mantém ambas no valor zero.

1. Metodologia

Um primeiro óbice a ultrapassar no que diz respeito à exploração de um *corpus* em língua latina, seja ele parte integrante de *corpora* paralelos ou não (ver BOWKER/PEARSON 2002), prende-se com o carácter reducionista e incipiente dos recursos lexicais em formato eletrónico disponíveis para esta língua, o que não permite uma etiquetagem satisfatória nem um posterior tratamento eficiente.

Efetivamente, excetuando um módulo mínimo no projeto *VISL* (Visual Interactive Syntax Learning)², que se limita à apresentação do resultado da etiquetagem e aná-

2 O *VISL* é um projeto de investigação do Institute of Language and Communication, da University of Southern Denmark, cujos alunos e professores, desde 1996, têm vindo a conceber e a implemen-

lise sintática efetuadas a algumas frases-exemplo, merece referência um projeto/instrumento especificamente criado para proceder à etiquetagem automática de textos latinos sob a responsabilidade de Jean SCHUMACHER 2001, na Universidade Católica de Louvain: *Itinera Electronica*³.

Este instrumento, além de permitir transportar os resultados da análise para uma folha de cálculo ou uma base de dados, o que facilita processos como a filtragem da informação ou cálculos estatísticos, revelou-se ainda muito pouco útil devido à incapacidade revelada pelo *Itinera Electronica* em trabalhar textos de maiores dimensões: além de necessitar de muito tempo para a computação, este programa não está preparado para trabalhar textos que excedam os 60.000 caracteres, o que é manifestamente muito reduzido e torna inviável o trabalho com *corpora* de dimensões razoáveis.

Por exemplo: para proceder à etiquetagem automática do livro terceiro da *Clavis Prophetarum* (VIEIRA 2000), constituída por 43.153 palavras (segundo contagem automática parcelar), foi necessário dividir o texto em 12 pequenos ficheiros e submeter cada um deles, alternadamente, no programa. Os resultados desta análise sofreram o processo descrito acima, e finalmente reuniram-se os resultados na mesma folha de cálculo, o que permitiu obter uma visão geral das capacidades de etiquetagem do *Itinera Electronica*, bem como da taxa de cobertura dos recursos lexicais acoplados.

De 43.153 formas, o *Itinera Electronica* deixa 18.827 por classificar, o que equivale a 44% de palavras que não se encontram documentadas no dicionário eletrónico. Relativamente às 24.326 formas classificadas, foram produzidas 51.433 etiquetas, o que significa que houve um acréscimo de 27.107 etiquetas provocado pela ambiguidade – o que se prende certamente mais com as características intrínsecas da própria língua do que propriamente com o instrumento de análise. A partir da consideração dos resultados destes dicionários procedemos à primeira anotação do *corpus*. Depois, para o caso das preposições, foi construída uma segunda anotação para as preposições latinas da *Clavis Prophetarum* (*Clavis III LA*). Relativamente à análise sintática automática, foram aplicadas as regras de desambiguação formalizadas para as preposições e respetivos sintagmas.

Com estas configurações, as formas desconhecidas são 607. Considerando que, destas, 240 correspondem a formas resultantes da falta de homogeneidade gráfica entre *jŷ/iI* e *uU/vV* e 66 abreviaturas (embora cinco destas se contabilizem também nas 240 antes citadas), temos um total de 306 formas claramente não constantes do dicionário, o que resulta numa taxa de cobertura dos recursos linguísticos (dicionário e gramáticas) de 97,5%.

De um total de 37.451 anotações resultantes da primeira anotação (34.474 entradas diferentes) passou-se agora a 36.138 anotações (33.285 entradas diferentes): mesmo

tar ferramentas linguísticas baseadas na Internet para a educação e a investigação (VISL 1996-2020).

3 O Projeto *ITINERA ELECTRONICA* pretende ser uma fonte de ambientes educativos interativos para o ensino e aprendizagem de línguas, literaturas e culturas clássicas, especialmente latinas (SCHUMACHER 2001).

considerando o acréscimo de anotações resultante da aplicação das gramáticas relativas aos nomes próprios e aos numerais romanos, as regras de desambiguação das preposições e sintagmas preposicionais correspondem a uma redução de 1.189 entradas da lista das anotações.

2. Caracterização do léxico do corpus

Para se fazer a caracterização do léxico do corpus, começou-se por definir para o Latim um conjunto de nove etiquetas básicas correspondentes *grosso modo* às tradicionais *partes orationis*:

| Etiqueta | Categoria | Exemplo |
|----------|-------------|---|
| A | Adjetivo | aeneo, aeneus, A +FLX= Aeneus+pos+ab+s+m |
| ADV | Advérbio | altius, alte, ADV +FLX=Alte+comp |
| CONJ | Conjunção | et, CONJ |
| INT | Interjeição | o, INT |
| N | Nome | rosarum, rosa, N +FLX=Rosa+gen+p |
| PREP | Preposição | ad, PREP |
| PRO | Pronome | me, ego, PRO +pes+ac+s+m |
| V | Verbo | amare, amo, V +FLX=Amo+INF+Prés+Act |
| NUM | Numeral | tribus,tres, NUM +card+FLX=Tres+ab+s+n |

Tabela 1: Etiquetas básicas das no *Clavis III LA*

Depois, aplicando o programa *Nooj*, sem ter sido feita qualquer desambiguação, obtiveram-se os seguintes dados relativos à distribuição das etiquetas pelas diferentes *partes orationis*:

| Partes orationis | Clavis III LA |
|-------------------------|----------------------|
| Nomes | 14.436 |
| Verbos | 12.302 |
| Adjetivos | 7.494 |
| Advérbios | 6.933 |
| Conjunções | 4.714 |
| Preposições | 4.349 |
| Pronomes | 4.112 |
| Interjeições | 438 |
| Numerais/Determinantes | 352 |

Tabela 2: As *partes orationis* no *Clavis III* não desambiguado

Os dados da tabela anterior correspondem à seguinte distribuição percentual:

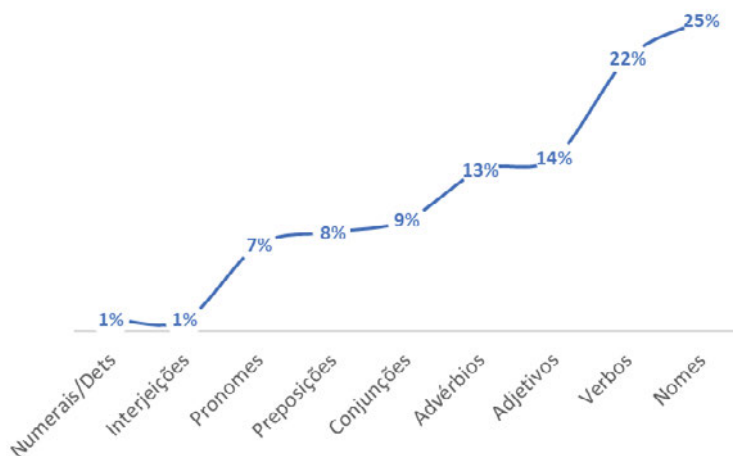


Gráfico 1: Distribuição percentual das *partes orationis* no *Clavis III* não desambiguado

Com o objetivo de avaliar a influência da ambiguidade nos resultados atrás expostos, tomámos uma parte do corpus, o primeiro capítulo, como elemento de controlo e procedemos à desambiguação manual de cada uma das ocorrências, de forma a podermos comparar os resultados da etiquetagem efetuada usando o dicionário que criámos e a gramática, com uma etiquetagem livre de ambiguidades.

Na tabela seguinte podemos ver as variações provocadas pela desambiguação:

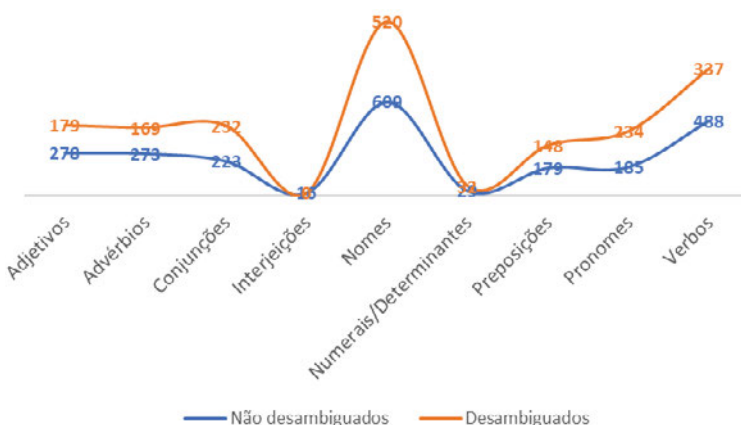


Gráfico 2: As *partes orationis* no *Clavis III* não desambiguado e desambiguado

No sub-corpus, a maior variação antes/após desambiguação verifica-se nos pronomes que, em termos percentuais, passam de 10% para 13%. As conjunções sobem, percentualmente, de 10% para 13%, ao passo que os advérbios e os verbos sofrem uma descida equivalente, de 12% para 9% e de 21% para 18%, respetivamente. A classe das preposições apresenta variação significativa: de 11% no corpus não desambiguado passamos para 17% após desambiguação. De resto, os pronomes descem 3%, os nomes descem 2% e os restantes ou variam 1% ou mantêm a percentagem.

Porém, se se quiser passar além desse objetivo, procedendo ao estudo aturado da utilização de determinada classe de palavras, a desambiguação torna-se imperativa, sob pena de (i) serem incluídos num determinado grupo itens que não lhe pertencem, (ii) os dados serem duplicados, devido às etiquetas que relacionam determinado item com várias classes, e (iii) de se correr o risco de se trabalhar em vão, sobre lemas que, na realidade, não são usados no corpus.

Esta é a razão pela qual se fez a desambiguação das preposições, como se poderia ter feito de uma outra qualquer classe de palavras, aplicando os dicionários e as gramáticas criadas para avaliar a eficácia da sua aplicação, de forma a permitir abordagens fiáveis no estudo desta categoria no corpus usando técnicas automáticas.

3. Desambiguação das preposições latinas

No *corpus* da *Clavis III LA* (VIEIRA 2000) anotado apenas com o dicionário desenvolvido e a gramática morfológica que permite identificar, perante as palavras não constantes no léxico, as formas com um dos quatro clíticos (-ve, -ne, -que e -cum), efetuada uma consulta solicitando todas as ocorrências etiquetadas com *PREP*, obteve-se um total de 4.349 ocorrências, sendo que as formas diferentes são 48: *a, ab, absque, ad, adversum, adversus, ante, apud, circa, citra, clam, contra, coram, cum, de, e, erga, ex, extra, in, infra, inter, intra, ob, per, post, prae, praeter, pro, procul, se, secum, secundum, sed, seque, simul, simulque, sine, sub, subter, subtus, super, supra, tenus, trans, ultra, usque, e versus*.

Destas, não são ambíguas, no sentido em que só têm uma etiqueta, as seguintes treze: *ab, absque, apud, de, e, erga, ex, in, inter, ob, per, sub* e *trans* que não podem ser senão preposições.

Das restantes, *se* recebe etiquetas de *PRO* e *PREP*, mas, considerando que *se* preposição é uma forma arcaica de *sine*, podemos com segurança eliminar deste *corpus* a etiqueta *PREP* da forma *se*. O mesmo se passa com *sed* que, além de conjunção, pode ser uma forma arcaica de *sine*.

Também *a* pode ser uma interjeição ou uma preposição, pelo que terão de ser analisados os contextos para verificar a possível utilização da forma como interjeição. O mesmo se passa com *pro*.

As formas seguintes podem ser advérbios ou preposições, que terão de ser desambiguadas: *ante, circa, clam, contra, infra, post, prae, praeter, procul, simul, subter, subtus, supra, ultra e usque*.

Cum ora é conjunção, ora é preposição, sendo necessária a análise contextual para desambiguação. Quanto às restantes formas da lista, todas podem pertencer a duas ou mais categorias:

- *adversum* pode ser nome, adjetivo, verbo, preposição ou advérbio;
- *adversus* pode ser adjetivo, advérbio, nome, verbo ou preposição;
- *citra* pode ser advérbio, nome ou preposição;
- *coram* pode ser advérbio, nome ou preposição;
- *extra* pode ser advérbio, verbo ou preposição, assim como *intra*;
- *secundum* pode ser advérbio, nome, adjetivo ou preposição;
- *sine* pode ser usado como verbo, nome ou preposição;
- *super* pode ser adjetivo, advérbio ou preposição;
- *tenus* pode ser um nome ou uma preposição;
- *versus* pode ser verbo, advérbio, nome ou preposição.

Na elaboração de regras para desambiguação das preposições procuraremos, dado o elevado número de etiquetas provenientes do facto de termos optado por um alto grau de pormenorização descritiva no dicionário eletrónico (sobretudo devido ao facto de ter sido produzida uma etiqueta diferente para cada caso, género, grau, ...), procuraremos também proceder à desambiguação dos termos ocorrentes no contexto das preposições, nomeadamente no que diz respeito à redução das etiquetas de casos homónimos, usando para isso as restrições de natureza sintática das próprias preposições. Por exemplo, sabendo que *tenus* é uma preposição que se pospõe ao seu complemento, e que este, por seleção de *tenus*, se encontra em ablativo, podemos reduzir as etiquetas de *memoria*, na expressão *centonibus memoria tenus inflatos*, de seis para uma, visto que apenas uma das seis etiquetas contém o traço *+ab*, ao mesmo tempo que se define que, nesta circunstância, *tenus* deve ser etiquetado como *PREP*, se aplicarmos uma regra como a seguinte:



Figura 1: *FST*⁴ de desambiguação do SP introduzido por *tenus*

4 *FST* é um transdutor de estados finitos (*finite-state transducer, FST*).

Tendo sido analisadas as ocorrências das diferentes formas ambíguas no *corpus Clavis III LA* e respetivos contextos, foram construídas várias gramáticas de desambiguação, algumas das quais, de carácter mais geral, poderão ser válidas para outros *corpora*, enquanto outras são específicas para o *corpus* em estudo.

Tenus ocorre apenas uma vez no *corpus*, precisamente na expressão citada acima, pelo que foi formalizada a regra já descrita.

Versus ocorre também apenas uma vez, como advérbio, tendo sido formalizada uma regra para manter apenas esta etiqueta (*versus*/*<ADV>*).

Super é usado 38 vezes, sempre como preposição, tendo sido elaboradas as seguintes regras: *super* é preposição quando: seguido de nome no acusativo (que, por sua vez, mantém apenas a etiqueta relativa a este caso) e, eventualmente, de um adjetivo no mesmo caso: *super lapidem*, *super sedem sanctam*; seguido de adjetivo e nome no acusativo (mantendo os primeiros apenas as etiquetas relativas a este caso): *super omnem impietatem*, *super omnes vicinos*; seguido de pronome no acusativo (que, por sua vez, mantém apenas a etiqueta relativa a este caso): *super eos*, *super utrumque*; seguido de nome no ablativo (que, por sua vez, mantém apenas a etiqueta relativa a este caso): *super exercitio*, *super salute*; seguido de pronome no ablativo (que, por sua vez, mantém apenas a etiqueta relativa a este caso): *super qua*.

Sine ocorre 57 vezes, sempre como preposição, e as regras elaboradas para a sua desambiguação foram as seguintes: *sine* é preposição quando: seguido de nome no ablativo (que, por sua vez, mantém apenas a etiqueta relativa a este caso) e, eventualmente, de um adjetivo no mesmo caso: *sine labore*, *sine Deo vero*; seguido de pronome no ablativo (que, por sua vez, mantém apenas a etiqueta relativa a este caso) e, eventualmente, de um nome no mesmo caso: *sine ulla lege*, *sine alio teste*; seguido de forma verbal no ablativo (que, por sua vez, mantém apenas a etiqueta relativa a este caso): *sine praedicante*; seguido de adjetivo e nome no ablativo (que, por sua vez, mantém apenas as etiquetas relativas a este caso): *sine certo magistratu*, *sine magna causa*.

Considerando que *se* e *sed*, enquanto formas arcaicas de *sine*, não ocorrem no *corpus* senão como pronome e conjunção, definiram-se regras que eliminassem as etiquetas *<PREP>* a elas associadas:

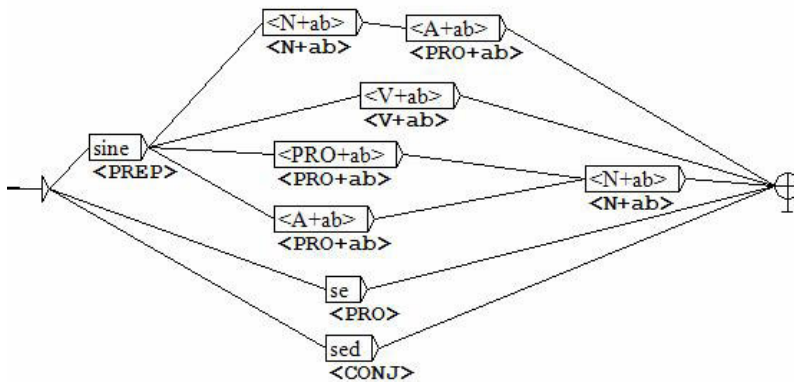


Figura 2: FST de desambiguação de *sine*, *se* e *sed*

Analisando os contextos das 17 vezes em que *secundum* ocorre, não foi possível generalizar regras de desambiguação, pelo que foram construídas as seguintes, especificamente para este *corpus*: *secundum* é um advérbio se seguido de uma vírgula: *Secundum, eas minime deperditas esse*; *secundum* é preposição quando seguido de *Apostolum, carnem, fidem, oraculum, Philosophum, sententiam* ou *extremam*: *secundum sententiam Domini*; seguido de um pronome no acusativo, que deve manter apenas esta etiqueta, e possivelmente de um nome no acusativo: *secundum quid, secundum suam misericordiam*; seguido de *Rhetoricae* (genitivo) e um nome no acusativo: *secundum Rhetoricae leges*.

Embora *intra*, nas onze ocorrências, seja sempre preposição, optámos por elaborar uma gramática para resolver as ambiguidades relacionadas com o sintagma preposicional que introduz, representada no grafo seguinte, que prevê a ocorrência de complementos em acusativo imediatamente à direita ou com um complemento em genitivo (*intra suscepti instituti cancellos*) ou outro preposicional (*intra illud a Christo saeculum*) de permeio:

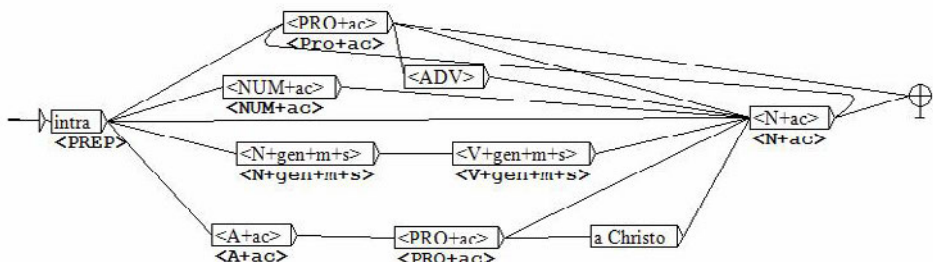


Figura 3: FST de desambiguação do SP introduzido por *intra*

O mesmo acontece em relação a *extra*, nas suas nove ocorrências:

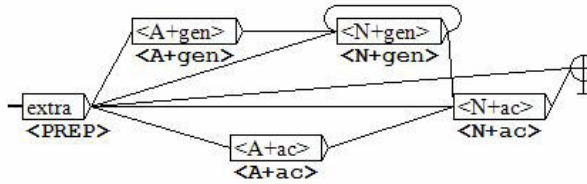


Figura 4: FST de desambiguação do SP introduzido por *extra*

Este grafo determina que *extra* é sempre preposição, prevenindo a desambiguação dos elementos do sintagma preposicional que introduz: um nome no acusativo, que pode ser antecedido por um adjetivo no mesmo caso – *extra mundum*; *extra omnem hyperbolem* – e por um complemento em genitivo – *extra Romanae ditionis limites*; *extra anni solisque vias*.

Coram não ocorre como nome no *corpus*, pelo que se definiram as seguintes regras para desambiguar a forma, entre advérbio e preposição: *coram* é preposição quando antecede um nome ou pronome no ablativo: *coram gentibus*, *coram te*. Nos outros casos, *coram* é advérbio: *et coram res uti sunt intuendo*. Eis o grafo correspondente:

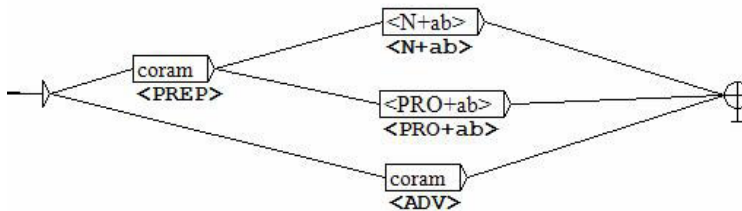


Figura 5: FST de desambiguação de *coram*

Citra ocorre sete vezes no *corpus*, sempre como preposição, e o grafo abaixo (relativo à desambiguação do sintagma preposicional) estabelece que a *citra* preposição podem seguir-se, etiquetados desse modo: um nome no acusativo: *citra piaculum*; um adjetivo e um nome no acusativo: *citra omnem hyperbolem*; um determinativo encaixado, em genitivo, e um nome no acusativo: *citra Apostoli mentem*; um adjetivo no acusativo, um genitivo encaixado e um nome no acusativo: *citra omne temeritatis offendiculum*; uma sequência de adjetivos no acusativo (unidos por vírgula, conjunção, ou ambas) e um nome no acusativo: *citra ingens, et inauditum miraculum*:

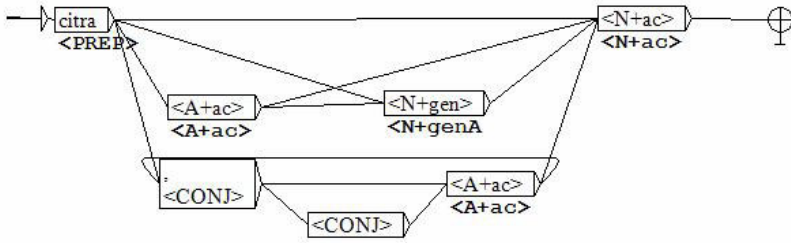


Figura 6: FST de desambiguação de *citra*

Adversum só ocorre uma vez, como advérbio. *Adversus* é preposição nas suas catorze ocorrências no *corpus*, e o grafo de desambiguação dos sintagmas respetivos é o seguinte:

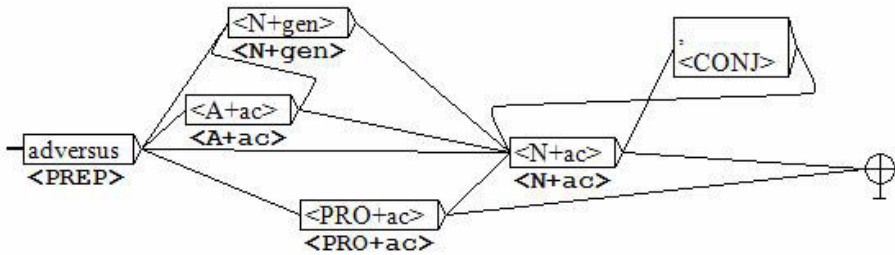


Figura 7: FST de desambiguação de *adversus*

Para *cum*, preposição ou conjunção subordinativa, o grafo seguinte permite desambiguar automaticamente 264 das 269 ocorrências:

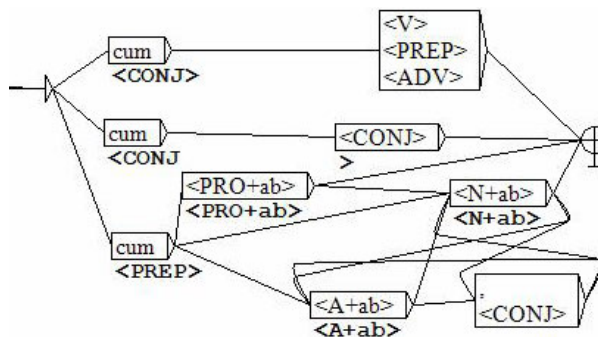


Figura 8: FST de desambiguação de *cum*

Este grafo define que: *cum* é conjunção quando seguido por outra conjunção. Neste caso, as duas palavras são consideradas uma única unidade, etiquetada como *CONJ*:

Para *clam*, *infra*, *procul* e *simul*, que no *corpus* só ocorrem como advérbios, foram construídas regras como a seguinte, que prevê a atribuição da etiqueta <ADV> a *clam*:

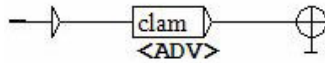


Figura 11: FST de desambiguação de *clam*

Contra é advérbio se antes estiver uma preposição (*e contra*) ou um verbo (*arguebant contra*), ou se depois estiver um verbo (*contra opponebantur*). Nos restantes casos, é preposição, tendo sido formuladas também regras de desambiguação do(s) sintagma(s) que introduz, no grafo seguinte:

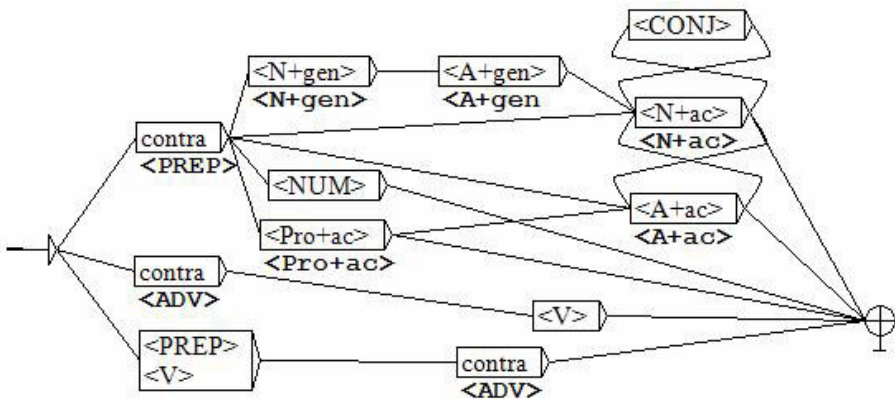


Figura 12: FST de desambiguação de *contra*

Post é preposição, exceto quando integrado na locução *paulo post*, que recebe a etiqueta <ADV>:

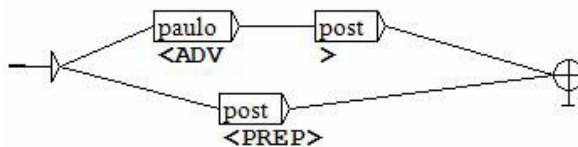


Figura 13: FST de desambiguação de *post*

Supra tanto é advérbio como preposição. Como advérbio, ocorre antes de verbo, de preposição, de nome que não esteja no acusativo ou de «, e *idem*». Nos restantes casos é preposição, tendo sido estabelecidas regras de desambiguação do sintagma que introduz:

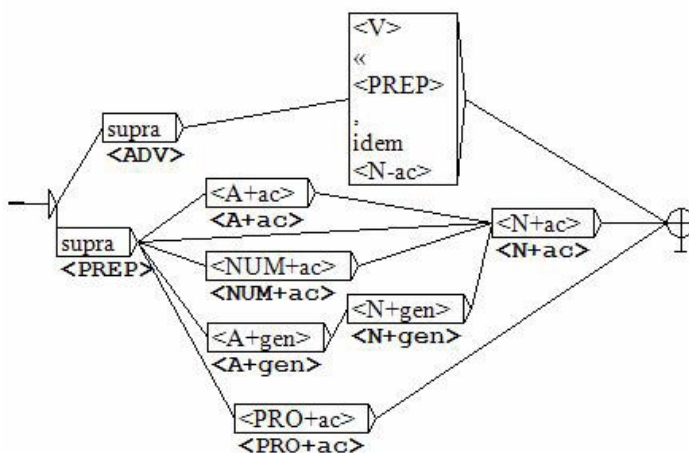


Figura 14: FST de desambiguação de *supra*

Usque tanto pode ser integrado na locução preposicional *usque ad* como ser preposição simples (no *corpus*, depois de *Ravennam*) ou ser advérbio:

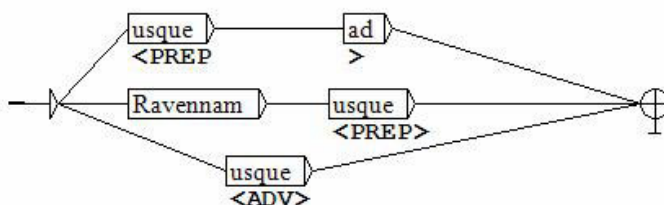


Figura 15: FST de desambiguação de *usque*

4. Conclusões

Atualmente, a utilização de programas de análise automática de textos não é ainda uma prática corrente, limitando-se a círculos restritos de investigação. Neste trabalho, elaboramos regras de desambiguação automática das preposições latinas para a versão latina da *Clavis Prophetarum* e procedemos à avaliação da eficácia da sua aplicação.

Para tal, e 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 sua versão latina, tendo efetuado um estudo de carácter estatístico-lexical para tentar ultrapassar as meras contagens de formas no sentido de elaborar e aplicar um conjunto de regras de desambiguação relativas ao Latim – algumas com

maior poder de generalização do que outras –, centradas na classe da preposição, que permitam análises mais fiáveis desta categoria.

O resultado foi muito interessante: relativamente às preposições, verifica-se um total de 2.990 ocorrências, num total de 34 preposições diferentes. Foram eliminadas todas as etiquetas que classificavam como preposições as formas *adversum*, *clam*, *coram*, *infra*, *procul*, *se*, *secum*, *secundum*, *sed*, *seque*, *simul*, *simulque*, *ultra* e *versus*. Das restantes que, além de poderem ser preposições, podem pertencer a outras categorias, foram completamente desambiguadas, através da aplicação de regras, a saber: *a*, *adversus*, *pro*, *circa*, *post*, *prae*, *praeter*, *subter*, *subtus*, *citra*, *extra*, *sine*, *super*, *tenuis* e *versus*.

Estas aplicações poderão ser utilizadas noutras obras latinas do P.e António VIEIRA e em qualquer texto latino, mas não resolvem a totalidade das regras de desambiguação dos textos latinos. Para que isso aconteça tornam-se necessários:

- (i) a criação de um sistema de etiquetagem linguística para o Latim,
- (ii) a formalização da morfologia flexional da mesma língua, a par da organização de um dicionário de formas canónicas cuja interação resulte num
- (iii) dicionário eletrónico de ampla cobertura,
- (iv) o desenvolvimento de alguns recursos, na forma de gramáticas locais que permitam a identificação e etiquetagem automática de formas complexas ou não registadas no dicionário,
- (v) incluir no dicionário unidades multipalavra
- (vi) e estabelecer relações de derivação entre as palavras, reorganizando o dicionário de formas canónicas e estabelecendo regras derivacionais que permitam a redução do número de entradas do dicionário e a etiquetagem automática de formas derivadas.

Este trabalho iniciou um trabalho de fornecimento das regras de desambiguação da classe de palavras a preposição e deve ser entendido como um singelo contributo nesse âmbito. O desenvolvimento de outro tipo de dicionários, contemplando, por exemplo, relações semânticas ou dicionários bilingues, é outra tarefa que se pode levar a cabo utilizando o *Nooj*.

Bibliografia

- BICK, E. 2000: *The Parsing System «palavras»*. Automatic grammatical analysis of Portuguese in a constraint grammar framework, Ph.D. Thesis, Aarhus University
- BOWKER, L./PEARSON, P. 2002: *Working with Specialized Language*. A practical guide to using corpora, London
- FIRTH, J. R. 1968: «A synopsis of linguistic theory, 1930-55», in: F. R. PALMER (ed.), *Selected Papers of J. R. Firth (1952-59)*, London: 168-205
- GAZDAR, G./MELLISH, C. 1989: *Natural Language Processing in Prolog*. An introduction to computational linguistics, Boston

- HALLIDAY, M. A. K./TEUBERT, W./YALLOP, C./CERMÁKOVÁ, A. 2004: *Lexicology and Corpus Linguistics*, London
- HUTCHINS, W. J./SOMERS, H. L. 1992: *An Introduction to Machine Translation*, London
- LabEL = Laboratório de Engenharia da Linguagem. (s. d.). Disponível em <http://label.ist.utl.pt/en/presentation.php>
- SCHUMACHER, J. 2001: «Les approches statistiques du Projet ITINERA ELECTRONICA: présentation et résultats». *Folia Electronica Classica*, 1. Disponível em <http://bcs.fltr.ucl.ac.be/FE/01/Stat.html>
- SILBERTZEIN, M. 2018 [2003-]: *NooJ Manual*. Disponível em <http://www.nooj-association.org/media/k2/attachments/app/NooJManual.pdf>
- SMALL, S./COTTRELL, G./TANENHAUS, M. (ed.) 1988: *Lexical Ambiguity Resolution*. Perspectives from psycholinguistics, neuropsychology and artificial intelligence, Palo Alto
- VIEIRA, A. 2000: *Clavis Prophetarum/Chave dos Profetas*, edição crítica de A. DO ESPÍRITO SANTO, Livro III, Lisboa
- VISL = *Visual Interactive Syntax Learning*. 1996-2020. Disponível em <https://visl.sdu.dk/>

Ambiguity and automatic disambiguation of the Latin prepositions in the third book of the *Clavis Prophetarum*

Abstract: The third book of *Clavis Prophetarum* [*The Prophets' Key*] by Father António VIEIRA, S. J. (1608-1697), in its Latin version, reedited by the Portuguese National Library in 2000 – a critical edition by Arnaldo ESPÍRITO SANTO – constitutes the *corpus* of this work. In order to be able to work on a *corpus* of a given language, we need formalized electronic linguistic resources in order to obtain the greatest possible coverage and that can be used in appropriate systems. If, for the Portuguese language, we already have trustworthy resources developed since the 1990s by *LabEL* (Laboratory of Linguistic Engineering), as to the Latin language, we cannot mention the same. In fact, the use of automatic text analysis software is not yet a common practice, it is limited to restricted research circles. However, this is an area whose growing importance and potential for language teaching fully justify all the effort to disseminate it, so that more people are interested in investing and making it an added value in research. This article is a small contribution to this aim and has as main goals to help elaborate rules of automatic disambiguation of prepositions and their syntagms, in the Latin version of *Clavis Prophetarum*. It is also intended to evaluate the effectiveness of its application, in order to allow later reliable approaches in the study of this category in the corpus, using automatic techniques.

Keywords: Father António VIEIRA, S. J. (1608-1697), *Clavis Prophetarum*, Linguistics, *Corpus*, Ambiguity, Electronic resources.