UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE ESTUDOS DA LINGUAGEM

DANIEL ALVES DA SILVA LOPES DINIZ

ASPECTOS SEMÂNTICOS E COMPUTACIONAIS DE RESOLUÇÃO DE PRONOMES *E-TYPE* EM UM CORPUS HISTÓRICO DO PORTUGUÊS

CAMPINAS

2016

DANIEL ALVES DA SILVA LOPES DINIZ

ASPECTOS SEMÂNTICOS E COMPUTACIONAIS DE RESOLUÇÃO DE PRONOMES *E-TYPE* EM UM CORPUS HISTÓRICO DO PORTUGUÊS

Monografia apresentada ao Instituto de Estudos da Linguagem da Universidade Estadual de Campinas como requisito parcial para a obtenção do título de Bacharel em Linguística.

Orientadora: Prof[a]. Dr[a]. Ruth E. V. Lopes

CAMPINAS
2016

Para todas as pessoas que enfrentam demônios do meio-dia, com votos de que, separados, de algum modo estejamos juntos.
E para meu pai, por ser a minha família.

Agradecimentos

À professora Ruth, pelos conselhos, pelo rigor na orientação, pelo encorajamento e sobretudo pela enorme paciência com meu ritmo de trabalho (ou sua inexistência) e meus inúmeros atrasos.

Boa parte desta pesquisa lida com programação em Python, algo com que tive meu primeiro contato ao cursar uma disciplina de tópicos de Ciência da Computação para Linguística ministrada pelo professor Tomasz Kowaltowski a convite da professora Charlotte Galves. Depois cursei uma eletiva de Linguística Computacional com o professor Pablo Faria na qual pude aprimorar meus conhecimentos sobre Python e, principalmente, sobre a biblioteca NLTK. Assim, agradeço aos três docentes pela iniciativa, pioneira no IEL, de oferecer aos alunos contato com Linguística Computacional.

O bacharelado a que corresponde essa monografia, especialmente sua segunda metade, foi permeado por muitos problemas relativos à minha saúde mental, intensificados pela rudeza e ignorância de algumas pessoas que conheci nesse tempo de curso. De fato, muitas vezes duvidei que conseguiria terminar esse trabalho ou me formar — ou me formar em 4 anos. Mas pude, de alguma forma, manter esses problemas razoavelmente controlados. Isso só foi possível graças ao meu pai e aos meus amigos, claro, mas também graças ao trabalho da Tânia e da Fabiana. Lhes devo muito. Sou grato ainda a vários médicos e enfermeiros que tantas vezes me atenderam no Hospital de Clínicas da Unicamp. Agradeço à minha orientadora e à professora Norma Wucherpfennig por terem sido compreensivas com essas dificuldades, mas sem perderem o rigor em suas avaliações.

Agradeço ao Lucas, por sua amizade que me acolheu mesmo nos momentos em que eu não estava amigável, que me entendeu quase sempre e não desistiu de mim quando não me entendeu.

À Giovanna, por estar sempre disposta a conversar comigo, mesmo quando a conversa era apenas um relato meu sobre quanto estava preocupado com esta monografia.

À Evelyn, por ser minha amiga há tanto tempo e mesmo agora me vendo tão pouco e estando tão longe! Sua disciplina e inteligência me motivaram a terminar esta monografia e me incentivam a continuar.

À Elisa, pelas conversas sobre assuntos esquisitos e pela grande consideração que tem por todas as pessoas.

À Gabriela e ao João, por terem me convidado para escalar e me encorajado a conti-

nuar praticando. Na época, não imaginei que persistir nisso seria tão recompensador e benéfico! Obrigado aos membros do GEEU por terem me recebido com entusiasmo.

Finalmente, agradeço ao meu pai, que passou por muitos sofrimentos por minha causa desde que saí de Curitiba, mas nunca mediu esforços para me ajudar a superá-los. Em tudo que faço há uma tentativa de retribuir seu amor.

Canto XXXIX

O mistério das cousas, onde está ele?
Onde está ele que não aparece
Pelo menos a mostrar-nos que é mistério?
Que sabe o rio disso e que sabe a árvore?
E eu, que não sou mais do que eles, que sei disso?
Sempre que olho para as cousas e penso no que
[os homens pensam delas,
Rio como um regato que soa fresco numa pedra.

Porque o único sentido oculto das cousas
É elas não terem sentido oculto nenhum,
É mais estranho do que todas as estranhezas
E do que os sonhos de todos os poetas
E os pensamentos de todos os filósofos,
Que as cousas sejam realmente o que parecem ser
E não haja nada que compreender.

Sim, eis o que os meus sentidos aprenderam sozinhos:
As cousas não têm significação: têm existência.
As cousas são o único sentido oculto das cousas.

O Guardador de Rebanhos, Alberto Caeiro
(heterônimo de Fernando Pessoa).

Resumo

Um pronome anafórico (isto é, em uma leitura não-dêitica) parece poder, à primeira vista, ter como antecedente ou uma expressão referencial ou um sintagma nominal quantificado. Neste último caso, é uma variável tanto sintática quanto semanticamente ligada, de modo a ter como contraparte na proposição expressa pela sentença em que ocorre uma variável lógica ligada a um quantificador (existencial ou universal). Evans (1980) observa, no entanto, que, em certos contextos, pronomes que não podem ser ligados admitem, ainda assim, leituras notoriamente anafóricas. Tais pronomes foram denominados "E-type". Esse fenômeno suscitou problemas sintático-semânticos que receberam dois tratamentos na literatura: i) na abordagem composicional da semântica formal, considerou-se que pronomes *E-type* podem ser substituídos por descrições correferentes; e ii) na Discourse Representation Theory (DRT), conforme Kamp (2002), os indivíduos referenciados e as proposições expressas em uma sequência de sentenças são adicionados ou removidos de conjuntos conforme o discurso é processado, o que permite um tratamento uniforme dos vários tipos de pronome. Esta pesquisa objetivou i) apresentar alguns dos problemas supra-citados; ii) comparar o tratamento a eles dispensado por essas duas semânticas; e iii) criar um algoritmo de resolução de anáforas que encontrasse o antecedente do pronome de terceira pessoa plural em contextos E-type ora pressupondo uma das semânticas analisadas em ii, ora pressupondo a outra, de modo a verificar se adotar uma ou outra tem algum efeito sobre a eficácia do algoritmo. Para este último objetivo, 30.126 sentenças de 19 textos do Corpus Histórico do Português Tycho Brahe foram analisadas com um programa de elaboração própria escrito em Python 3 que importa várias ferramentas da biblioteca Natural Language Toolkit (NLTK). Desse total, 234 sentenças tinham pronomes anafóricos do tipo procurado. Conseguiu-se fazer com que o programa detectasse relações de dominância, irmandade e c-comando entre sintagmas e selecionasse, de uma lista de possíveis antecedentes para um dado pronome "eles", aquele com a maior probabilidade de realmente sê-lo. Não se pôde, no entanto, implementar as semânticas composicional e DRT de modo relevante para os procedimentos de decisão executados pelo programa. Considerando que esta pesquisa situa-se em uma interface pouco explorada, especialmente em um âmbito que não o da Ciência da Computação, entre Sintaxe Gerativa, Semântica Formal e Linguística Computacional, mesmo estes resultados bastante preliminares e aquém do inicialmente planejado parecem promissores. Espera-se que pesquisas subsequentes consigam implementar computacionalmente os modelos teóricos comentados aqui.

Palavras-chave: linguística computacional, resolução de anáfora, pronomes E-type.

Abstract

An anaphoric pronoun (that is, in a non-deictic reading) seems, at first sight, to be able to have as its antecedent either a referential expression or a quantified nominal phrase. In the latter case it will be a variable both syntactically and semantically bound, so that it will have as its counterpart in the proposition expressed by the sentence in which it occurs a variable logically bound to a quantifier (existential or universal). Evans (1980) observes, however, that in certain contexts, pronouns which cannot be bound nevertheless admit notoriously anaphoric readings. Such pronouns were named "E-type". This phenomenon has aroused syntactic-semantic problems which received two treatments in the literature: i) in the compositional approach of formal semantics, it has been thought that E-type pronouns can be replaced by coreferential definite descriptions; and ii) in the Discourse Representation Theory (DRT), according to Kamp (2002), referred individuals and expressed propositions in a sequence of sentences are added or removed from sets as the discourse is processed, which allows for a uniform treatment of the many types of pronouns. This research aimed i) to introduce some of the problems aforementioned; ii) to compare the treatment given to them by those two semantics; and iii) to create an algorithm of anaphora resolution that could find the antecedent of a third person plural pronoun in E-type contexts either assuming one of the semantics analyzed in ii, or the other, in such a way as to verify whether adopting either has any effect upon the algorithm's efficiency. In order to accomplish this latter goal, 30,126 sentences from 19 texts of the Historical Corpus of Portuguese Tycho Brahe were analyzed by a self-made program written in Python that imports several tools from the Natural Language Toolkit (NLTK) library. From this total, 234 sentences had anaphoric pronouns such as those that were been looked for. It was managed to have the algorithm detecting dominance, sisterhood and c-command relations between phrases and to select, among a list of possible antecedents for a given "they" pronoun, that with the greatest probability of being it. It could not be implemented in a relevant way, however, the compositional and DRT semantics for the decision procedures executed by the program. Considering that this research is situated in a poorly explored interface, especially in a medium other than Computer Science, between Generative Syntax, Formal Semantics and Computational Linguistics, even such preliminary results that fall short from what was originally expected appear to be promising. It is expected that further research may computationally implement the theoretical models addressed here.

Keywords: computational linguistics, anaphora resolution, E-type pronouns.

Lista de ilustrações

Lista de abreviaturas e siglas

CHPTB	Corpus Histórico do Português Tycho Brahe
Compl	complemento
DRS	Discourse Representation Structure
DRT	Discourse Representation Theory
fbf	fórmula bem formada
GT	Gramática Transformacional
GU	Gramática Universal
LF	Logical Form
NLTK	Natural Language Toolkit
PF	Phonetic Form
PLN	Processamento de Linguagem Natural
POS	part of speech
QR	Quantifier Raising
Spec	especificador
TRL	Teoria da Regência e Ligação
vbl	variável

SUMÁRIO

1 Introdução

Um pronome parece, à primeira vista, poder ser dêitico ou anafórico. Neste último caso, operará semanticamente ou correferindo com seu antecedente ou sendo uma variável por ele ligada. Por volta dos anos 1960, no entanto, notou-se que certos pronomes parecem escapar a essa classificação por terem uma forte leitura anafórica, mas não admitirem ligação devido a restrições tanto sintáticas quanto lógico-semânticas. Um possível nome para designá-los é "pronome *E-type*". Esse fenômeno suscitou, por um lado, uma revisão ou ampliação de semânticas formais composicionais em trabalhos como Evans (1980) e, por outro, deu fôlego para o desenvolvimento de semânticas formais ditas *dinâmicas* no começo dos anos 1980. À proposta de interpretação dos pronomes E-type avançada em Evans (1980) se dará o nome de "análise E-type"[1], e a semântica dinâmica comentada aqui será a Discourse Representation Theory (doravante DRT) conforme descrita por Kamp (2002).

Dado esse fundo teórico, o objetivo que norteou esta pesquisa foi o de implementar um algoritmo de resolução de anáfora que resolvesse os antecedentes do pronome "eles" em um corpus do português ora pressupondo a análise *E-type*, ora pressupondo a DRT. Idealmente, esperava-se responder as seguintes perguntas:

- Esses dois modelos semânticos são passíveis de terem suas formalizações e previsões para fenômenos linguísticos implementadas computacionalmente? De que maneira e por quais motivos?
- A maior parte dos programas dedicados a resolução de anáfora[2] emprega métodos de força bruta, isto é, heurísticas de orientação estatística ou que procuram treinar o programa previamente dando a ele um "gabarito". Como tentar implementar modelos de semântica formal difere dessas técnicas? Quais as implicações dessa opção?
- Adotar um dos modelos impacta o desempenho do programa? Como? Por quê?

Como se verá no capítulo 8, não se pôde implementar esses modelos de modo a compará-los e poder responder essas perguntas: conseguiu-se transformar as sentenças anotadas sintatica-

[1] Esse aspecto terminológico será discutido propriamente na seção 5.2, mas já é importante adiantar que uma certa confusão aqui é esperada: a expressão "E-type" pode designar tanto um certo fenômeno sintático-semântico quanto a interpretação a ele dada em trabalhos como Evans (1980). Até que essa ambiguidade seja comentada mais rigorosamente, convencione-se que "pronome(s) *E-type*" refere-se ao problema, e "análise *E-type*", à teoria a ele dedicada.

[2] O termo "resolução de anáfora" é usado aqui num sentido estrito, como nome de uma área do campo de Processamento de Linguagem Natural (PLN), que é, por sua vez, uma área da Ciência da Computação e da Linguística Computacional.

mente do corpus adotado em objetos de Python que representam árvores sintáticas e que o programa detectasse corretamente dominância e c-comando entre constituintes dessas "árvores".

A parte I desta pesquisa consiste em uma revisão da literatura que enfatiza os conceitos e dispositivos formais mais cruciais para a descrição do problema dos pronomes *E-type*.

O primeiro capítulo descreve brevemente a sintaxe e a semântica[3] da lógica de predicados clássica, em uma tentativa de fazê-lo com rigor formal e, ao mesmo, simplicidade, sem entrar em detalhes que não concernam esta pesquisa. Isso é importante porque é preciso entender como a quantificação é entendida na Lógica para que se entenda alguns detalhes do funcionamento de pronomes *E-type*.

No capítulo 3 definem-se alguns conceitos fundamentais como "composicionalidade", "sentido" e "referência" (conforme Frege) e apresenta-se, muito breve e resumidamente, o modelo semântico introduzido em Heim e Kratzer (1998). Nomeadamente, define-se a linguagem empregada por tal modelo e algumas das mais elementares de suas regras de interpretação, como a função denotação ($[\![\ ]\!]$).

O capítulo 4 é dedicado à teoria sintática adotada aqui, a Teoria da Regência e Ligação, parte do programa gerativista. Atenção especial é dedicada às previsões dessa teoria para o funcionamento da ligação de sintagmas nominais e aos motivos pelos quais ela postula que deve haver um nível de interpretação da Gramática Universal dedicado ao processamento semântico: a Forma Lógica. Essa discussão leva, como se verá, ao movimento de alçamento de quantificadores, que é crucial para que descreva corretamente o problema de pronomes *E-type* abordado aqui.

O problema tematizado aqui é apresentado propriamente no capítulo 5. Começa-se com as definições formal e informal de "pronome"; prossegue-se a uma caracterização dos pronomes ligados e então ao problema relativo à anaforicidade em certas sentenças, o problema dos pronomes *E-type*. São comparadas, então, duas modelagens semânticas para o problema: a análise *E-type* e a Discourse Representation Theory.

Na parte II é desenvolvida a "parte prática" desta pesquisa, isto é, a tentativa de implementar em Python um algoritmo de resolução de anáfora dedicado ao pronome "eles".

O capítulo 6 descreve as ferramentas empregadas e explica as motivações para o uso de cada uma delas.

No capítulo 7 as intuições que orientaram a escrita do programa são apresentadas, de modo que se explique em linhas gerais quais estratégias de resolução de anáfora foram implementadas e por quais motivos.

[3] Deve-se entender "sintaxe" e "semântica" aqui no sentido que os termos têm na Lógica, e não na Linguística.

O capítulo 8 expõe alguns resultados preliminares que puderam ser obtidos.

Por fim, algumas conclusões, ou, o que soa mais preciso, considerações finais são feitas no capítulo 9, e discute-se quais desenvolvimentos futuros podem ser antecipados.

Parte I

Fundamentação teórica

2 Aspectos lógicos

A linguagem[1] de uma lógica de predicados de primeira ordem clássica[2] consiste de:

1. Símbolos lógicos:

   conectivos $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$ (negação, conjunção, disjunção, implicação material e equivalência, respectivamente);

   quantificadores $\forall, \exists$ (os quantificadores universal e existencial, respectivamente);

   variáveis individuais os elementos de um conjunto enumerável $\{x_1, x_2, x_3, \ldots, x_n\}$;

   símbolos de pontuação "(" e ")".

2. Assinatura, uma tripla $\Sigma = \langle \mathscr{E}, \mathscr{P}, \mathscr{F} \rangle$ em que:

   $\mathscr{E}$ conjunto de constantes individuais;

   $\mathscr{P}$ conjunto de símbolos de predicados, com as respectivas aridades[3];

   $\mathscr{F}$ conjunto de símbolos de funções, com as respectivas aridades.

Os símbolos lógicos são comuns a qualquer linguagem, ao passo que a assinatura (que contém os símbolos não lógicos) é única para cada uma (propriedade que a denominação "assinatura" tenta evidenciar), dado que cada linguagem formaliza objetos e relações entre objetos diferentes.

---

[1] Note-se que, em Lógica, "linguagem" é um conjunto de símbolos, que, devidamente concatenados e interpretados, devem formalizar corretamente argumentos, de modo a exprimir significados (isto é, condições de verdade) sem ambiguidade. Em linguística (ou melhor, na vertente da linguística na qual este trabalho se insere), "linguagem" é não um construto artificial, mas uma faculdade mental resultante da evolução da espécie humana cuja estrutura se supõe dedutível a partir de suas manifestações empiricamente observáveis. Espera-se que o contexto de uso da palavra seja o bastante para que os sentidos distintos que ela tem para a Lógica e para a linguística não se confundam.

[2] Por vezes se usam os sinônimos "cálculo de predicados (de primeira ordem)" e "lógica de primeira ordem". "Primeira ordem" distingue essa lógica daquelas de segunda ordem, nas quais predicados podem ser quantificados como termos: "in first-order predicate calculus only 'individual' variables 'x', 'y'... etc., may be bound by quantifiers; in second-order calculi 'F', 'G'... etc. may also be bound, as in '(x)(F)Fx' " (HAACK, 1978, p. 40). Observe-se que "(x)" equivale, na notação empregada por Haack, a "$\forall(x)$". É importante ressaltar que a linguagem descrita aqui é a clássica porque a linguagem de uma lógica de predicados não clássica, como a lógica modal, por exemplo, pode conter símbolos adicionais.

[3] Aridade é um número natural que representa a quantidade de parâmetros (ou "argumentos", uma expressão mais afeita à computação) requerida por um predicado ou uma função para que um valor de verdade, no primeiro caso, ou um objeto do contradomínio, no segundo, seja referenciado.

Um "termo" é um símbolo da linguagem cuja interpretação é um objeto e pode, portanto, ser tomado como parâmetro por um predicado ou função. Assim, as constantes e as variáveis de indivíduo são termos, e, se $t_1, \ldots, t_n$ são termos e $f \in \mathscr{F}$ têm aridade n, então $f(t_1, \ldots, t_n)$ é um termo[4] (SILVESTRE, 2011, p. 224).

Os termos combinam-se uns aos outros segundo regras sintáticas de boa formação de modo a formarem concatenações de símbolos denominadas "fórmulas bem formadas" (por vezes abreviadas *wff*, do inglês "well-formed formula", ou fbf, uma connvenção que será adotada a partir deste ponto). Assim, as fbf de uma linguagem de uma lógica de predicados são definidas da seguinte maneira:

- Se $t_1, \ldots, t_n$ são termos e $P \in \mathscr{P}$ é um símbolo de predicado de aridade n, então $P(t_1, \ldots, t_n)$ é uma fbf;
- Se α é um termo, então $(\neg\alpha)$ é uma fbf;
- Se α, β são termos, então $(\alpha \wedge \beta)$ é uma fbf;
- Se α, β são termos, então $(\alpha \vee \beta)$ é uma fbf;
- Se α, β são termos, então $(\alpha \rightarrow \beta)$ é uma fbf;
- Se α, β são termos, então $(\alpha \leftrightarrow \beta)$ é uma fbf;
- Se x é uma variável de indivíduo e α é um termo, então $(\forall x\alpha)$ é uma fbf;
- Se x é uma variável de indivíduo e α é um termo, então $(\exists x\alpha)$ é uma fbf.

  Adaptado de Silvestre (2011, p. 224–225).

Falta, no entanto, uma interpretação para a linguagem, isto é, um modo de determinar quais de suas fbf serão verdadeiras e quais falsas. Seja L uma linguagem gerada por uma assinatura Σ. Uma "interpretação para L", "modelo para L" ou "L-estrutura" é um par $\mathfrak{A} = \langle D, I \rangle$ no qual D é um conjunto não vazio (o domínio da interpretação) e I é uma função de interpretação dos símbolos de Σ, tal que:

[4] Embora tanto os predicados quanto as funções tenham aridades, uma função com o devido número de parâmetros é um termo, mas um predicado com seus parâmetros não é um termo. Isso porque satisfazer a aridade de uma função significa justamente fazer referência ao objeto do contradomínio da função correspondente aos parâmetros fornecidos, que pertencem ao domínio. Já relacionar termos por meio de um predicado constitui uma proposição concernente a eles, mas não referencia um objeto.

I(c) é um elemento de D, para cada constante c de $\mathscr{E}$;

I(P) é um subconjunto de D^m para cada símbolo P de predicado de aridade m;

I(f) é uma função de D^m em D para cada f símbolo de função de $\mathscr{F}$ de aridade $m \geqslant 1$.

Isto é, I informa qual objeto de D corresponde a cada nome de $\mathscr{E}$, quais objetos de D são relacionados por cada predicado de $\mathscr{P}$ e qual objeto de D é retornado por cada função de $\mathscr{F}$. Diz-se que L tem sua assinatura extendida por $\mathfrak{A}$.

As regras de boa formação das linguagens de predicados exigem que as variáveis individuais ocorram sempre no escopo de um quantificador, o que faz com que "$\forall x_1(Px_1, x_2)$", por exemplo, não seja uma fórmula bem formada por x_2 ocorrer livremente. Por esse motivo, as variáveis individuais são também chamadas "variáveis ligadas" e frequentemente contrastadas com as constantes de indivíduo, nesse contexto chamadas "variáveis livres" — a possibilidade de ocorrerem livremente decorre justamente de terem um referente fixo.

3 Aspectos semânticos

A semântica como entendida neste trabalho parte de alguns pressupostos que serão apresentados brevemente a seguir. Primeiramente, supõe que "[...] language, including our very own home language, is in principle freely reinterpretable like a calculus, at least for the purposes of a semanticist" (HINTIKKA, 1979, p. 717). Ou seja, uma língua é, em si, semanticamente indeterminada; cabe a um *modelo teórico* especificar sua relação com o significado. Analogamente, como já comentado no capítulo 2, uma linguagem lógica L precisa ser associada a uma assinatura $\mathfrak{A}$.

Supõe também que o significado de uma sentença são suas condições de verdade, isto é, o estado de coisas no qual se sabe que ela é verdadeira — ainda que seu real valor de verdade possa ser desconhecido (BORGES NETO, 2002, p. 17). Um modelo semântico deve, portanto, associar corretamente sentenças da língua-objeto a condições de verdade expressas na metalíngua, isto é, formular esquemas como " 'S' sse S", os chamados *esquemas T*[1]. Sendo as condições de verdade de uma sentença bastante autoevidentes, o objetivo, esse sim não trivial, de uma teoria semântica é explicar como os falantes entendem (e produzem) sentenças que até então lhes eram inéditas (HEIM; KRATZER, 1998, p. 1–2).

Considerando-se que este modelo semântico se baseia nas ideias de Frege sobre o significado, é importante apresentá-las aqui, em especial a célebre distinção entre sentido e referência. Um primeiro conceito importante é o de *composicionalidade*.

Frege conjecturou que as condições de verdade de uma sentença são determinados pelos sentidos de suas partes e o modo com que estes se combinam, uma propriedade denominada "composicionalidade":

> *It is held by many grammarians that the relation of syntax and semantics is characterized by the Fregean principle of compositionality. The principle can be stated in various ways; let us now adopt the following formulation: the literal meaning of an expression is uniquely determined by the literal meanings of its subexpressions and their mode of composition (SZABOLCSI, 1981, p. 141).*

Assumida a composicionalidade, resta especificar quais os significados das partes de uma sentença (ou seja, de sintagmas) e como eles se combinam.

[1] A ideia de que esquemas da forma " 'S' sse S" expressam as condições nas quais uma sentença é verdadeira se deve principalmente a Alfred Tarski, que a propôs em trabalhos como *The Concept of Truth in Formalized Languages* (o original em polonês foi publicado em 1933, e a ele se seguiram traduções para o alemão, em 1935, e para o inglês, em 1983).

Em trabalhos como *Begriffsschrift* ("Conceitografia"), Frege supusera que nomes próprios[2] são rótulos para objetos, o que exige afirmar que optar por "Terra" ou "o único planeta habitado por seres humanos" para denotar o terceiro planeta mais próximo do Sol não tem qualquer efeito lógico-semântico. Entretanto, em seu célebre *Über Sinn und Bedeutung* ("Sobre sentido e referência"), ele observa que igualdades como "$a = b$" são mais informativas que aquelas como "$a = a$", o que não deveria acontecer caso nomes fossem meros rótulos, ou seja, caso o significado de uma expressão consistisse apenas de seu referente. Note-se que a segunda igualdade é autoevidente por ser conhecida de maneira *a priori* (e *analítica*). Assim, ele concluiu que expressões significam não simplesmente referenciando objetos, mas *denotando um referente* e *expressando um sentido*. Embora o significado inclua um sentido, ou seja, um objeto necessariamente distinto do referente, o modelo semântico adotado aqui é *extensional*. Isso significa que a denotação de cada nó terminal em uma estrutura sintática é seu referente, e não seu sentido ou seu significado (isto é, seu referente somado a seu sentido)[3].

Algumas expressões, entretanto, parecem não ter significado completo em si mesmas (e, por consequência, não ter um referente fixo). Esses significados *insaturados*, isto é, incompletos, precisam ser complementados por significados *saturados*, completos. São, portanto, funções (no sentido matemático do termo) cujos argumentos são significados saturados (HEIM; KRATZER, 1998, p. 3). À hipótese de que a composicionalidade consiste da saturação de significados se dá o nome de "aplicação funcional" (HEIM; KRATZER, 1998, p. 13).

Como esta semântica, por ser extensional, está interessada na correspondência entre as expressões linguísticas e os objetos do mundo, é preciso especificar as referências de expressões linguísticas, uma vez que elas serão os objetos modelo-teóricos desta semântica: "se o procedimento da semântica de modelos consiste na associação das expressões a 'objetos do mundo', nenhum modelo de interpretação de uma linguagem pode prescindir de 'objetos' " (BORGES NETO, 2002, p. 21). O referente de uma sentença é um valor de verdade, e indivíduos (objetos do mundo) são os referentes dos sintagmas nominais (ou "nomes próprios", na terminologia empregada por Frege). É importante ainda saber que o sentido de uma sentença é uma proposição, passível de formalização em sistemas lógicos como os descritos no capítulo 2. As demais expressões têm significados insaturados, isto é, são funções que tomam valores de verdade, indivíduos ou outras funções como argumentos. Haack (1978, p. 61) esquematiza os referentes e sentidos que expressões têm segundo

[2] "Nome próprio", para Frege, é qualquer nome ou sintagma com propriedades referenciais: "by 'proper name' Frege understands *both* ordinary names and definite descriptions (he says that a name is any expression that refers to a definite object, though in fact he also envisages the possibility of names, like 'Odysseus', that don't denote a real object)" (HAACK, 1978, p. 61–62). Esse uso é o mesmo que o da expressão "termo singular" na filosofia da linguagem contemporânea.

[3] Muitos fenômenos linguísticos exigem uma semântica na qual, em certos contextos, a denotação de um sintagma seja seu sentido. Tais contextos são denominados *intensionais*. Um exemplo simples é "Ana acredita que João fuma". Para os propósitos deste trabalho, no entanto, um sistema extensional é suficiente.

Frege na tabela 1.

expression	*sense*	*reference*
proper name	meaning of the name	object
predicate	meaning of the predicate expression	concept
sentence	proposition	truth-value

Tabela 1 – Sentido e referência segundo Frege.

Fonte: Haack (1978, p. 61).

Quais entidades linguísticas referenciam objetos diretamente, isto é, quais são as contrapartes na linguagem natural das constantes de indivíduo lógicas, era e ainda é motivo de debate. Delinearam-se, historicamente, duas linhas de pensamento sobre esse problema.

Antes de comentá-las deve-se ressaltar, no entanto, que em filosofia da linguagem há conceitos definidos e nomeados de maneira bastante divergente do que é usual na linguística, de modo que são necessários alguns esclarecimentos. Em particular, é preciso definir "descrição", em particular a "descrição definida", e "nome".

Uma *descrição* é um sintagma no qual um determinante dos classificados pela gramática tradicional como "artigo" precede um sintagma nominal. Exemplos são "o rei da Inglaterra", "uma criança" e "a pessoa mais velha que já viveu". Ela é *definida* caso o artigo em questão seja definido (ainda na nomenclatura da gramática tradicional): "by a 'description' I mean any phrase of the form 'a so-and-so' or 'the so-and-so.' A phrase of the form 'a so-and-so' I shall call an 'ambiguous' description; a phrase of the form 'the so-and-so' (in the singular) I shall call a 'definite' description" (RUSSELL, 1912, p. 82).

Para autores como Kripke e Meinong, nomes operam na linguagem natural exatamente como as constantes de indivíduo operam na Lógica: um nome denota um indivíduo diretamente, isto é, não por ele ter determinadas propriedades, mas *per se*. Isto é, um nome não descreve seu referente. Assim, nomes são *designadores rígidos*, isto é, denotam o mesmo indivíduo em todos os mundos possíveis, independentemente de quão diferentes sejam as propriedades de seus respectivos objetos em cada um deles (HAACK, 1978, p. 58–59).

Já para Russell, nomes próprios "camuflam" descrições e são, assim, contextualmente elimináveis. As descrições são, semanticamente, proposições lógicas quantificadas. Desse modo, "o atual rei da França" não referencia diretamente um objeto, mas sim expressa a proposição $\exists x\Big(\text{REI}(x) \wedge \forall y\big(\text{REI}(y) \rightarrow y = x\big)\Big)$[4]. É justamente por nenhum objeto corresponder a essa ca-

[4] A paráfrase em português dessa proposição lógica é: "existe um x tal que x é atual rei da França e tal que, para todo y, se y é atual rei da França, y é x". O primeiro termo da conjunção afirma que há *pelo menos* um atual rei da França;

racterização que a sentença "o atual rei da França é careca" é falsa[5], como observa Russell (1905, p. 483–484) ao comentar a hipótese fregeana de que nomes expressam um sentido e denotam um referente:

> *One of the first difficulties that confront us, when we adopt the view that denoting phrases express a meaning and denote a denotation, concerns the cases in which the denotation appears to be absent. If we say "the King of England is bald," that is, it would seem, not a statement about the complex meaning "the King of England," but about the actual man denoted by the meaning. But now consider "the King of France is bald". By parity of form, this also ought to be about the denotation of the phrase "the King of France". But this phrase, though it has a meaning provided "the King of England" has a meaning, has certainly no denotation, at least in any obvious sense. Hence one would suppose that "the King of France is bald" ought to be nonsense; but it is not nonsense, since it is plainly false.*

A literatura de filosofia (analítica) da linguagem sobre o *Frege's puzzle*, o processo de referência em sentenças contendo existenciais negativos (qual o referente de "Pégaso" em "Pégaso não existe"?) e problemas afins é bastante extensa, de modo que não se entrará em detalhes sobre o assunto aqui — essa seção teve como objetivo meramente reconhecer a existências desses problemas.

Seja como for, há apenas dois significados saturados: valores de verdade e indivíduos (isto é, objetos do mundo). Os *tipos* básicos do modelo semântico serão, assim, indivíduos (tipos e) e valores de verdade (tipos t). Os outros tipos, significados insaturados[6], são definidos recursivamente[7] a partir deles:

Definição 1 (Linguagem-tipo (BORGES NETO, 2002, p. 22))**.** Seja D o conjunto de todas as denotações possíveis para as expressões linguísticas. Os tipos de seus elementos são definidos pelas seguintes regras:

Regra 1 e é um tipo;

Regra 2 t é um tipo;

---

o segundo, que há *exatamente* um. Por simplicidade, em Lógica, quando se quer afirmar a existência de exatamente um objeto, por vezes escreve-se $\exists \iota x(\ldots)$ ou $\exists! x(\ldots)$. Mais rigorosamente, ι (a letra grega *iota*) ou "$\exists!$" representam o quantificador lógico de unicidade, que pode ser definido a partir dos quantificadores existencial e universal: $\iota x\big(P(x)\big) := \exists x\big(P(x) \wedge \forall y(P(y) \rightarrow y = x)\big)$. O símbolo ":=" expressará aqui, como em Haack (1978), que o símbolo à esquerda está sendo definido em termos da fbf à direita — ao passo que "=" denotará uma mera coincidência de valores entre os *relata*.

[5] É importante ressaltar que julgar "falsa" uma sentença como essa não é trivial, mas sim uma posição teórica, qual seja, a visão não pressuposicional adotada em Russell (1905). Não se entrará no mérito aqui de discutir os pormenores dos valores de verdade de proposições correspondentes a sentenças que contêm descrições sem referente.

[6] "As is well known, according to Frege, the ontological furniture of the universe divides into objects and functions" (VAN HEIJENOORT, 1967, p. 325).

[7] A recursividade da definição 1 faz com que haja infinitos tipos; note-se, no entanto, que poucos têm uma contraparte semântica.

Regra 3 se α e β são tipos, $\langle \alpha, \beta \rangle$ é um tipo.

$D_e \subset D$ é o conjunto dos tipos e. $D_t \subset D$ é o conjunto dos tipos t (isto é, D_t é o conjunto $\{0,1\}$).

$[\![\ ]\!]$ é a função que retorna a denotação de uma expressão linguística. Por exemplo, $[\![\text{“Daniel”}]\!] = \text{Daniel}$. Isso significa que a denotação, o referente, de “Daniel” (uma palavra da língua-objeto) é Daniel, um indivíduo. Já $[\![\text{Daniel}]\!]$ é uma expressão semântica sem sentido, uma vez que um indivíduo não denota nada (HEIM; KRATZER, 1998, p. 23).

Um verbo intransitivo como “fuma” é interpretado semanticamente como $[\![\text{“fuma”}]\!] = \lambda x \in D_e : \text{x fuma}$. Ou seja, a denotação de “fuma” é uma função que, para cada indivíduo, retorna 1 caso ele fume e 0 do contrário. O tipo de $[\![\text{“fuma”}]\!]$, é, portanto, $\langle e,t \rangle$.

Especificados os tipos da semântica, é preciso esclarecer quais regras os combinam, uma vez que se adota aqui o princípio da composicionalidade:

Definição 2 (Princípios de composição (HEIM; KRATZER, 1998, p. 43-44))**.**

Nós terminais se α é um nó terminal, então $[\![\alpha]\!] \in D_e$;

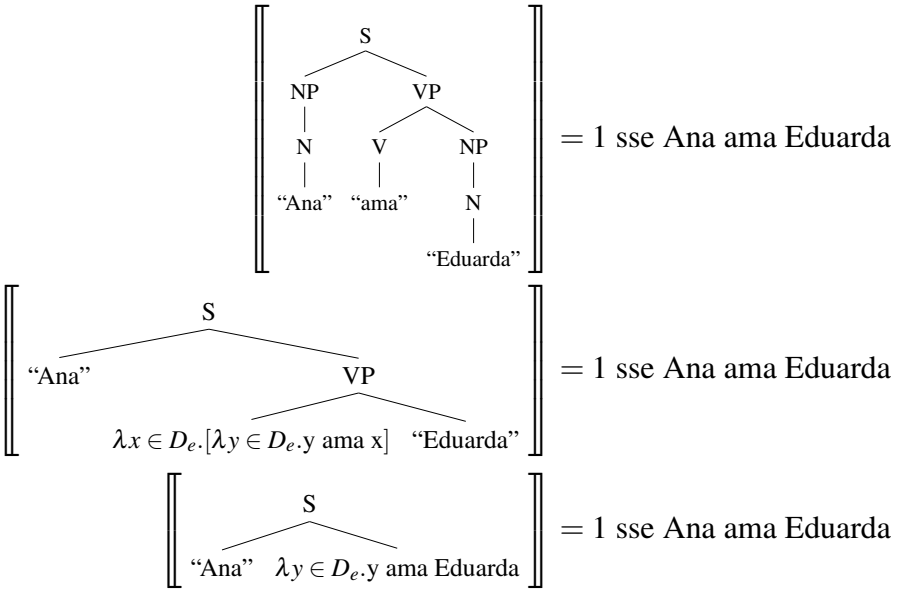
Nós não ramificados se α é um nó não ramificado e β é seu filho, então $[\![\alpha]\!] = [\![\beta]\!]$;

Aplicação funcional se α é um nó não terminal, $\{\beta, \gamma\}$ é o conjunto dos filhos de α e $[\![\beta]\!]$ é uma função em cujo domínio está $[\![\gamma]\!]$, então $[\![\alpha]\!] = [\![\beta]\!]([\![\gamma]\!])$.

Por exemplo, suponhamos um modelo com o seguinte conjunto D de denotações: $D = \left\{ \begin{array}{c} \text{Ana,} \\ \text{Eduarda,} \\ \lambda x \in D_e.[\lambda y \in D_e.\text{x ama y}], \\ 0, \\ 1 \end{array} \right\}$. Nesse caso, a composição semântica da sentença (1) se dá como na figura 1:

(1) Ana ama Eduarda.

Demonstração.



□

Figura 1 – A composição semântica da sentença (1).

Fonte: adaptado de Heim e Kratzer (1998, p. 99–100).

4 Aspectos sintáticos: a Teoria da Regência e Ligação

May (1985, p. 1) comenta que desde o projeto fregeano tornou-se comum pensar que, sendo a sintaxe em última análise a interface entre o som e o significado (isto é, ouvir uma sentença implica estabelecer relações hierárquicas entre seus segmentos e conhecer, assim, suas condições de verdade), a gramática de uma língua deve ser a responsável por mapear estruturas sintáticas nas unidades lógicas descritas pela semântica. Como na Sintaxe gerativa e para May, aqui supor-se-á que entender uma sentença é submetê-la a diferentes "níveis de interpretação", de modo que cada um tome a estrutura do nível anterior como *input* e gere uma estrutura possível mas não necessariamente distinta como *output*. "Gramática" e conceitos afins, como "gramaticalidade", serão entendidos aqui, portanto, como em May (1985, p. 2–3):

> *A* grammar *is understood as a function that specifies for each sentence of a language its formal description at each level of representation. A grammar (strongly)* generates *a class of* structural descriptions*, whose members are sets of representations* $\{a_1, \ldots, a_n\}$*, where each* a_j $(1 \leq j \leq n)$ *is a representation at level* A_j*. A "gramatical" sentence, then, is one that is assigned a structural description each of whose members are well-formed; an ungrammatical sentence is one that is assigned a structural description with at least one ill-formed member.*

Aqui faz-se necessário comentar quais mudanças epistemológicas o advento da Teoria da Regência e Ligação (abreviada TRL a partir daqui) provocou no interior da Sintaxe gerativa, visto que lidar com relações anafóricas é precisamente lidar com regência e ligação. A TRL é tanto sucessora direta da Gramática Transformacional (doravante GT) quanto uma revisão de vários de seus pressupostos e métodos.

Propusera-se na GT que uma sentença assume, entre sua gênese no léxico e sua forma final, duas formas intermediárias: uma estrutura profunda e uma estrutura superficial. Nesse paradigma, os itens lexicais se organizam na estrutura profunda como descrito pelas teorias X' e de papéis-θ. Uma série de movimentos (isto é, transformações) gera uma estrutura superficial tomando a estrutura profunda como *input*. Cada tipo de movimento tem sua ocorrência e operação determinadas por regras próprias que garantem a gramaticalidade da estrutura superficial gerada: o movimento-QU, por exemplo, sempre alça palavras-QU do complemento de VP para o especificador de CP.

A TRL introduz dois níveis de representação para as sentenças posteriores à estrutura superficial e reanalisa as várias transformações regidas pelas mais diferentes regras da GT como uma única transformação, Mova-α, que tem sua aplicação restringida pela própria Gramática Uni-

versal: “GB [Government and Binding] proposes that the grammar itself consists of a series of ‘modules’ that contain constraints and principles which govern the well-formedness of the output [de Mova-α]” (SELLS, 1985, p. 23–24).

As estruturas profunda e superficial da GT têm como contrapartes na TRL a estrutura-d e a estrutura-s, respectivamente. Essa nomenclatura procura evidenciar o fato de que elas tiveram seus papeis na derivação sintática revistos, mas não completamente modificados.

Uma das representações que se segue à estrutura-s, isto é, que a toma como *input*, é a PF (*phonetic form*), que contém uma cadeia (isto é, a estrutura-s linearizada) a ser submetida às regras fonológicas da língua em questão. De modo geral, pouca atenção teórica é dispensada aos detalhes da forma de PF e das propriedades dos movimentos que a geram.

A outra estrutura gerada a partir da estrutura-s é a LF (*logical form*). May (1985, p. 2) a descreve da seguinte maneira: “it represents whatever properties of syntactic form are relevant to semantic interpretation — those aspects of semantic structure that are expressed syntactically”. A figura 2 esquematiza a articulação dos módulos da GU prevista pela TRL.

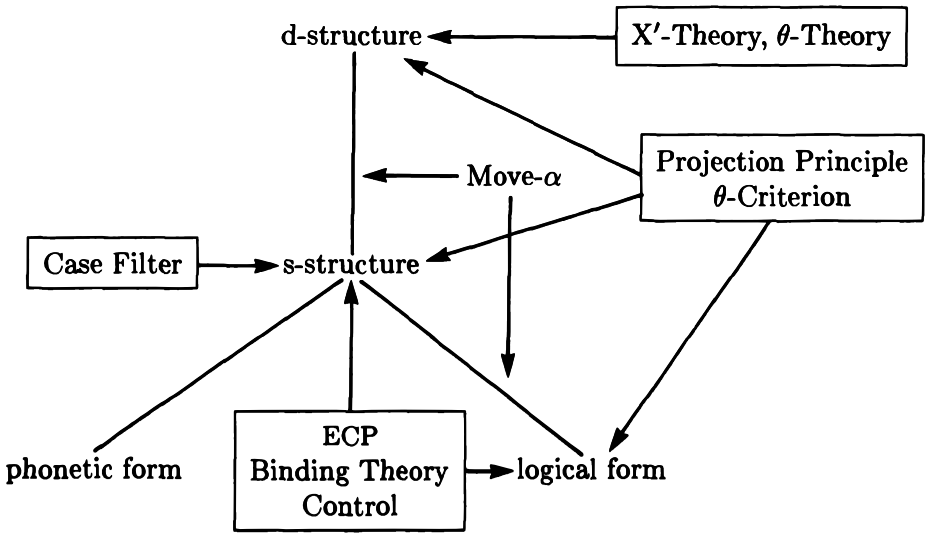


Figura 2 – A estrutura da GU segundo a TRL.

Fonte: Sells (1985, p. 24).

4.1 Ligação de sintagmas nominais

A TRL, como seu nome já sugere, tem como um de seus objetivos descrever e prever como sintagmas nominais[1] se distribuem sintaticamente quanto à correferência. *Ligação* é definida em 3 e 4:

Definição 3 (C-comando (MIOTO; SILVA; LOPES, 2013, p. 210))**.** α *c-comanda* β sse β é o irmão de α ou se β é dominado pelo irmão de α.

Definição 4 (Ligação (MIOTO; SILVA; LOPES, 2013, p. 211))**.** Um nó α *liga* um nó β sse:

i. α c-comanda β; e

ii. α e β estão coindexados.

As restrições sintáticas para a gramaticalidade de sentenças com sintagmas coindexados foram resumidas em três princípios, listados em 6. Para entendê-los, é preciso antes definir *categoria de regência* (definição 5).

Definição 5 (Categoria de regência (MIOTO; SILVA; LOPES, 2013, p. 215))**.** A categoria de regência de α é o XP mínimo que contém α, o regente de α e

i. um sujeito que é distinto de α e que não contém α; ou

ii. a flexão que atribui Caso nominativo para α.

Definição 6 (Princípios de ligação (MIOTO; SILVA; LOPES, 2013, p. 211))**.**

Princípio A Uma anáfora deve estar ligada em sua categoria de regência;

Princípio B Um pronome deve estar livre em sua categoria de regência; e

Princípio C Uma expressão-R deve ser livre.

A figura 3 ilustra que "a mãe d[o Pedro]$_i$ se$_i$ adora" é agramatical porque a anáfora deve estar ligada em sua categoria de regência, mas [$_{DP}$ o Pedro] não pode satisfazer essa condição porque não c-comanda o nó I^0.

[1] A expressão "sintagma nominal" deve ser entendida em um sentido lato, isto é, como abrangendo tanto NPs quanto DPs. Quando essa distinção for necessária, serão empregadas as expressões "NP" e "DP".

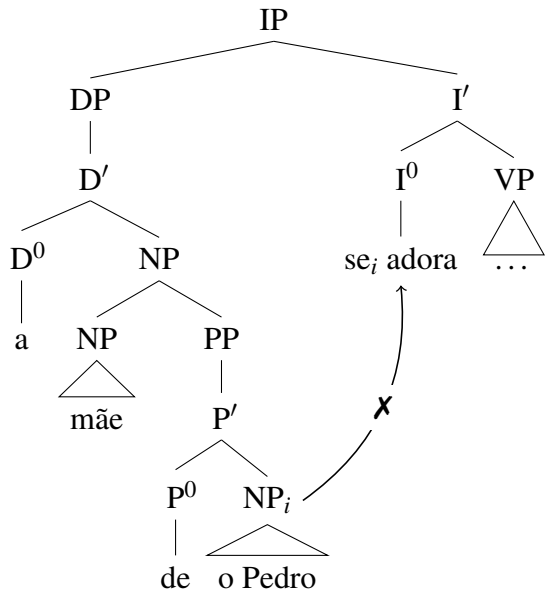


Figura 3 – A importância do c-comando para a ligação.

Fonte: Adaptado de Mioto, Silva e Lopes (2013, p. 210).

4.2 A existência da Forma Lógica

Na Sintaxe gerativa, como em qualquer outra teoria científica, a existência de componentes da GU (e da própria GU, na verdade) deve ser afirmada (ainda que hipoteticamente) somente se houver fatos que o justifiquem. Seguir esse princípio evita explicações *ad hoc* para os fenômenos linguísticos e uma multiplicidade de objetos formais com constituições e funcionamentos arbitrários.

A existência de PF parece pouco controversa: como é inegável que sentenças têm uma dimensão fonético-fonológica linear e a estrutura-s é hierárquica e não prevê qualquer interação dos constituintes com a fonotaxe da língua em questão, deve haver um nível de interpretação gramatical reservado a fenômenos morfossintáticos e fonológicos. Por exemplo, a possibilidade de contrair um sintagma nominal e a cópula em inglês (como em "he is" contraído em "he's") e a concordância verbal em "Lucy sings" (e não "*Lucy sing") "are syntactically conditioned in the sense that their s-structure representation will contain the information necessary to their operation, but that operation itself (e.g., the 'spelling-out') takes place in the PF part of the grammar" (SELLS, 1985, p. 21).

Analogamente, isto é, pelos mesmos motivos e por métodos semelhantes, a existência de LF deve ser deduzida pela empiria. O caso de LF é mais controverso que o de PF, mas um argumento forte a favor de sua existência é o fenômeno de *alçamento de quantificadores* (*quantifier raising* ou QR, na literatura anglófona).

A descrição de QR exige observar que sintagmas nominais quantificados e palavras-QU não são referenciais e não podem, portanto, receber um papel-θ na posição A da estrutura-d em que são gerados. Assim, para que o critério θ não seja violado, esses sintagmas devem se mover para uma posição A' ao se adjungir a um CP, IP ou VP (CYRINO, 1994, p. 13). Tal movimento, sendo uma instância A' de mova-α, deixa na posição de origem do sintagma movido uma variável (*vbl*) com ele coindexada[2]. A figura 4 ilustra a estrutura-s e a LF de uma sentença com um DP quantificado[3] em posição de sujeito.

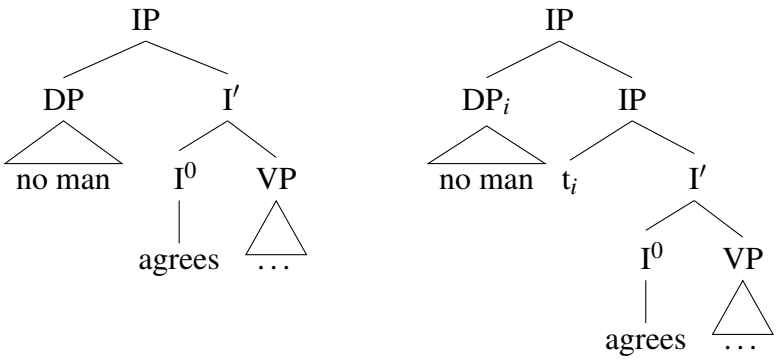


Figura 4 – O alçamento de quantificadores.

Fonte: adaptado de Larson e Segal (1995, p. 228).

Na sentença (2) o sintagma nominal quantificado ocupa a posição de objeto:

(2) Jorge viu todo mundo.

Com efeito, "todo mundo" em (2) não denota, como talvez se pudesse supor, algo como o conjunto das pessoas salientes no discurso no momento em que a sentença é emitida[4]. Esse DP opera semanticamente como um quantificador universal (no sentido lógico do termo) que itera sobre os elementos desse conjunto[5]. Isto é, entender o sentido de (2) exige processar uma quantificação

[2] Na notação usada em Heim e Kratzer (1998), o índice do sintagma movido ocupa uma posição arbórea própria. Optou-se aqui, por simplicidade, por índices anexados ao próprio objeto sintático movido, como em Larson e Segal (1995).

[3] Tais sintagmas quantificados serão considerados aqui, conforme Heim e Kratzer (1998) e Larson e Segal (1995), como DPs. Note-se, no entanto, que em obras como Mioto, Silva e Lopes (2013), sintagmas nominais adjungidos desse modo são rotulados com "QP" (*quantifier phrase*).

[4] De fato, vários pronomes não referenciam objetos. Um exemplo intuitivo é "ninguém", mas observe-se que "um" em "um estudante disse ser inteligente" não referencia nenhum objeto em específico; meramente informa que o conjunto de estudantes não é o conjunto vazio (MÜLLER, 2003; CANN; KEMPSON; GREGOROMICHELAKI, 2009; EVANS, 1980).

[5] Larson e Segal (1995) discutem em mais detalhes o significado de afirmar isso. Parece importante, por exemplo, especificar se essa iteração se dá sobre nomes, indivíduos ou apontamentos (isto é, o gesto pragmático que torna um

lógica. Sendo, semanticamente, um quantificador universal, este DP precisa ligar (no sentido lógico, e não sintático, do termo) uma variável que corresponda, a cada iteração, a um dos elementos do conjunto sobre o qual se está iterando — um quantificador universal que não ligue uma variável impedirá que a proposição em que ocorre seja uma fbf. O objeto sintático (no sentido linguístico do termo "sintaxe") que corresponde à variável logicamente ligada ao quantificador é, justamente, o vestígio que o sintagma nominal quantificado deixou ao ser alçado. Assim, é preciso que uma variável seja c-comandada por seu antecedente quantificado no nível da interpretação semântica (que, como se procura sustentar aqui, é LF) para que sua interpretação como variável logicamente ligada seja possível (LARSON; SEGAL, 1995, p. 249). Mesmo em sentenças como a da figura 4 esse requisito não é satisfeito na estrutura-s, uma vez que nesse nível da interpretação da sentença não há nenhum vestígio ligado ao sintagma nominal quantificado. E, o que é ainda mais grave, como a figura 5 explicita, "todo mundo" não c-comanda nenhum nó na estrutura-s da sentença (2).

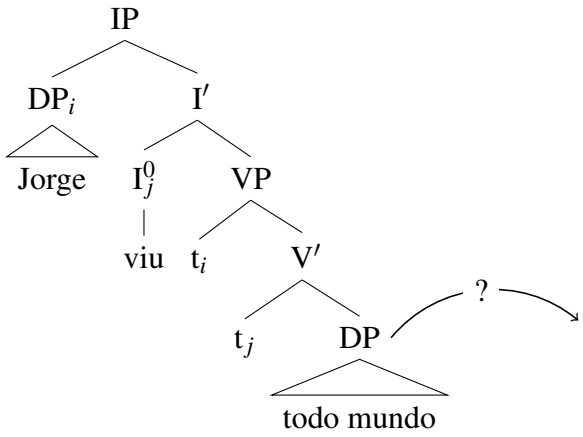


Figura 5 – O c-comando do sintagma quantificado na estrutura-s da sentença (2).

Disso se depreende que, ainda que o critério-θ não requeresse a existência de LF, os sintagmas nominais quantificados muitas vezes se encontram, na estrutura-s, em posições que não lhes dão o escopo necessário para uma correta derivação da proposição expressa pela sentença, isto é, suas condições de verdade. Por esses motivos, postular a existência de LF parece razoável o bastante pelo menos para os propósitos desta monografia.

QR (e, por consequência, a existência de LF) prevê que estruturas-s contendo mais de um sintagma nominal quantificado terão mais de uma contraparte em LF, uma vez que não há uma ordem específica para o alçamento de ambos. Na figura 6, por exemplo, "every man" e "some woman" podem ser alçados de modo que i) o quantificador universal tenha escopo sobre o quantificador existencial; ou ii) o quantificador existencial tenha escopo sobre o quantificador

indivíduo linguisticamente saliente). Esses aspectos da natureza da quantificação linguística não serão discutidos aqui, no entanto, por estarem fora do escopo desta pesquisa.

universal. No primeiro caso, diz-se que o quantificador universal tem "escopo amplo", e o existencial, "escopo estreito"; no caso ii, o quantificador universal tem "escopo estreito", e o universal, "escopo amplo". Essa previsão é correta porque, com efeito, há uma leitura de "every man admires some woman" na qual, para todo homem, é correto afirmar que este homem admira alguma mulher, e outra na qual há pelo menos uma mulher amada por todos os homens. A primeira leitura é expressa por $\forall x\Big(\text{HOMEM}(x) \rightarrow \exists y\big(\text{MULHER}(y) \wedge \text{ADMIRA}(x,y)\big)\Big)$ e a segunda, por $\exists y\Big(\text{MULHER}(y) \wedge \forall x\big(\text{HOMEM}(x) \rightarrow \text{ADMIRA}(x,y)\big)\Big)$. Esse tipo de ambiguidade é análoga a outras ambiguidades estruturais resolvidas em outros níveis da gramática, como "[$_{\text{DP}}$ mad dogs] and Englishmen go out in the noonday sun" versus "[$_{\text{DP}}$ mad dogs and Englishmen] go out in the noonday sun" (LARSON; SEGAL, 1995, p. 230).

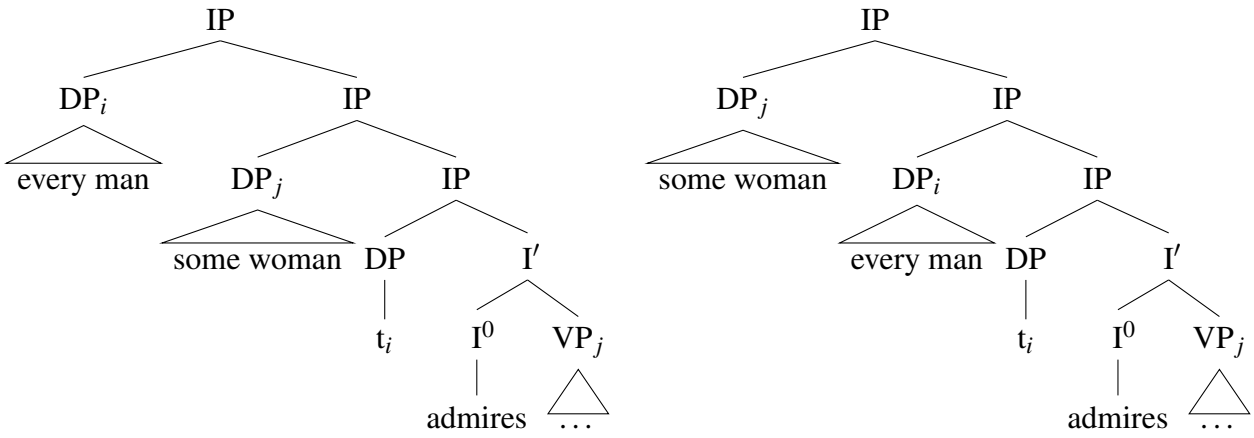


Figura 6 – QR como explicação de ambiguidades relativas a escopo.

Fonte: adaptado de Larson e Segal (1995, p. 229).

Assim, QR é exatamente um alçamento do sintagma nominal quantificado, na passagem da estrutura-s para LF, que lhe permite ter uma variável a ele ligada em seu escopo sintático, como requer sua interpretação lógico-semântica. Nesse caso, "todo mundo" é movido para o Spec de CP[6]. Como em qualquer outra instância A' de Mova-α, uma variável é deixada no nó irmão de V^0, de modo que, após QR, "todo mundo" c-comanda, entre outras coisas, a variável a ele ligada.

Ou seja, assumir que a estrutura-s já é o nível de interpretação semântica implica afirmar que a estrutura representada na figura 5 é o *input* para a interpretação semântica da sentença (2), o que, por sua vez, impossibilita a ligação sintática que permite a ligação lógico-semântica que se sabe ser parte da interpretação da sentença (2). Deve haver, então, um nível gramatical posterior à

[6] Considera-se aqui que toda sentença tem como sintagma máximo um CP, o que é mais comum na literatura de Sintaxe que na de Semântica Formal; nesse último campo, seria mais comum a duplicação do IP máximo.

estrutura-s no qual movimentos encobertos[7], isto é, sem realização fonética, garantam que a sintaxe licencie a interpretação semântica correta.

[7] No português e no inglês, as ocorrências de QR de sintagmas nominais quantificados que ligam pronomes são, de fato, encobertas, mas note-se que há instâncias de QR descobertas. O alçamento de um sintagma para o Spec de CP com o objetivo de focalizá-lo constitui um exemplo de movimento aberto entre a estrutura-s e LF. Além disso, há línguas, como o húngaro, em que QR é consistentemente aberto (LARSON; SEGAL, 1995, p. 230–231), (CYRINO, 1994, p. 13).

5 Pronomes

Por sua própria natureza interdisciplinar, o problema tematizado aqui é discutido tanto pela TRL, parte do programa gerativo-transformacional, quanto pela Semântica Formal. Essas teorias empregam as palavras "pronome" e "anáfora" de modo que certos objetos linguísticos recebem o mesmo nome nas duas abordagens e outros, nomes diferentes. Ou seja, os usos desses dois termos nessas teorias se sobrepõem, mas não coincidem. Faz-se necessário, portanto, convencionar quais objetos linguísticos receberão quais nomes nesta pesquisa. Como em Müller (2003, p. 18), o termo "pronome" será empregado aqui de modo "a abranger tanto o que a Teoria da Ligação (Chomsky 1981) chama de pronome quanto o que essa mesma teoria chama de anáfora. Também quanto ao termo *anáfora*, vamos nos afiliar a essa tradição mais antiga e utilizá-lo para expressar relações de dependência referencial entre sintagmas nominais e não apenas para os pronomes do tipo reflexivo como faz a Teoria da Ligação" (itálico original). Cabe observar, ainda, que a gramática tradicional distingue pronomes "dêiticos", aqueles cujos referentes são determinados por meios extra-linguísticos (por exemplo, se se diz "ninguém gosta dele" apontando-se para um homem visível ao locutor e ao interlocutor), de "anafóricos", aqueles cujos referentes covariam com o referente de uma expressão referencial (um sintagma nominal, por exemplo). Em Sintaxe gerativa e Semântica Formal, no entanto, geralmente se considera que as ocorrências dêiticas e anáforicas são casos particulares de um mesmo fenômeno, a saber, uma classe de palavras que, ao invés de terem referentes fixos, referenciam objetos salientes no contexto discursivo no momento em que os referentes devem ser atribuídos na interpretação da sentença (MÜLLER, 2003, p. 21), (HEIM; KRATZER, 1998, p. 239).

Um *pronome* é informalmente definido como um item lexical que tem a mesma distribuição sintática dos sintagmas nominais, mas cuja denotação não é dada por uma entrada lexical (MÜLLER, 2003, p. 18). Observe-se que pronomes com realização fonética podem ser anafóricos, mas vestígios de movimento necessariamente o são, motivo pelo qual também devem ser abrangidos pelo termo "anáfora". Isto é, vestígios são inerentemente anafóricos, mas os pronomes com realização fonética por vezes compartilham dessa propriedade:

> Os pronomes, sob certos aspectos, comportam-se como vestígios: dão origem a interpretações em que funcionam como variáveis ligadas. Mas, à diferença dos vestígios, que, pela própria maneira como são construídos, têm sempre um elemento que os liga, os pronomes podem ter ou não um antecedente linguístico. Quando não, sua interpretação precisa ser reconstruída a partir de informações presentes no contexto (CHIERCHIA, 2003, p. 401).

A definição formal de pronome será feita segundo Heim e Kratzer (1998). Para tanto,

é preciso revisar a definição 2 de modo que $[\![\ ]\!]$ seja complementada por uma função auxiliar que atribua um referente para cada símbolo terminal que não o tenha fixado no léxico (isto é, como um elemento do conjunto D_e). Esses símbolos serão denominados *variáveis*: "a variable denotes an individual, but only relative to a choice of an assignment of a value" (HEIM; KRATZER, 1998, p. 92). Assim, uma *atribuição (de variável)* é uma função parcial $a := \mathbb{N} \rightarrow D_e$ cujo domínio é expresso por $\mathrm{dom}(a)$. Exemplos de atribuições de variáveis incluem $\left[\begin{smallmatrix}1 \rightarrow \text{Ana}\\ 2 \rightarrow \text{Ricardo}\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}1 \rightarrow \text{Ana}\\ 2 \rightarrow \text{Ana}\end{smallmatrix}\right]$ e $[1 \rightarrow \text{Ana}]$. Definida a função a, os referentes de variáveis são atribuídos conforme as definições 7 e 8.

Definição 7 (Regra de pro-formas e vestígios (HEIM; KRATZER, 1998, p. 292))**.** Se α é uma pro-forma ou vestígio, a é uma atribuição de variável e $i \in \mathrm{dom}(a)$, então $[\![\alpha_i]\!]^a = a(i)$.

Definição 8 (Condição de adequação (HEIM; KRATZER, 1998, p. 292))**.** Um contexto c é *apropriado* para uma LF ϕ somente se c determina uma atribuição de variável g em cujo domínio ocorre como elemento cada índice que tem ocorrência livre em ϕ.

Assim, $[\![\text{"ele"}_2]\!]^{\left[\begin{smallmatrix}1 \rightarrow \text{Ana}\\ 2 \rightarrow \text{Ricardo}\end{smallmatrix}\right]} = \left[\begin{smallmatrix}1 \rightarrow \text{Ana}\\ 2 \rightarrow \text{Ricardo}\end{smallmatrix}\right](2) = \text{Ricardo}$ é uma atribuição de variável que será um contexto apropriado para uma sentença caso não haja outros índices de variáveis nesta sentença a não ser 1 e 2. Um símbolo terminal α é uma variável sse há atribuições a e a' tais que $[\![\alpha]\!]^a \neq [\![\alpha]\!]^{a'}$. *Constantes* são símbolos terminais que não são variáveis[1].

5.1 Pronomes ligados

Dados os objetivos visados e os temas abordados nesta pesquisa, há pouco a se dizer sobre pronomes dêiticos, de modo que se volta agora à anaforicidade pronominal e alguns problemas inerentes a esse fenômeno. Considerem-se as sentenças (3-a) e (3-b):

(3) a. João$_i$ disse que *pro*$_i$ é inteligente.
b. Um aluno$_i$ disse que *pro*$_i$ é inteligente.
Adaptado de Cann, Kempson e Gregoromichelaki (2009, p. 141).

Uma primeira impressão de *pro* nas sentenças (3) poderia ser a de que eles funcionam semanticamente da mesma forma: *pro* na sentença (3-a) toma como referente o referente de João, $[\![\text{"João"}]\!]$, e, na sentença (3-b), toma como referente o referente de "um aluno", $[\![\text{"um aluno"}]\!]$.

Como já exposto na seção 4.2, no entanto, "um aluno" não referencia nada; meramente expressa que o conjunto de alunos não é o conjunto vazio. O significado de (3-b) parece ser o de que

[1] Não confundir esse uso do termo "constante" com o uso que este tem na Lógica, descrito no capítulo 2.

há um aluno que afirmou a própria inteligência, isto é, que de um objeto $x \in D_e$ está sendo afirmado tanto que ele é um aluno, quanto que ele disse que x é inteligente. Deste modo, *pro* em (3-b) deverá ser uma variável ligada em três sentidos do termo "ligada": sintaticamente, na estrutura-s da sentença, semanticamente, em sua LF, e logicamente, na proposição expressa por (3-b). Assim, a proposição expressa por (3-a) é $D\big(j, I(j)\big)$, mas a expressa por (3-b) é $\exists x\Big(A(x) \wedge D\big(x, I(x)\big)\Big)$[2] (CANN; KEMPSON; GREGOROMICHELAKI, 2009, p. 141).

Com efeito, substituir a anáfora por seu antecedente em (3-b) produz uma sentença que tem uma proposição distinta associada a si, e, portanto, condições de verdade também diferentes das originais:

(4) *[$_{\text{DP}}$ Um estudante]$_i$ disse que [$_{\text{DP}}$ um estudante]$_i$ é inteligente.

Esse fenômeno, resultado de uma violação do princípio-C descrito em 6, na página 16, corrobora a previsão de que pronomes antecedidos por uma expressão-R funcionam de um modo e pronomes antecedidos por sintagmas nominais quantificados, de outro. A sentença (5) exemplifica uma sentença com um pronome cujo antecedente é um sintagma nominal quantificado por "todo":

(5) Every male skier loves his mother.
todo esquiador ama dele mãe
"Todo esquiador$_{\text{i}}$ ama a mãe dele$_{\text{i}}$ ".
$\forall x\Big(\big(\text{ESQUIADOR}(x) \wedge \text{HOMEM}(x)\big) \rightarrow \text{AMA}\big(x, \text{MÃE}(x)\big)\Big)$
Adaptado de King e Lewis (2016).

É importante definir o que se entende por "variável ligada", uma vez que esse conceito, na Linguística, tem uma definição distinta daquela que tem na Lógica. Linguisticamente, a ligação é tanto sintática quanto semântica: "sintaticamente temos uma anáfora ligada quando um sintagma nominal está ligado a seu antecedente. Semanticamente, temos uma anáfora ligada quando uma forma pronominal é interpretada como uma variável presa por um operador" (MÜLLER, 2003, p. 30–31).

---

[2] Considere-se a seguinte glosa para essas duas proposições:

j	João
$A(x)$	x é aluno
$D(x,y)$	x disse y
$I(x)$	x é inteligente

5.2 Pronomes *E-type*

"Sentenças *donkey*" são sentenças da seguinte forma: "the typical donkey sentence has an indefinite DP in an 'if' clause or relative clause, and a pronoun anaphoric to it in the main clause" (HEIM; KRATZER, 1998, p. 295). Elas foram introduzidas na literatura por Peter Geach em seu *Reference and Generality*, de 1962, e devem seu nome a geralmente envolverem burrinhos. Ganharam muita atenção de semanticistas, sintaticistas e filósofos da linguagem desde então por evidenciarem alguns problemas particularmente controversos envolvendo anaforicidade. Eis algumas sentenças *donkey*:

(6) a. Every man who owns a donkey feeds it.
b. If a man owns a donkey, he feeds it.

A pergunta relevante aqui é: qual o referente de "it" nas sentenças (6)? O pronome está certamente relacionado semanticamente a "a donkey", mas de que maneira? Como já discutido anteriormente, definições indefinidas como essa não referenciam nenhum objeto. "It" não pode correferir com "a donkey", portanto, simplesmente porque não há referente a ser retomado.

Assim, esse pronome deve funcionar como uma variável ligada (descartando aqui, é claro, a leitura dêitica que sempre é possível e exigiria um gesto ostensório por parte de quem proferisse as sentenças (6)). Então, analogamente, a sentença (7-a) deve expressar ou a proposição (7-b) ou a proposição (7-c):

(7) a. If Pedro owns a donkey he is rich.
b. $\forall x\big(\text{DONKEY}(x) \wedge \text{OWNS}(\text{Pedro}, x) \rightarrow \text{RICH}(\text{Pedro})\big)$
c. $\exists x\big(\text{DONKEY}(x) \wedge \text{OWNS}(\text{Pedro}, x)\big) \rightarrow \text{RICH}(\text{Pedro})$

A proposição (7-b) relaciona a descrição indefinida a um quantificador universal. Explicar como um tipo de sintagma que sempre pareceu expressar existência pode expressar universalidade exigiria algumas previsões bastante contraintuitivas, de modo que parece melhor descartar essa leitura em favor da hipótese de que "a donkey" de fato expressa existência, como em (7-c). Observe-se, no entanto, que em (7-c) o quantificador existencial não tem escopo sobre os termos lógicos que correspondem à segunda oração subordinada de (7-a). Analogamente, o quantificador existencial relacionado a "a donkey" em (6-a) não terá escopo sobre a oração "feeds it". Isso, no entanto, deve ser impossível porque faz com que "it" não seja devidamente ligado, de modo a ficar sem nenhum referente (KAMP, 2002, p. 190–191). Observe-se que este problema de escopo pode ser apresentado também de modo menos lógico-semântico e mais sintático: mesmo após QR, "a

donkey" não c-comanda "it" em (6-a), de modo que não pode ligá-lo (HEIM, 1990, p. 140).

Chega-se então a um paradoxo: "it" certamente tem uma forte relação semântica com "a donkey", como o comprovará a intuição linguística de qualquer falante; não obstante, não pode nem correferir com esse sintagma nem por ele ser ligado. As sentenças (6) evidenciam, portanto, que alguns sintagmas que não podem estar ligados por violarem princípios sintáticos (os elencados na definição 6) considerados válidos por motivos independentes, ainda assim parecem correferir como se estivessem ligados:

> *[nas sentenças (6)] the quantificational expression is said to be too deeply embedded within the relative clause or the conditional antecedent to bind the pronouns he and it. Nevertheless the reference of the pronouns seems to co-vary with the choice of individuals quantified over, suggesting some kind of binding by the quantifier [...] (CANN; KEMPSON; GREGOROMICHELAKI, 2009, p. 143).*

Nesse sentido, Evans (1980) argumenta que esses pronomes em contextos nos quais nem são dêiticos, nem herdam o referente de uma expressão-R, nem são variáveis ligadas, constituem uma categoria própria[3], a de *pronomes E-type*.

Neste ponto faz-se necessário um breve interlúdio terminológico. O problema que consiste na correferência intuitiva de pronomes que, estranhamente, não podem estar ligados por vezes é metonimicamente denominado "(problema dos) pronomes *E-type*"[4]. Mas, num sentido talvez mais preciso do termo, "pronome *E-type*" é a categoria de pronomes postulada por Evans (1980) cuja definição e descrição procura resolver esse problema. Em outras palavras, a expressão "pronome *E-type*" pode designar tanto um fenômeno linguístico que deve ser modelado semanticamente, quanto um desses modelos. Numa tentativa de evitar ambiguidades, a expressão "pronome *E-type*" designará nesta pesquisa o problema linguístico; já a solução para ele proposta em Evans (1980) será denominada "análise *E-type*", à maneira de Heim (1990).

São comparadas, a seguir, a análise *E-type* e a análise da Discourse Representation Theory dos pronomes *E-type*.

---

[3] Não se pretende implicar, com isso, que Evans (1980) é o primeiro registro de pronomes *E-type* na literatura linguística. De fato, o próprio autor reconhece que observações nesse sentido já haviam sido feitas quando da publicação de seu trabalho (EVANS, 1980, p. 344). Esse trabalho é tomado aqui como uma referência tanto do que vem a ser pronomes *E-type* quanto de um certo modo de interpretá-los semanticamente porque, concorda-se aqui com o autor, ele é a primeira tentativa de investigar esse fenômeno sistematicamente. Cabe ainda notar que Evans publicara, em 1977, um artigo sobre o mesmo assunto intitulado "Pronouns, Quantifiers and Relative Clauses" no Canadian Journal of Philosophy, direcionado primariamente a filósofos (da linguagem).

[4] Outros nomes incluem "anáforas não ligadas" (*unbound anaphora*, "anáfora *donkey*" (*donkey anaphora*), e "anáfora problemática" (*problematic anaphora*) (KING; LEWIS, 2016).

5.2.1 Pronomes *E-type* segundo a análise *E-type*

Evans (1980) argumenta que pronomes *E-type* são, na superfície, muito similares aos ligados (o que explica a distinção ter demorado a ser feita), mas há certos contextos linguísticos que salientam suas diferenças: "[...] the distinction between a bound and an *E-type* interpretation of a pronoun comes out most clearly when the pronoun has a plural quantifier as antecedent" (EVANS, 1980, p. 339). Como essa passagem já sugere, a classificação de um pronome é determinada pela relação hierárquica entre ele e seu antecedente, de modo que, a rigor, não são os pronomes que são *E-type*, dêiticos, ligados etc., mas sim suas leituras: "[...] whether or not a pronoun is interpreted as bound by a quantifier phrase depends upon the grammatical relation in which it stands to that quantifier phrase" (EVANS, 1980, p. 340–341). Por questões de simplicidade, no entanto, muitas vezes expressões como "pronome *E-type*" ou "pronome ligado" serão empregadas metonimicamente. As sentenças usadas para introduzir a questão são reproduzidas a seguir:

(8) a. Few congressmen admire only the people they know.
b. Few congressmen admire Kennedy, and they are very junior.
(EVANS, 1980, p. 339).

Na sentença (8-a), a leitura de "they" é a de uma variável ligada ao quantificador relativo a "few": o que está sendo afirmado é que o conjunto de congressistas não é vazio e que, para alguns poucos elementos desse conjunto, é correto afirmar que eles admiram somente quem conhecem. À primeira vista, (8-b) parece admitir uma leitura análoga, mas, segundo Evans (1980), há dois motivos para isso ser impossível.

O primeiro problema está relacionado a escopo. Se o quantificador em (8-b) tiver escopo sobre "they" e ligá-lo, a interpretação da sentença será a de que alguns poucos congressistas são tais que tanto admiram Kennedy quanto são muito jovens. Essa não é, contudo, a interpretação intuitiva dessa sentença. O que realmente se afirma em (8-b) é que poucos congressistas admiram Kennedy e que os indivíduos que são congressistas que admiram Kennedy são muito jovens.

Ainda que o problema de escopo não existisse, outro, que para Evans é ainda mais grave, se imporia. A leitura intuitiva de (8-b) é não apenas a de que os congressistas admiradores de Kennedy são muito jovens, mas de que *todos* os congressistas admiradores de Kennedy são muito jovens. Em outras palavras, "they" em (8-b) referencia todos os elementos de D_e que verificam o predicado da cláusula anterior, isto é, todos os objetos que são congressistas que admiram Kennedy:

> *It looks as though the role of the pronoun in these sentences is that of referring to the object(s), if any, which verify the antecedent quantifier-containing clause. If this is the role of these* E-type *pronouns, we explain why the truth of the clause containing them requires that all the relevant objects satisfy the predicate [...] (EVANS, 1980, p. 340).*

Essa pressuposição de totalidade imposta por pronomes E-type é codificada corretamente pelo quantificador universal, mas não pelo existencial, de modo que pronomes E-type não podem ser variáveis ligadas.

Há alguns métodos para se distinguir pronomes ligados de *E-types*. Como pronomes ligados são, por definição, antecedidos por sintagmas nominais quantificados, perguntar que objetos eles referenciam não tem sentido, mas a mesma questão é perfeitamente razoável para sentenças como (8-b) (no caso dela, em específico, perguntar o que "they" referencia leva à resposta "os poucos congressistas que admiram Kennedy"). De fato, como já comentado, sintagmas nominais quantificados não referenciam nada. Também por esse motivo, tornar negativo o antecedente quantificado de um pronome *E-type* torna agramatical a sentença que o contém, mas o mesmo não acontece ao se negar o antecedente de um pronome ligado:

(9) a. No congressmen admire only the people they know.
b. *No congressmen admire Kennedy, and they are very junior.
c. No sheep are such that John owns them and Harry vaccinates them in the Spring.
d. *John owns no sheep, and Harry vaccinates them in the Spring.
(EVANS, 1980, p. 340).

Se um pronome *E-type* plural (como "they" ou "eles") retoma *todos* os objetos que verificam a sentença que contém seu antecedente, o que um pronome *E-type* singular retoma? Evans (1980, p. 343) argumenta que, no caso de pronomes singulares, a exigência é não de totalidade, mas de unicidade:

> *If it is the role of* E-type *pronouns in general to refer to the object or objects which verify the antecedent clause, and if an* E-type *pronoun is singular, then we would predict that the use of that pronoun will convey the implication that there is just one object verifying the antecedent clause — an implication which is not carried by the use of the existential quantifiers themselves.*

Se um pronome E-type referencia o(s) objeto(s) que verifica(m) a sentença que contém seu antecedente e fortemente pressupõe totalidade ou unicidade desse(s) objeto(s), o problema de sua referência é resolvido ao se postular que ele funciona como uma descrição definida correferente (HEIM, 1982, p. 25). Assim, as sentenças "just one man drank champagne. He was ill" são interpretadas como "just one man drank champagne. The man who drank champagne was ill": "he [Evans] assumes that a sentence containing a singular definite description with the predicate 'F' implies that there is exactly one object that is F and is true if that unique F satisfies whatever the sentence predicates of 'the F' " (HEIM, 1982, p. 26).

A hipótese de Evans (1980) de que anáforas não ligadas podem ser explicadas por uma

categoria extra de pronomes que funcionam como descrições sustenta-se sobre dois argumentos, portanto: i) problemas relativos a escopo de quantificadores; e ii) pressuposições de totalidade ou unicidade que pronomes ligados não acarretam. O segundo argumento é bem mais importante para a hipótese que o primeiro. De fato, críticas à análise *E-type* tenderam a enfatizar que pronomes *E-type* não exigem, ou, pelo menos, não com a força descrita por Evans, a totalidade ou unicidade do conjunto de objetos retomado pelo pronome, de modo que não há motivos para considerá-los uma categoria à parte de pronomes: "it is this commitment to uniqueness presuppositions that Heim (1982) and other DR [Discourse Representation] theorists perceive as the Achilles heel of the *E-type* analysis" (HEIM, 1990, p. 142).

5.2.2 Pronomes *E-type* segundo a Discourse Representation Theory

A *Discourse Representation Theory* (DRT) foi desenvolvida de maneira paralela por Kamp (2002)[5] e Heim (1982). Ela será apresentada tal qual descrita pelo primeiro autor e não se entrará em detalhes aqui quanto às semelhanças e diferenças entre os dois modelos.

Kamp (2002, p. 189) comenta que a semântica, à época da publicação original do artigo, pelo menos, se divide entre modelos que se propõe a, para cada sentença, especificar suas condições de verdade, e modelos que objetivam, para cada sentença, formalizar o que um falante apreende quando a entende. A DRT deve ser, então, justamente uma teoria que, ao anular essa divisão, traga mais poder explicativo para a Semântica. Para tanto, precisa formalizar a interpretação semântica de uma sentença considerando não só seus componentes, mas também o fato de que ela é afetada pelas outras sentenças do discurso no qual se encontra. "Discurso", deve-se lembrar aqui, é meramente um certo número de sentenças que devem ser interpretadas conjuntamente e numa certa ordem. Essa ênfase no processamento *online* do significado é o que caracteriza a DRT como um modelo *dinâmico*:

> *The principal respect in which DRT differs from the formal approaches to the analysis of meaning in natural language that existed at the time when it was conceived is the attention it pays to the systematic ways in which the interpretation of words and sentential constructions may depend on the discourse context, such as it is given by the sentence or sentences with which the given sentence co-occurs in a connected discourse or text, and to the intersentential semantic relations that are created by such dependences (HAMM; KAMP; VAN LAMBALGEN, 2006, p. 3).*

Para tanto, emprega dispositivos formais, as *Discourse Representation Structures* (DRS), que mantêm um inventário de referentes discursivos e proposições que é atualizado à medida em que o discurso é processado. De fato, a DRT almeja ter uma realidade psicolinguística:

> *[...] discourse representations can be regarded as the mental representations which speakers form in response to the verbal inputs they receive. [...] I conjecture that*

[5] O hoje célebre artigo em que Kamp introduz a DRT foi originalmente publicado em 1981. 2002 é o ano de publicação do livro no qual ele foi republicado e que foi consultado para esta pesquisa.

> *the structures which speakers of a language can be non-trivially described as forming to represent verbal contents are, if not formally identical, then at least very similar to the representations here defined (KAMP, 2002, p. 192–193).*

Os referentes discursivos representados nas DRSs são "aquilo sobre o qual um dado discurso versa, independentemente de sua real ou concreta existência" (SIMÃO, 2015, p. 46).

A DRT é formalmente composta por i) uma definição formal recursiva do conjunto de todas as DRSs bem formadas e um modelo semântico para os membros desse conjunto, e ii) um procedimento de construção que especifique como estender ou expandir uma DRS a partir de uma determinada sentença (GEURTS; BEAVER; MAIER, 2016), (SIMÃO, 2015).

Assim, ao discurso (10-a) corresponde a DRS (10-b):

(10) a. Pedro owns Chiquita. He beats her.

b.

u	*v*
Pedro owns Chiquita	
u = Pedro	
v = Chiquita	
u owns v	
he beats her	
u beats her	
u beats v	

(KAMP, 2002, p. 194–195).

Daí se infere que a maioria dos sintagmas nominais, incluindo nomes próprios e descrições indefinidas[6], introduzem referentes em uma DRS; já pronomes só podem correferir com os elementos de uma DRS (KAMP, 2002, p. 197). Nesse sentido, o que é mais importante observar aqui é que o problema *E-type* é resolvido pela DRT simplesmente porque não se coloca para ela: o próprio modo pelo qual referentes são retomados nas DRS postuladas pela DRT impede que surjam as restrições aparentemente paradoxais à referência de pronomes em certos contextos.

[6] Note-se que a ideia de que descrições indefinidas têm poder referencial é diametralmente oposta a tudo que se supôs sobre sua semântica até aqui (e em toda a bibliografia consultada para esta pesquisa, também). Esse é um ponto de discordância entre a DRT e outros modelos semânticos interessante, mas que não será realmente discutido aqui.

Parte II

Resolução de anáfora

6 Materiais

Python é uma linguagem de programação orientada a objetos criada no começo da década de 1990 por Guido van Rossum que, embora mais lenta que linguagens como C, tem sido bastante empregada em PLN devido a seu alto nível[1], que a torna particularmente amigável a usuários pouco familiares com programação, e a sua dinamicidade e praticidade. Nesta pesquisa utilizou-se a versão 3.5.2 da linguagem. As versões 3 e posteriores de Python são significantemente mais adequadas a análises do português que as anteriores porque codificam cadeias de caracteres em UTF-8 por padrão. Até Python 2, cadeias em UTF-8 eram objetos da classe `unicode`, e cadeias em ASCII, objetos `str`. Considerando que a ortografia do português utiliza muitas letras que contêm caracteres não ASCII, como aqueles com diacríticos, operar com cadeias de caracteres em UTF-8 nativamente, isto é, sem precisar convertê-las para cadeias ASCII, é bastante prático e eficiente. A documentação de Python 3.5.2 pode ser encontrada em <`https://docs.python.org/3.5/index.html`> (acesso em 12 de novembro de 2016).

Natural Language Toolkit (NLTK) é uma biblioteca para Python dedicada a PLN desenvolvida por Bird, Klein e Loper (2009).

Esta pesquisa tem como um de seus objetivos implementar em Python um algoritmo de resolução de anáfora para o português brasileiro. Há vários trabalhos sobre resolução de anáfora, inclusive alguns para o português, e a maior parte deles emprega métodos de força bruta. Isso tem como vantagens a exigência de pouca ou nenhuma anotação do corpus sendo processado (e, portanto, um número maior de textos passíveis de serem analisados) e uma taxa de sucesso bastante alta. Esse tipo de estratégia não representa, contudo, a maneira com que falantes efetivamente resolvem anáforas. Essa não é uma falha, limitação ou demérito desses algoritmos; simplesmente não é esse o propósito deles. São desenvolvidos para resolver anáforas com a maior taxa de acerto possível, sem preocupações quanto à realidade sintático-semântica ou psicolinguística dos métodos empregados, e nisso têm sido exitosos. No entanto, como esta é uma pesquisa primeiramente linguística, não se pretendeu aqui que a resolução do pronome "eles" alcançasse uma alta taxa de sucesso (embora este seja certamente um objetivo secundário), mas sim que procedesse de modo a poder aspirar a representar pelo menos parcialmente a resolução de anáforas feita por falantes reais.

Dados esses propósitos, considerou-se necessário ter acesso a informações sintáticas sobre as sentenças lidas, além de suas respectivas POS tags, uma exigência que reduziu bastante o

[1] Grosseiramente falando, o "nível" de uma linguagem de programação é o "quão distante" ela está das operações feitas pelo *hardware*. Operar em um nível alto acarreta um aumento no tempo de processamento, mas melhora a legibilidade do código e suaviza a curva de aprendizado do programador.

número de corpora adequados a serem processados aqui. Eventualmente, optou-se pelo Corpus Histórico do Português Tycho Brahe (CHPTB) (GALVES; FARIA, 2010), que é "um corpus eletrônico anotado, composto de textos em português escritos por autores nascidos entre 1380 e 1881"[2].

O programa foi desenvolvido e testado em um computador com 3,8GB de memória RAM operando com o sistema operacional Ubuntu GNOME 16.04.1, uma distribuição Linux.

[2] Segundo o *site* no qual o corpus está hospedado: <http://www.tycho.iel.unicamp.br/corpus/index.html>. Acesso em 15 de fevereiro de 2017.

7 Métodos

Os arquivos sintaticamente anotados do CHPTB podem ser baixados como um arquivo comprimido no repositório *online* do projeto. Sendo assim, o programa desenvolvido nessa pesquisa espera ser executado como nesse exemplo:
`$python3 anafora.py tycho_psd a_003_psd`. Com efeito, esse foi exatamente o comando utilizado para gerar os resultados descritos aqui. Sua sintaxe é melhor explicada a seguir:

1. O primeiro argumento indica que o segundo argumento, o arquivo `anafora.py`, deve ser interpretado pela versão 3 da linguagem Python.

2. O terceiro argumento é o diretório no qual estão os arquivos do corpus: qualquer arquivo com extensão `.txt` dentro dele será processado, e outras extensões serão ignoradas. Essa seleção pela extensão do arquivo pode ser facilmente reconfigurada. É opcional incluir "/", o caractere que delimita diretórios em um caminho de arquivo, nesse argumento. Como os caminhos completos do arquivo `.py` e desse diretório não estão especificados, deve-se estar, quando da execução desse comando, no diretório que os contém (ou, alternativamente, especificar os caminhos completos).

3. O último argumento é uma sequência de nomes de arquivos que, embora localizados dentro do diretório especificado no argumento anterior, devem ser ignorados pelo programa. Esse argumento é opcional e tem um tamanho arbitrário.

A estrutura dos arquivos sintaticamente anotados do CHPTB é tal que sentenças são separadas por uma linha em branco, e seus sintagmas são demarcados por parênteses, conforme exemplificado na figura 7.

Embora a biblioteca NLTK disponha, em seu módulo `tree`, de classes de objetos dedicadas a estruturas hierárquicas, considerou-se necessário criar uma classe customizada, denominada "TBtree", de modo que a cada árvore sintática do CHPTB correspondesse exatamente um objeto `TBtree`. Cada objeto `TBtree` tem uma lista, denominada `.plain`, associada a si. Inicialmente vazia, essa lista deve ser preenchida com as cadeias de caracteres que compõe a árvore. Uma outra classe própria, "Corpus", foi criada para conter o CHPTB inteiro. `Corpus` tem um atributo, `.forest`, que é uma lista, inicialmente vazia, a ser preenchida com os objetos `TBtree` correspondentes às sentenças do CHPTB. Assim, o programa lê os arquivos do CHPTB do seguinte modo:

```
( (IP-MAT (NP-SBJ (NP (NPR Dona) (NPR Maria) (NPR Antónia))
  (CONJP (CONJ e)
 (NP (NPR Dona) (NPR Bernarda) (NPR Campos))))
  (, ,)
  (IP-GER (VB-G mostrando-)
  (NP-SE (CL -se))
  (NP-ACC (NP (Q muito) (PRO$-F-P minhas) (N-P amigas))
  (CONJP (CONJ e)
 (NP (Q muito) (PRO$-F-P minhas) (VB-AN-F-P interessadas)))))
  (, ,)
  (VB-D quiseram)
  (IP-INF (VB saber)
  (ADVP (ADV logo))
  (NP-ACC (CP-FRL (WNP-1 (D o) (WPRO que))
  (IP-SUB (NP-ACC *T*-1)
  (NP-SBJ (PRO eu))
  (TR-D tinha)
  (VB-PP passado)
  (PP (P com)
      (NP (PRO$-F Sua) (NPR Majestade)))))))
  (. .))
  (ID A_004,60.820))
```

Figura 7 – Um exemplo de sentença sintaticamente anotada do CHPTB.

Fonte: Galves e Faria (2010).

1. Atribua à variável `c` um objeto `Corpus` criado a partir do diretório cujo caminho foi passado como argumento para a execução do programa ignorando eventuais arquivos que tenham sido passados como o último argumento.

2. Abra um dos arquivos do corpus (seguindo ordem alfabética). Caso não haja mais arquivos a serem lidos, termine essa execução;

3. Atribua à variável `t` um novo objeto `TBtree`;

4. Leia uma linha do arquivo de texto atual (caso o arquivo tenha terminado, volte ao passo 2). Se a linha contém apenas o caractere nova linha[1] ou se começa por "(CODE", considere-a uma linha vazia;

5. Se a linha lida é vazia e `t` não é uma árvore vazia, "feche" `t`, adicione `t` a `c.forest`, "abra uma nova árvore" (isto é, atribua a `t` um novo objeto `TBtree`) e volte ao passo 4;

[1] Embora a composição do caractere nova linha varie conforme o sistema operacional, convenientemente, em Python, ele pode ser sempre representado por "`\n`".

6. Se a linha lida é vazia mas `t` ainda está vazia, apenas volte ao passo 4;

7. Se a linha lida não é vazia, adicione-a à árvore `t` e volte ao passo 4.

"Fechar" uma árvore `TBtree` é transformar as cadeias de caracteres armazenadas na lista `plain` em um objeto da classe `ParentedTree`, implementada pela biblioteca NLTK. Objetos dessa classe são "árvores" nas quais informações sobre parentesco e dominância entre nós podem ser codificadas com facilidade (mais facilmente do que a classe `Tree` permite). Assim, quando uma árvore `TBtree` está sendo "fechada", os seguintes passos são seguidos:

Seja `t` um objeto `TBtree`:

1. a última linha de cada árvore `TBtree` (ou seja, aquela imediatamente anterior a uma linha em branco) é uma informação meta-textual, uma etiqueta que identifica a qual arquivo aquela sentença sintagmatizada pertence e qual sua posição dentro dele. Sendo assim, a última cadeia na lista `t.plain` é removida da lista e armazenada, sem a palavra "ID" e parênteses, em `t.ID`;

2. as cadeias de caracteres na lista `t.plain` são concatenadas, separadas por um espaço em branco, em uma grande cadeia guardada na própria variável `t.plain`;

3. observe-se que o passo anterior remove, além da etiqueta da sentença, também o parênteses que corresponde ao primeiro parênteses da árvore. Isto é desejável, na verdade, porque a notação usada no CHPTB e o modo pelo qual as árvores da classe `ParentedTree`, de NLTK, são geradas fazem com que a raiz de cada árvore seja um nó vazio, o que é redundante, embora não represente um prejuízo sério. Assim, o parênteses mais externo da cadeia de caracteres em `t.plain` é removido;

4. nós que representem sinais de pontuação são removidos. Embora sinais de pontuação sejam relevantes para certas análises, concluiu-se que, para essa pesquisa, mantê-los pioraria a legibilidade das árvores e obrigaria o programa a fazer certas checagens e procedimentos inúteis (o que o deixaria mais lento, também). Considera-se nó de pontuação uma cadeia de caracteres do tipo "(X X)" ou "(PUNC X)", sendo "X" um sinal de pontuação qualquer. Tais nós foram identificados por meio de expressões regulares;

5. tenta-se gerar um objeto `ParentedTree` a partir de `t.plain`. Isso é possível porque essa classe de NLTK tem um método, `fromstring`, que, dada uma cadeia de caracteres com uma certa sintaxe, retorna uma instância da classe. Caso essa operação seja bem sucedida, a árvore gerada é armazenada em `t.tree`. Do contrário, um objeto `None` é retornado imediatamente;

6. uma versão binária de `t.tree` é armazenada em `t.cnf`. A binarização é feita por um método de NLTK, `chomsky_normal_form`, que, para tanto, executa basicamente dois procedimentos: i) nós n-ários, sendo $n > 2$, são expandidos de modo que os nós à direita do nó mais à esquerda tornem-se um único sintagma, seu irmão, e o procedimento é repetido até que não haja nenhum nó com mais de 2 filhos; e ii) ao nome de cada nó são adicionados os nomes dos respectivos filhos;

7. as palavras que integram a sentença são salvas como uma lista em `t.words`, e `t.pos` é uma lista contendo duplas (tuplas de tamanho 2) na qual o primeiro elemento é um nó terminal e o segundo é sua POS tag. Note-se que os elementos de `t.words` e os nós terminais de `t.pos` nem sempre são exatamente as palavras que compõe a sentença, embora toda palavra sempre pertença a ambas as listas. Isso porque certas categorias vazias são anotadas nos arquivos do CHPTB, de modo que não constituem palavras da sentença, mas integram `t.words` e `t.pos`;

8. `t`, o objeto `TBtree`, é retornado.

Resumidamente, um objeto `Corpus` é criado e começa-se a ler os arquivos do CHPTB. Os arquivos são "parseados", ou seja, transformados em objetos `TBtree` que são inseridos em `Corpus.forest` conforme são lidos.

Há duas funções, `contains` e `ccomands`, que representam os conceitos sintáticos de dominância e c-comando, respectivamente.

O código de 8 é bastante intuitivo: o método `.subtrees`, implementado por NLTK, retorna, para uma dada árvore, a lista de todas as árvores que ela contém (inclusive ela própria, que será sempre o primeiro elemento dessa lista). Assim, para verificar se uma árvore `x` contém uma `y`, basta iterar sobre os elementos de `x.subtrees()`: caso um deles seja y, a função retornará `True`. Do contrário, precisa retornar `False`. A aparente complexidade se deve ao fato de que antes dos testes pertinentes a c-comando são executados alguns testes para evitar checagens desnecessárias, como verificar se um sintagma c-comanda outro que sequer ocorre em sua árvore.

```
def contains(partree1, partree2):
    for s in partree1.subtrees():
        if s == partree2:
            return True
    return False
```

Figura 8 – A função `contains`.

A implementação de 9 parece mais complexa mas na verdade não o é. Como explicado na definição 3 da página 16, um sintagma `x` c-comanda um sintagma `y` caso eles sejam irmãos ou o irmão de `x` domine `y`. Assim, `ccomands(x, y)` verifica se `y` pertence a uma dupla cujos elementos são os irmãos de `x`. Em caso afirmativo, a função retorna `True`. Do contrário, verifica se o irmão de `x` domina `y`. Isso é feito com a função `contains` já descrita.

```
def ccomands(cmd, obj, cnf=True):
    if type(cmd) is not type(obj):
        raise ValueError("Árvores de tipo diferente.")

    if type(cmd) is TBtree and type(obj) is TBtree:
        if cnf:
            a, b = cmd.cnf, obj.cnf
        else:
            a, b = cmd.tree, obj.tree
    else:
        a, b = cmd, obj

    if any((
    a.root() != b.root(),
    contains(a, b),
    contains(b, a)
    )):
        return False
    # irmãos
    if b in (a.right_sibling(), a.left_sibling()):
        return True
    # irmão domina
    for s in (a.right_sibling(), a.left_sibling()):
        if s and contains(s, b):
            return True
    return False
```

Figura 9 – A função `ccomands`.

Sendo o programa capaz de detectar relações de dominância e c-comando, espera-se que já consiga fazer algumas previsões sobre quais sintagmas podem ser o antecedente de um pronome “eles” para uma dada árvore do CHPTB. Para tanto, adota-se o seguinte procedimento, cujo código é reproduzido em 10:

1. Armazena-se na lista `Corpus.woods` as sentenças em `Corpus.forest` que contêm exa-

tamente uma ocorrência da palavra “eles”[2];

2. Itera-se sobre `Corpus.woods`, “pegando” cada árvore `TBtree` de uma vez e colocando-a na variável `elestree`;

3. Atribui-se à variável `anafora` o sintagma de `elestree` que contém o pronome de terceira pessoa plural;

4. A variável `current` recebe o valor da mãe de `anafora`;

5. Verifica-se se o nó à esquerda de `current` contém algum sintagma cuja POS tag seja “N-P”, que identifica sintagmas nominais plurais. A mãe de um sintagma rotulado por “N-P” é adicionada à lista `antecedents` de `elestree`[3]. Adiciona-se a mãe do nó, e não o próprio nó, porque espera-se, considerando-se a sintaxe do português e o padrão de anotação morfológica usado no CHPTB, que constituintes rotulados por “N-P” sejam NPs que tenham como irmão à esquerda um determinante, ou um pronome possessivo etc.;

6. Caso não haja mais nenhum nó à esquerda, `current` passa a ser a mãe do atual `current`. Caso o topo da árvore já tenha sido atingido, passa-se ao próximo elemento de `Corpus.woods`.

A função `extractor` é, provavelmente, o coração do programa aqui desenvolvido. Sua maneira de proceder, “olhando” à esquerda e acima (nessa ordem) em busca de sintagmas cuja POS tag seja “N-P”, é inspirada pela aplicação do algoritmo de Hobbs ao português brasileiro descrita em Santos (2008, p. 23): “a resolução [de anáfora] é realizada [pelo algoritmo de Hobbs] através de uma busca em largura, na árvore sintática da sentença, da esquerda para a direita, procurando por sintagmas nominais compatíveis em gênero e número com o pronome”.

Comenta-se na próxima seção que tipo de resultado essa análise consegue retornar.

[2] O ideal, certamente, é que o programa consiga lidar com qualquer quantidade de ocorrências de “eles”, mas optou-se por simplificar esse objetivo a princípio, uma vez que presume-se ser mais fácil obter êxito com casos mais simples e depois estender o procedimento para os mais complexos do que tentar, já de início, lidar com toda a complexidade sintática de sentenças com múltiplos “eles”. .

[3] Há duas simplificações bastantes fortes aqui: i) “olhar” somente à esquerda impossibilita a análise de catáforas; e ii) nem sempre o antecedente de um pronome plural é também plural. Considere-se por exemplo este discurso: “João e Fabiana se casaram ontem. Eles optaram por uma cerimônia simples”. O antecedente aqui é um sintagma nominal coordenado no qual há dois sintagmas nominais singulares, o que indica que restringir a busca por antecedentes de “eles” a sintagmas cuja POS tag contenha a marca “P”, de pluralidade, é uma estratégia insuficiente. Como já comentado, num primeiro momento pretende-se tornar o algoritmo tão simples quanto possível e aumentar seu poder explicativo gradativamente em pesquisas futuras.

```
def extractor(self, cnf=True):
        F = self.forest
        woods = []

        for tbt in F:
            pos = tbt.pos
            words = [ w.lower() for w in tbt.words ]
            if words.count(eles) != 1:
                continue
            woods.append(tbt)
            if cnf:
                t = tbt.cnf
            else:
                t = tbt.tree
            for sub in t.subtrees(lambda x: x.height() == 2):
                if sub[-1].lower() == eles:
                    anafora = sub
                    break
            # olhar à esquerda e depois subir
            current = anafora.parent()
            while current != anafora.root():
                try:
                    nps = get_NPs(current.left_sibling())
                except AttributeError:
                    current = current.parent()
                    continue
                tbt.antecedents.extend(nps)
                for n in nps:
                    print(t.pformat_latex_qtree())
                    print("Antecedente:", n.parent())
                    print("C-comanda:", ccomands(n, anafora))
                    if ccomands(n, anafora):
                        self.bound += 1
                    else:
                        self.unbound += 1
                current = current.parent()
        return woods
```

Figura 10 – A função `extractor`.

8 Resultados

Como explicado no capítulo 7, foram analisadas sentenças com exatamente uma ocorrência de “eles”. Assim, de um total de 30.126 sentenças, 234 foram processadas, e desta parcela, 34 “antecedentes” do pronome “eles” o c-comandam e 221 não o c-comandam. Colocou-se “antecedentes” entre aspas porque, dadas as capacidades atuais do programa, não há nenhuma garantia de que um sintagma nominal identificado como antecedente de um pronome de fato o seja; o que se pode afirmar seguramente de um desses sintagmas é que de fato se trata de um sintagma nominal plural (assim o garante sua POS tag) e que ele precede linearmente o pronome, podendo ou não c-comandá-lo (relação que o programa consegue detectar com sucesso). Esses dados simples são o que de mais quantitativo se pode concluir dos resultados do programa. A seguir, numa tentativa de exibir os resultados dessa pesquisa de maneira mais qualitativa, toma-se uma das 234 sentenças processadas como exemplo:

(11) *pro* como *pro* via os nossos religiosos na igreja chegava-se a eles

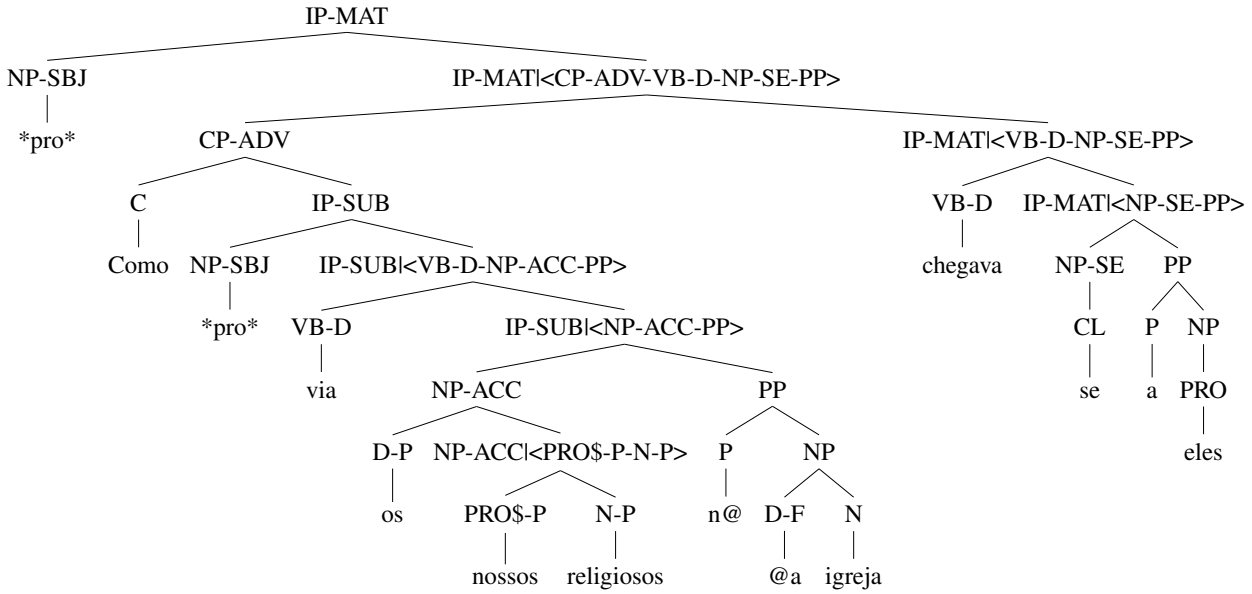


Figura 11 – A anotação sintática da sentença (11) no CHPTB.

Note-se que a anotação morfológica do CHPTB (isto é, que atribui POS tags aos tokens do corpus) registra entre asteriscos certos constituintes que têm um papel sintático, mas não

são pronunciados. É o caso de *pro*, por exemplo. Além disso, a formatação do CHPTB é tal que clíticos são tratados como tokens independentes e preposições são separadas dos constituintes a que se aglutinam, como determinantes; uma arroba marca a contração que foi desfeita. Como comentado no capítulo 7, as árvores n-árias do CHPTB são binarizadas antes que se prossiga à resolução de anáfora propriamente dita; a árvore na figura 11 é a representação dessa estrutura já binarizada. Note-se que, sendo uma reestruturação automática, essa operação inevitavelmente agrupará constituintes sem relação entre si e desconectará constituintes claramente relacionados. Isso pode ser observado mesmo na figura 11: há um sintagma constituído por um clítico, um determinante e um pronome ("chegava-[se a eles]"). Assim, a binarização descrita anteriormente é um ponto a ser melhorado: haveria modos mais precisos para converter as árvores n-árias do CHPTB em árvores binárias, mais adequadas ao tipo de análise sintática executada pelo programa? Considerou-se aqui, contudo, que os ganhos oriundos dessa operação automática compensam e justificam as imprecisões que ela acarreta.

O programa informa que o antecedente de "eles" na sentença (11) é o sintagma "nossos religiosos", cuja POS tag é "NP-ACC"[1]. De fato, nesta sentença, esse é o único sintagma que, por concordância de número, pode ser o antecedente de "eles"[2]. É informado também que "eles" não é c-comandado por tal antecedente, o que se confirma pela relação posicional entre "nossos religiosos" e "eles".

Essas conclusões são alcançadas pela execução dos procedimentos descritos no capítulo 7: identifica-se em que nó se encontra o pronome procurado; a partir dessa posição, percorre-se a árvore da direita para a esquerda, até que não haja constituintes à esquerda para serem checados, e de baixo para cima, até que se atinja o nó máximo da árvore. Verifica-se em cada nó se sua POS tag é "N-P", e, em caso positivo, a mãe do nó é considerada um antecedente do pronome.

Embora esses resultados sejam bastante preliminares e estejam aquém do que se pretendia alcançar inicialmente, é importante observar que eles dão pistas sobre como ampliar o programa com estratégias baseadas na análise *E-type* e na análise da DRT.

Evans (1980, p. 341) afirma, sobre a distribuição sintática dos pronomes ligados, que "a pronoun will be interpreted as bound by a quantifier phrase only if it precedes and c-commands the pronoun". Isto é, avaliar se dois constituintes em uma árvore sintática estão em relação de c-comando é, segundo a análise *E-type*, uma capacidade indispensável para que se distingua um pronome *E-type* de um pronome ligado, o que por sua vez é necessário para que a resolução de anáfora proceda substituindo os pronomes do primeiro tipo por descrições definidas correferentes.

[1] Na realidade, "NP-ACC|<PRO$-P-N-P>", mas o conteúdo do caractere "|" em diante é apenas subproduto da binarização previamente executada pelo programa.

[2] Como já comentado, o programa exige antecedentes plurais para "eles", o que leva ao descarte de antecedentes perfeitamente adequados, como a coordenação entre dois sintagmas nominais singulares.

Quanto à análise da DRT, elaborar uma lista com candidatos a antecedente de um determinado pronome “eles” é condição necessária para a elaboração de uma DRS que represente corretamente a sentença (ou, melhor, o discurso) sendo processada — embora certamente esteja longe de ser uma condição suficiente para isso.

Mais comentários sobre os possíveis desenvolvimentos futuros que os resultados permitem esperar são feitos no próximo capítulo.

9 Considerações finais

Procurou-se discutir o problema dos pronomes E-type, duas modelagens semânticas distintas para eles e meios pelos quais cada uma poderia ser implementada computacionalmente num programa de resolução de anáfora, de modo que se pudesse verificar se a adoção de uma em ou outra impactaria a eficiência do programa — e, em caso positivo, em que medida e por quais motivos. Para isso considerou-se necessário apresentar, antes, alguns aspectos rudimentares de Lógica, Semântica Formal e TRL (especialmente aquilo que concerne o módulo LF da teoria) comuns a ambos os modelos.

Não foi possível, pelo menos por enquanto, fazer com que o programa implementasse estratégias de resolução de anáfora conforme a análise E-type ou conforme a DRT, e nem mesmo que conseguisse discriminar, de uma lista de possíveis antecedentes para uma dada ocorrência de "eles", aquele com a maior probabilidade de sê-lo. Conseguiu-se que o programa transformasse as sentenças sintaticamente anotadas do Corpus Histórico do Português Tycho Brahe em objetos de fácil manipulação pela linguagem Python enriquecida pela biblioteca NLTK, detectasse relações de dominância e c-comando, e construísse, para cada instância de "eles", uma lista de possíveis antecedentes.

O trabalho desenvolvido aqui entrevê a possibilidade de pesquisas futuras que revisem falhas conceituais e metodológicas e ampliem o programa desenvolvido de modo a torná-lo mais poderoso e eficiente. São inventariados a seguir alguns pontos que parecem particularmente propícios a melhoramentos vindouros:

- A descrição da análise *E-type* foi feita com base em apenas Evans (1980) e Heim e Kratzer (1998), mas pode ser enriquecida pelo cotejamento com Evans (1997), Cooper (1979) e Parsons (1978)[1]. Evans (1977) discute em mais detalhes certos aspectos da análise *E-type* que em Evans (1980) são apenas mencionados (EVANS, 1980, p. 337). Parsons (1978) e Cooper (1979) avançam modelos semânticos para os pronomes *E-type* que mantêm, em linhas gerais, as ideias de Evans[2]. Neste último trabalho, por exemplo, propõe-se que um pronome *E-type* seja composto por

[1] Cooper, R. The interpretation of pronouns. In: HENY, F.; SCHNELLE, H. (ed.) **Syntax and Semantics 10**, Selections from the Third Groningen Round Table. Nova York: Academic Press, 1979.
Evans, G. Pronouns, quantifiers, and relative clauses (I). In: **Canadian Journal of Philosophy**, v. 7, n. 3, 1977, pp. 467–536.
Parsons, T. **Pronouns as paraphrases**. Manuscrito. University of Massachusetts Amherst, Amherst, 1978.

[2] Cooper (1979) e Evans (1977, 1980) são descritos por Heim (1990, p. 137) como "a family of related proposals", embora a autora reconheça nuances entre as propostas: Cooper (1979) seria um modelo "mais pragmático" que Evans (1979), por exemplo (HEIM, 1990, p. 169).

> *a definite article and a predicate that is made up of two variables. The first variable is of type $\langle e, \langle e,t \rangle \rangle$ and remains free in the sentence as a whole. The second variable is of type e, and typically gets bound in the sentence [...]. We may think of these variables as unpronounced pro-forms (HEIM; KRATZER, 1998, p. 290).*

Assim, considerar esses trabalhos pode ser importante na medida em que eles se propõem a descrever precisamente de que modo um pronome *E-type* é interpretado como uma descrição definida, exatamente o que se procura se se quer implementar a análise *E-type* computacionalmente.

- Podem ser discutidas as diferenças entre as DRTs propostas por Kamp (2002) e Heim (1982), um ponto que aqui foi meramente mencionado.

- Pode-se aprofundar a discussão dos argumentos contrários à análise *E-type* levantados em trabalhos como Heim (1982), Heim (1990) e Kamp (2002), bem como de eventuais réplicas por parte de adeptos da análise *E-type*. Nesta pesquisa, optou-se por não entrar no mérito de argumentar em favor de uma das análises. Na verdade, mesmo a dicotomia entre elas não pode ser tomada como óbvia: elas são mutuamente exclusivas ou podem se complementar ou mesmo combinar?

- É já bastante conhecido que a anaforicidade por vezes exige concordância de número e gênero entre o pronome e seu antecedente. Isso pôde ser devidamente implementado porque a própria POS tag dos sintagmas já codifica essas informações. Outros fatores importantes para a resolução de anáfora que não se conseguiu considerar são a animacidade e a distância linear entre antecedente e pronome. A animacidade talvez possa ser verificada recorrendo-se a uma ferramente como o WordNet, uma ferramenta inclusa em NLTK capaz de, por exemplo, listar os hiperônimos de um vocábulo, o que pode talvez conter pistas semânticas relativas a animacidade. Sobre o segundo item, sabe-se que uma distância linear pequena entre um possível antecedente e o pronome torna mais provável que ele seja de fato o antecedente correto. Podem haver estratégias computacionais que permitam codificar isso corretamente.

- O algoritmo implementado nesta pesquisa parece assemelhar-se em alguns pontos ao algoritmo de Hobbs, bastante conhecido no campo da resolução de anáfora. Pesquisas futuras podem aprimorar o programa desenvolvido aqui ao incorporar estratégias bem sucedidas desse algoritmo (SANTOS, 2008).

- Considerando a gravidade das falhas estruturais de algumas árvores resultantes da binarização aplicada sobre as árvores n-árias do CHPTB e a importância dessa operação para as análises executadas pelo programa, é preciso investigar em que medida essa reestruturação pode ser otimizada.

Referências

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. O'Reilly Media, 2009. Disponível em: <http://www.nltk.org/book/>. Acesso em: 12 nov. 2016.

BORGES NETO, J. Semântica de modelos. In: FOLTRAN, M. J.; MÜLLER, A.; NEGRÃO, E. V. (Ed.). **Semântica formal**. São Paulo: Editora Contexto, 2002.

CANN, R.; KEMPSON, R. M.; GREGOROMICHELAKI, E. Anaphora, discourse and context. In: ____. **Semantics**: an introduction to meaning in language. Nova York: Cambridge University Press, 2009. (Cambridge Textbooks in Linguistics, v. 198), p. 137–168.

CHIERCHIA, G. **Semântica**. Tradução de Luiz Arthur Pagani, Lígia Negri e Rodolfo Ilari. Campinas: Editora da UNICAMP, 2003.

CYRINO, S. M. L. **O objeto nulo do português do Brasil**: um estudo sintático-diacrônico. 217 f. Tese (Doutor em Ciências) — Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, 1994. Disponível em: <http://www.bibliotecadigital.unicamp.br/document/?code=vtls000081626>. Acesso em: 5 nov. 2016.

EVANS, G. Pronouns. **Linguistic Inquiry**, MIT Press, v. 11, n. 2, p. 337–362, 1980. Disponível em: <http://www.jstor.org/stable/4178164>. Acesso em: 11 mar. 2016.

GALVES, C. M. C.; FARIA, P. P. F. de. **Tycho Brahe Parsed Corpus of Historical Portuguese**. 2010. Base de dados na Internet. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html>. Acesso em: 12 nov. 2016.

GEURTS, B.; BEAVER, D. I.; MAIER, E. Discourse representation theory. In: ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Spring 2016. Stanford: Metaphysics Research Lab, Stanford University, 2016. Disponível em: <http://plato.stanford.edu/archives/spr2016/entries/discourse-representation-theory/>. Acesso em: 23 nov. 2016.

HAACK, S. **Philosophy of logics**. Cambridge: Cambridge University Press, 1978.

HAMM, F.; KAMP, H.; VAN LAMBALGEN, M. There is no opposition between formal and cognitive semantics. **Theoretical Linguistics**, v. 32, n. 1, p. 1–40, set. 2006. Disponível em: <https://doi.org/10.1515/TL.2006.001>. Acesso em: 23 nov. 2016.

HEIM, I. R. **The Semantics of Definite and Indefinite Noun Phrases**. 426 f. Tese (Doctor of Philosophy) — Universidade de Massachusetts Amherst, Amherst, 1982. Disponível em: <http://scholarworks.umass.edu/dissertations/AAI8229562>. Acesso em: 13 set. 2016.

HEIM, I. R. E-type pronouns and donkey anaphora. **Linguistics and Philosophy**, Springer, v. 13, n. 2, p. 137–177, 1990. Disponível em: <http://www.jstor.org/stable/25001382>. Acesso em: 12 set. 2016.

HEIM, I. R.; KRATZER, A. **Semantics in generative grammar**. Malden: Blackwell, 1998. v. 13. (Blackwell Textbooks in Linguistics, v. 13).

HINTIKKA, J. Frege's hidden semantics. **Revue Internationale de Philosophie**, v. 33, n. 130, p. 716–722, 1979. Disponível em: <http://www.jstor.org/stable/23944067>. Acesso em: 22 out. 2016.

KAMP, H. A theory of truth and semantic representation. In: PARTEE, B. H.; PORTNER, P. (Ed.). **Formal Semantics**: The essential readings. Oxford: Blackwell Publishers Ltd, 2002. cap. 8, p. 189–222. Disponível em: <http://dx.doi.org/10.1002/9780470758335.ch8>. Acesso em: 13 set. 2016.

KING, J. C.; LEWIS, K. S. Anaphora. In: ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Summer 2016. Stanford: Metaphysics Research Lab, Stanford University, 2016. Disponível em: <http://plato.stanford.edu/archives/sum2016/entries/anaphora/>. Acesso em: 12 set. 2016.

LARSON, R. K.; SEGAL, G. M. A. **Knowledge of meaning**: an introduction to semantic theory. Cambridge: MIT Press, 1995.

MAY, R. **Logical form**: its structure. Cambridge: MIT Press, 1985. v. 12. (Linguistic Inquiry Monographs, v. 12).

MIOTO, C.; SILVA, M. C. F.; LOPES, R. E. V. **Novo manual de sintaxe**. São Paulo: Editora Contexto, 2013.

MÜLLER, A. Pronomes e anáfora — o estado da arte. **Linha D'Água**, v. 0, n. 16, p. 17–37, 2003. Disponível em: <http://www.revistas.usp.br/linhadagua/article/view/37247>. Acesso em: 9 mar. 2016.

RUSSELL, B. On denoting. **Mind**, Oxford University Press, v. 14, n. 56, p. 479–493, 1905. Disponível em: <http://www.jstor.org/stable/2248381>. Acesso em: 21 fev. 2014.

____. **The problems of philosophy**. Nova York: Henry Holt and Company, 1912. (Home University Library of Modern Knowledge, 35). Disponível em: <http://www.archive.org/details/problemsofphilo00russuoft>. Acesso em: 21 fev. 2014.

SANTOS, D. N. de A. **Resolução de anáfora pronominal em português utilizando o algoritmo de Hobbs**. 65 f. Dissertação (Mestre em Ciência da Computação) — Instituto de Computação, Universidade Estadual de Campinas, Campinas, 2008. Disponível em: <http://www.bibliotecadigital.unicamp.br/document/?code=000431264>. Acesso em: 12 nov. 2016.

SELLS, P. **Lectures on contemporary syntactic theories**. Chicago: The University of Chicago Press, 1985. (Stanford University Center for the Study of Language and Information Lecture Notes, 3).

SILVESTRE, R. S. **Um curso de lógica**. 1. ed. Petrópolis: Vozes, 2011.

SIMÃO, D. **Nonada, ou uma travessia entre semânticas**. 80 f. Dissertação (Mestre em Letras) — Setor de Ciências Humanas, Letras e Artes, Universidade Federal do Paraná, Curitiba, 2015. Disponível em: <http://hdl.handle.net/1884/40930>. Acesso em: 21 nov. 2016.

SZABOLCSI, A. Compositionality in focus. **Folia Linguistica**, De Gruyter, v. 15, n. 1–2, p. 141–162, jan 1981. ISSN 1614-7308. Disponível em: <http://www.nyu.edu/projects/szabolcsi/Szabolcsi_compositionality_in_focus.pdf>. Acesso em: 5 jun. 2016.

VAN HEIJENOORT, J. Logic as calculus and logic as language. **Synthese**, Springer, v. 17, n. 3, p. 324–330, 1967. Disponível em: <http://www.jstor.org/stable/20114564>. Acesso em: 24 out. 2016.