Capítulo 4

O modelo semântico usado no Primeiro HAREM

Diana Santos

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 4, p. 43–57, 2007.

Este capítulo fundamenta o modelo semântico desenvolvido para o Primeiro HAREM. Em primeiro lugar, são levantadas algumas questões de fundo sobre a semântica da linguagem natural, e a sua aplicação no caso específico do REM. Segue-se uma apologia relativamente longa, tentando rebater diversos contra-argumentos levantados em algumas ocasiões (como por exemplo o Encontro do HAREM), e justificando as bases teóricas do modelo adoptado.

Como mencionado no capítulo 1, a razão por que começámos a tratar de REM na Linguateca foi porque nos pareceu a tarefa mais básica possível a nível de semântica. Contudo, isto não significa que o REM seja propriamente uma tarefa fácil, ou que a maior parte das questões associadas ao PLN não acabe por surgir, quando se pretende delimitar rigorosamente o âmbito e o propósito desta tarefa.

4.1 O que é semântica?

Sendo um capítulo sobre a definição de uma tarefa semântica, é preciso começar por lembrar que não há realmente um grande consenso entre o que é a esfera da semântica, o que leva a que seja necessário que, até a esse nível!, estabeleçamos uma definição para que o capítulo possa fazer sentido.

Muito simplificadamente, a semântica ocupa-se da relação entre a forma (a língua) e o "mundo exterior à língua". Deixemos neste momento de parte a questão complexa de o que é este mundo e se ele existe realmente na Natureza ou apenas nas nossas mentes (ver Santos (1940)). Por outras palavras, a semântica tenta relacionar objectos linguísticos com objectos não linguísticos. Visto que o mundo em si não está acessível para as nossas análises, existe sempre um modelo ou conceptualização que medeia entre ele e a língua, ou seja, os investigadores em semântica constroem modelos que pretendem representar a realidade e tentam mapear a língua nesses modelos.

Devido à grande complexidade da tarefa, um mapeamento directo é raramente sugerido (mesmo quando se está a falar da relação entre a língua e um modelo conceptual parecido, como por exemplo a lógica de primeira ordem). As teorias semânticas recorrem a estruturas intermédias (como a DRT de Kamp e Reyle (1993)), a tipos especiais de raciocínio (como lógica não monotónica, ver Ginsberg (1987)) ou a representações especificamente desenhadas para emparelhar propriedades conhecidas da linguagem natural, tais como mundos possíveis (Hughes e Cresswell, 1968) para interpretar modalidade, ou guiões (Schank e Rieger, 1974) para fazer sentido de algum tipo de descrições esperadas.

Seja qual for a teoria que nos agrade mais, estou convencida de que ninguém discordará do seguinte: delimitar o conceito de entidade mencionada, como conceito semântico, tem a ver com a relação entre a língua e o mundo exterior à língua, mundo esse que é mediado/representado por um conjunto de símbolos que representam esse mundo. A tarefa

de REM, como qualquer tarefa semântica, passa por um conjunto de categorias, sobre as quais se tenta chegar a um entendimento partilhado.

Existem duas grandes escolas de análise semântica: a **denotacionalista**, onde os símbolos são um substituto de objectos exteriores, e a **funcionalista**, em que os símbolos representam a relação entre os objectos, ainda dentro da própria língua. Assim, uma parte importante do significado de um texto (ou sintagma, ou palavra) é a função que desempenha relativamente aos outros elementos do texto. Pode ver-se esta análise como mais um nível entre a língua (forma) e o mundo; em paralelo com a denotação, deve também ter-se em conta a função. (E a função é geralmente obtida de um conjunto de poucos valores, tais como os casos de Fillmore (1968)). Esta é uma forma de tentar explicar sistematicamente porque é que uma mesma expressão em contextos diferentes tem ou pode ter significados diferentes, que é uma das propriedades mais básicas e mais importantes da linguagem natural. Por outro lado, existe ainda outra escola a que chamarei **pragmática**, que defende que é o contexto que define o sentido, e que não há denotação fixa. Ou seja, as funções de cada elemento no texto dão-lhe um significado, juntamente com o contexto real de produção da frase.

Em qualquer caso, a análise semântica pressupõe sempre uma classificação em categorias, e essa classificação não é nada consensual na forma como é estruturada: são conjuntos baseados em semelhanças, ou em diferenças (Ellis, 1993)? Todos os membros de uma categoria são iguais, ou há membros mais fortes do que outros? Quais os limites e as relações entre as categorias? São mutuamente exclusivas ou, pelo contrário, hierarquicamente ou funcionalmente definidas?

Para não tornar este capítulo demasiado geral, vou apenas discutir estas questões na subtarefa de dar sentido aos nomes próprios, o REM. Antes disso, vou fazer uma digressão necessária pela questão da vagueza.

4.1.1 A importância da vagueza para a semântica

Um dos meus cavalos de batalha é a questão da vagueza na língua. Ao contrário de uma concepção bastante divulgada, que considera a vagueza como uma fraqueza da linguagem natural que deve ser reparada, reduzida ou pelo menos tratada (como doença), eu considero que a vagueza é uma das qualidades mais importantes e positivas da linguagem natural, que deve ser admirada e tratada (com respeito) de forma a não se perder o seu conteúdo.

Ao contrário de outras abordagens que apenas reconhecem o fenómeno da vagueza em ocorrências concretas da língua, eu considero que a vagueza existe tanto ao nível da competência como ao nível do desempenho, ou seja, quer globalmente como propriedade dos itens lexicais e das estruturas da língua (fora do contexto) – a competência – , quer ao nível da língua concreta, das frases em contexto — o desempenho.

Felizmente, existem vários linguistas e filósofos que partilham esta opinião, donde não é necessário começar por argumentar longamente sobre a necessidade de lidar com este tema. Basta-me remeter para maiores autoridades (Burns, 1991; Pustejovsky, 1995; Lakoff, 1987; Buitelaar, 1998; Cruse, 2004) que lidam com a vagueza, se bem que sob perspectivas diferentes, ou para outros textos meus (Santos, 1997, 2006d) que já tratem a vagueza em pormenor.

De forma a restringir o âmbito do presente capítulo, discutirei apenas a questão da vagueza associada a abordagens computacionais relacionadas com a formalização dos nomes próprios, portanto directamente relacionadas com a questão do HAREM.

4.2 O que é o REM?

Qualquer definição de REM depende fortemente do modelo semântico adoptado, e em particular, do seu lado extra-linguístico. No MUC definiram-se três conceitos principais que representam generalizações que se supunha existirem no mundo real: pessoas (`PERSON`), organizações (`ORGANIZATION`) e locais (`LOCATION`), e a tarefa de REM propunha reconhecer nomes próprios (uma restrição de forma) que apontassem ou correspondessem a essas categorias (fixadas de princípio) em textos jornalísticos escritos em inglês. Quando os nomes próprios encontrados no texto se referiam a outro tipo de entidades que não locais, pessoas ou organizações, não deviam ser reconhecidos, e assumiu-se que uma pessoa, uma organização e um local nunca poderiam coincidir (o que não é propriamente surpreendente).

No HAREM, nós estávamos interessados em **todos** os nomes próprios (definidos de forma bastante liberal), ao contrário de apenas um subconjunto de nomes próprios que tivessem uma dada denotação, para ter uma ideia do que a tarefa de REM significava para o português. Por isso começámos por tentar categorizar todos essas ocorrências em vários tipos de texto.

4.2.1 Metonímia

Porque não estávamos só à procura de casos simples, depressa nos demos conta do que muitos outros investigadores já tinham notado e formalizado antes de nós: que há muitos casos em que um nome originalmente usado para denotar um certo objecto é usado como substituto para outros objectos (ou entidades) que pertencem a um tipo completamente diferente. Por exemplo, em *Fontes próximas do Palácio de Belém desmentiram que...*, a entidade *Palácio de Belém* não se refere a um edifício, mas sim ao Presidente da República português, eventualmente secundado também pelo seu gabinete.

Ao contrário das opções que muitos seguiram, de formalizar e sistematizar essas substituições, nós adoptámos uma solução mais radical, ao marcar a entidade final de acordo

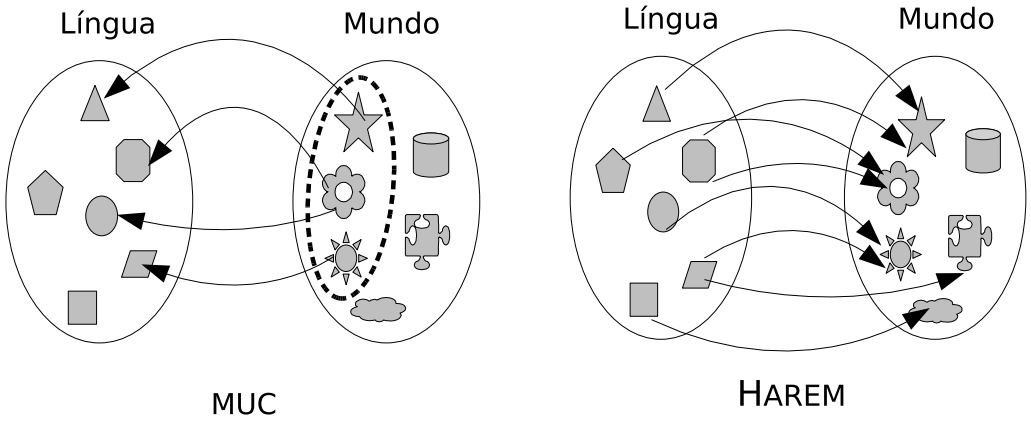

Figura 4.1: Dois pontos de partida diferentes para abordar a questão da semântica (do REM).

com o objecto denotado. (Que, no caso anterior, seria uma pessoa ou grupo de pessoas).

Este fenómeno é vulgarmente chamado **metonímia**, e pode ser definido como o caso em que uma expressão é usada para referir outro referente relacionado (veja-se Lakoff e Johnson (1980)). Exemplos conhecidos na literatura são o uso de *Vietname* para a guerra do Vietname, ou a *tosta mista* para o cliente que a encomendou, respectivamente nos seguintes exemplos:

> *Vietname nunca mais.*
>
> *A tosta mista queixou-se.* (dito por um criado ao cozinheiro do mesmo restaurante, e referindo-se, naturalmente, ao cliente que encomendou a tosta mista).

Qualquer pessoa que se debruce sobre a interpretação de nomes próprios em texto defronta-se com estes casos, muito comuns em textos jornalísticos. Markert e Nissim (2002) listam um número apreciável de padrões metonímicos associados a lugar ("place-for-event", "place-for-people", "place-for-product", etc.), assim como critérios detalhados para classificar "nomes de locais" nos vários padrões. Além disso, propõem uma organização hierárquica das metonímias, a existência de uma categoria "mista"[1] e o (implícito) reconhecimento de vagueza. Também Leveling e os seus colegas (Leveling e Veiel, 2006; Leveling e Hartrumpf, 2006) estudam a metonímia em termos de recolha de informação geográfica (RIG) em textos jornalísticos em alemão e concluem que, se retirarem os casos em que os "locais" são usados metonimicamente, obtêm resultados melhores no GeoCLEF[2]. Ou seja,

[1] Para tratar de casos como o seguinte exemplo (inventado), em que *Moçambique* aparece como local para a primeira oração, e como governo para a segunda: *quando chegou a Moçambique, que até essa altura se tinha mostrado contra as sanções, recebeu a desagradável notícia de que...*

[2] O GeoCLEF é uma avaliação conjunta de recolha de informação geográfica, integrada no CLEF (Rocha e Santos, 2007), e que desde 2005 inclui também o português (Gey et al., 2007; Mandl et al., 2007).

se o sistema só considerar casos em que os locais são mesmo locais, obtém maior precisão nos resultados relativos a tópicos geográficos.

Note-se que também os organizadores do MUC estavam conscientes deste fenómeno, embora as directivas do MUC ditassem que a marcação devia ser feita de acordo com a categoria original. Por outras palavras, o facto de um nome próprio em inglês ser usado numa construção metonímica "place-for-product" não impedia que fosse classificado como `LOCATION`. (Para sermos totalmente precisos, Chinchor e Marsh (1998, secção A.1.6) discute de facto os casos de metonímia, dividindo-os entre "proper" e "common", e trata-os diferentemente, mas sem explicar a razão.)

No HAREM optámos precisamente pela abordagem oposta: em casos de "place-for-product", o nome próprio seria marcado como `PRODUCT`, e não como `LOCATION`.

Mais do que isso, e dada a nossa aderência ao modelo da vagueza como propriedade fundamental da língua, o modelo semântico que abraçámos não recorre a metonímia, mas sim a vagueza de nomes próprios, que podem (e costumam) ter mais de uma interpretação associada.

Veja-se o caso mais comummente discutido, o nome de um país. Na minha opinião, o conjunto de interpretações "pessoas/povo, governo administrativo, local, e cultura/história" em relação a um nome de país fora de contexto é indissociável.

Ou seja, os usos mais prototípicos de *país* incluem ou podem incluir todas estas vertentes. É certo que, em alguns casos, um país/nação pode não ter um local, ou não ter um governo reconhecido, ou não ter mesmo ainda uma cultura/história.[3] Por outro lado, o facto de conter estas quatro (e mais) vertentes no seu significado não quer dizer que o seu nome não possa ser usado apenas numa vertente (em particular apenas como lugar, mas não especialmente como lugar), como ilustram os seguintes exemplos:

Portugal orgulha-se dos descobrimentos. (história/cultura)[4]
Portugal tem um clima ameno (local, geografia física)
Portugal tem uma taxa de natalidade baixa (povo)
Portugal decretou o feriado do Carnaval. (governo, administração, geografia política)

A diferença fundamental entre a abordagem de Markert e Nissim e de Leveling, por um lado, e a abordagem do HAREM, por outro, é que a primeira considera que primariamente países ou cidades são locais, e só por um processo mais complicado (metonímia) deixam a sua interpretação "básica" e passam a exprimir outras coisas não básicas, enquanto que a segunda abordagem, seguida no HAREM (assim como por outras correntes de semântica

[3] Só não consigo imaginar um país deserto, ou seja, que nunca tenha tido pessoas.
[4] Agradeço ao Nuno Cardoso o exemplo mais garantidamente histórico/cultural de: *A influência de Portugal foi grande no Japão.* Mantive, contudo, o exemplo original por causa da argumentação que se segue, à qual convém o uso de *Portugal* como sujeito.

mencionadas acima, muito particularmente a de Pustejowsky[5]) não privilegia a interpretação local em relação às demais interpretações. Como argumento para não privilegiar a vertente lugar, note-se que todos os casos mencionados acima podem ser anaforicamente relacionados usando a palavra país, mas não usando a palavra *local* ou *lugar* (só o segundo):

Portugal é um país com tradições (ou *é um país que se orgulha dos seus Descobrimentos*)
Portugal é um país de clima ameno.
Portugal é um país com taxa de natalidade abaixo de...
Portugal foi o único país da EU que decretou feriado na terça feira passada.

A perfeita aceitabilidade de *Portugal, local de sonho para muitos turistas, orgulha-se dos seus Descobrimentos* foi apresentada por Cristina Mota (c.p.) como uma prova de que a palavra *Portugal*, mesmo noutras acepções/vertentes, pode ser identificado como *local*. Eu discordo. Para mim, o autor da frase apenas está a ligar duas vertentes num argumento que se espera coerente, e não a referir-se à segunda vertente como `LOCAL`.

Voltando ao REM, o HAREM requer uma distinção entre `ABSTRACCAO`, `LOCAL`, `PESSOA` (povo), `PESSOA` (governo), ou mais do que uma vertente simultaneamente, ao contrário do MUC, que classificaria todos os casos acima como `LOCATION`.

No modelo semântico subjacente ao HAREM, portanto, a palavra *Portugal* não significa imediatamente um lugar. O contexto no qual o nome *Portugal* se insere é vital para seleccionar a vertente da palavra. Além disso, e aqui está a importância da vagueza para o modelo, pode muitas vezes significar mais do que uma única vertente. Se apenas classificássemos *Portugal* como `País`, que é uma alternativa por vezes sugerida (que será debatida mais abaixo), ficava muito por compreender. E se classificássemos `País` como (apenas) `Lugar` (como se fez no MUC), estávamos a deitar fora mais de metade do significado de *Portugal*.

4.2.2 REM como aplicação prática

Ninguém discorda que, para determinar a vertente semântica em que é empregue qualquer expressão, é preciso compreender o texto em questão, e que haverá diferentes casos em que um utilizador estará interessado em diferentes vertentes de um mesmo conceito (por exemplo, política portuguesa contemporânea vs. aspectos da natureza em Portugal).

Poucos compreendem, contudo, que isso significa que ao nome *Portugal* não pode então ser associada sempre a mesma classificação se se quer distinguir entre as várias vertentes.

[5] Pustejovsky (1995) sugere um mecanismo complicado de formalização semântica, estruturado em quatro eixos/estruturas (argumental, de acontecimentos, de modos de explicação (*qualia*), e de herança lexical), separando além disso o que ele chama tipos unificados (*unified types*) e tipos complexos (*complex types*). Pustejovsky (1995, p. 141–157) analisa, por exemplo, *book* e *newspaper* como tipos complexos `informação.matéria-impressa`, sendo além disso *newspaper* um tipo complexo (`informação.matéria-impressa`)`.organização`. Conforme o contexto, um dado texto pode referir-se a modos particulares de explicação (os quatro que ele considera são forma, propósito, constituição e criação), ou a mais do que um desses modos.

Ou seja, se o REM pretende ajudar os sistemas e as pessoas a distinguir entre diferentes significados, é preciso que estejam separados e não aglomerados.

Questões a que o modelo semântico do HAREM (com a correspondente criação da colecção dourada) permite responder é, por exemplo, quantas vezes é que a palavra *França* foi utilizada na acepção `LOCAL` – ao contrário da pergunta, a que o MUC responde, de quantas vezes é que *França* (ou *France*) foi usada como um país (assumindo que os países são considerados `LOCATION` no MUC).

Em ambos os casos, não se está a entrar em conta com *França* quando classifica pessoas, ou organizações (fábrica de roupa, ou de sapatos), claro. Por isso é que ambas as tarefas são correctamente compreendidas como análise semântica dos textos, visto que requerem mais do que uma classificação de acordo com um dicionário de nomes próprios ou almanaque. Note-se contudo que se for a pastelaria França a referida na frase *Encontramo-nos hoje às duas na França*, neste caso *França* seria classificada como um `LOCAL` no HAREM, e como `ORGANIZATION` no MUC.

4.2.3 REM como classificação semântica tradicional

É muitas vezes apresentado como alternativa ao REM, ou como outro modelo de REM, a classificação directa dos nomes próprios nas classes mais descritivas que os compõem, tal como país, artista, político, monumento, jornal, para evitar escolher em que vertente cada uma destas classes deverá ser colocada. Ou seja, livro é um objecto ou uma obra de arte? Jornal é uma organização, um porta-voz, ou um papel? País é um lugar, um povo, ou um conceito? Não interessa, dizem os defensores deste modelo, o que interessa é ter classificado um nome próprio como `Livro`, ou `Jornal`, ou `País`.

Isto na minha opinião é simplesmente escamotear o problema. Primeiro, porque acaba por não se atribuir uma classificação segundo uma grelha pré-determinada mutuamente exclusiva (como é o caso da divisão do MUC entre `LOCATION/PERSON/ORGANIZATION` ou da categorização do HAREM com 10 grupos). O REM deixa assim de ser um problema especificável *a priori*, porque em princípio há um número infinito de classes a que cada expressão pode ser atribuída. E ainda há outra objecção importante, relacionada outra vez com a vagueza essencial da língua, que é mais facilmente compreendida por um exemplo. Atentemos nas seguintes frases:

> *Património de Sintra ameaçado por construção selvagem*
> *Freixo de Espada à Cinta atrai turismo com festival de música*
> *Douro com problemas de poluição*

Todos os "lugares" com nome podem ser empregues para denotar um conjunto de pessoas, uma cultura, etc., mas exactamente que tipo de lugar (ou entidade) referida não é geralmente tornado explícito na comunicação, porque não é necessário. Nas frases acima,

Sintra refere-se a concelho, a vila, ou a serra? *Freixo de Espada à Cinta* descreve a cidade ou a região? E *Douro* é o rio, a região, ou a população ribeirinha?

Ou seja, não é claro que classificações semânticas se devem atribuir a estas entidades mencionadas (`Concelho`, `Vila`, `Serra`, `Cidade`, `Região`, etc.?), bem como continua a não ser óbvia qual a acepção (ou vertente) em que elas são usadas nos contextos dados (`Rio` não parece poder nunca englobar a população ribeirinha desse mesmo rio, embora para `País` isso pareça ser aceitável).

Isto demonstra que a opção de classificar as EM segundo os seus tipos semânticos imediatos (`País`, `Rio`, `Cidade`, etc.) causaria mais problemas do que os que resolveria.

Na minha opinião, a maior objecção a este modelo é que, em muitos casos, senão na sua esmagadora maioria, o falante não quer decidir se se está a referir à cidade, à serra ou a todo o concelho... quanto mais a pessoa que recebe a informação e não sabe o que passa (ou passou) na mente do falante. *Sintra*, na maior parte das vezes, é vaga entre as três interpretações "cidade", "população" e "serra".

4.3 O ACE como uma alternativa ao MUC: outras escolhas

Para que fique mencionado, a inspiração do HAREM foi o MUC. Não nos debruçámos na altura suficientemente sobre o ACE (Doddington et al., 2004), convencidos de que representava um estádio mais elevado, demasiado complexo para principiantes na tarefa do REM.

Agora, estou convencida de que foi um erro grave não termos estudado aturadamente o processo seguido no ACE, pois parece que, em paralelo, chegámos independentemente a muitas conclusões semelhantes, embora também enveredado por caminhos diferentes.

Comecemos por salientar que a questão da metonímia (ou várias vertentes de, principalmente, nomes de lugares) foi resolvida no ACE através da introdução da categoria "locais geopolíticos" (para países ou cidades que são comummente mencionadas como actores políticos). Esta é uma forma um pouco original de lidar com a questão da vagueza na língua, mas apenas neste caso particular (criando a categoria `LOCAL+ORG`, que pode além disso ser especializada através da escolha de uma das possibilidades).

Segundo a interpretação de Maynard et al. (2003a), repetida em Cunningham (2005), o ACE teve a intenção de melhorar o processo seguido pelo MUC de uma forma semelhante ao HAREM: nas palavras de Maynard, em vez de análise "linguística", tentaram uma análise "semântica": *where the MUC task dealt with the **linguistic** analysis, ACE deals with its **semantic** analysis*. Nas palavras do ACE (Doddington et al., 2004, p. 838), este está interessado no reconhecimento de "entidades, não apenas nomes" (*entities, not just names*). Pese embora a imprecisão desta terminologia (que opõe linguístico a semântico), o que eles querem dizer é que o MUC partiu da forma, e o ACE do conteúdo (denotação). Algo surpreendentemente, até mencionam a versão inglesa da nossa terminologia: *In ACE these*

names are viewed as ***mentions*** *of the underlying entities.* Não podíamos ter confirmação mais evidente para a nossa escolha de nome em português, nem demonstração mais óbvia de que o HAREM e o ACE identificaram o mesmo problema no MUC. Contudo, abordaram-no de uma forma diametralmente oposta.

O problema do MUC, que refinamos aqui, é que partia de uma definição arbitrária com base nos dois campos ligados pela semântica (a língua/forma, e o conteúdo/denotação), delimitada por um subconjunto deste último: a tarefa do MUC tinha como alvo nomes próprios (forma) com significado de organização, local, etc. (denotação), como se aprecia nas palavras de Chinchor e Marsh (1998): «*the expressions to be annotated are "unique identifiers" of entities (organizations, persons, locations), times [...] and quantities [...] The task requires that the system recognize what a string represents, not just its superficial appearance*».

O ACE escolheu o **lado do conteúdo** e pediu para -– independentemente da forma — os sistemas marcarem tudo o que fosse organização, local, pessoa, etc., sem restrições de forma (podiam ser realizados linguisticamente como substantivo, pronome, nome próprio, sintagma nominal, etc.).

O HAREM, ao contrário, **escolheu o lado da forma**: partiu de tudo o que é nome próprio em português (ver capítulo 16) e pediu para os sistemas identificarem e classificarem – sem restrições de sentido numa primeira fase, mas, depois de um estudo empírico inicial – com base na classificação proposta pela organização. (Note-se, no entanto, que aceitamos uma categoria `OUTRO`, ou seja, não garantimos que todas e quaisquer ocorrências de nomes próprios no texto podem ser enquadrados no produto cartesiano das categorias do HAREM.)

A parecença entre as duas extensões ao MUC (ambas reconhecem o MUC como inspiração) é também visível no aumento da variedade em tipo de textos: em vez de alargar em género como fizemos no HAREM, contudo, o ACE alargou em qualidade de texto ou meio de obtenção desse texto. Além de notícias impressas, usou textos obtidos a partir de reconhecimento óptico, e de reconhecimento automático de fala. Também alargou o assunto (em vez de um único domínio, passou a ter notícias sobre vários domínios ou assuntos). Interessante que, no caso do HAREM, usámos a extensão em termos de variante e sobretudo de estilo/género textual, alargando em termos de meio ou de qualidade apenas quando tal derivava de um género textual diferente: em particular, para cobrirmos a Web, tivemos de incluir textos de pouca qualidade, e para incluir entrevistas, tivemos de recorrer à transcrição da linguagem oral.

Outra semelhança entre o ACE e o HAREM foi o aumento significativo da complexidade na anotação humana, que, de acordo com (Maynard et al., 2003a), atingiu apenas 82,2% de consenso no ACE.

Outra diferença em relação ao MUC partilhada pelo HAREM e pelo ACE é a utilização neste último duma métrica baseada em custo (Maynard et al., 2002), que, embora mais geral do que a do HAREM, tem pontos de semelhança com a medida da classificação se-

mântica combinada do HAREM, permitindo a quantificação de uma dificuldade *a priori*.

Contudo, há diferenças entre o ACE e o HAREM, que nos impedem de rotular este de "o ACE português", mesmo de forma aproximada.

Em primeiro lugar, o ACE mistura a tarefa de reconhecimento de EM com a de reconhecimento de co-referências, o que significa que a forma de avaliar a identificação e/ou classificação é diferente. Desse ponto de vista, o HAREM emparelha com o MUC, ao separar (e no caso do HAREM, ignorar) a tarefa de co-referência da da identificação.

Mas a distinção mais importante é filosófica mesmo, e está relacionada com o tema principal do presente texto: o ACE exige uma única resposta correcta (através da possível criação de categorias vagas, tal como as entidades geopolíticas ou as instalações), enquanto no HAREM estamos interessados, não numa cristalização oficial dessas categorias, mas na detecção empírica de todas as perspectivas possíveis oferecidas pela língua. Ou seja, em vez de resolver o problema da vagueza do lado da organização com categorias fixas codificando essa vagueza (ou tipos complexos, na terminologia de Pustejowsky) aceitámos *a priori* qualquer conjunto de categorias como sendo representável pelo HAREM e que os anotadores decidiram atribuir como tal no contexto.

Para sermos completamente justos, convém realçar, mais uma vez agradecendo à Cristina Mota por nos ter tornado cientes desse facto, que o ACE permite, opcionalmente, a marcação da vertente (local, pessoa, organização) para as entidades geopolíticas[6]. Embora isso seja uma forma de resolver (para um conjunto limitado) a questão das múltiplas vertentes, parece-nos que a diferença é maior que a semelhança: por um lado, no HAREM não é só a categoria `<LOCAL | ORGANIZACAO>` que pode ser vaga, mas todas; por outro, quando uma expressão é só `LOCAL`, deve ser marcada como tal no HAREM, e não duplamente como "`<LOCAL | ORGANIZACAO>` vertente `LOCAL`", como no ACE.

4.4 A abordagem do HAREM como processamento da linguagem natural em geral

Um modelo conceptual ingénuo de um reconhecedor de EM é concebê-lo como um sistema com listas de nomes próprios previamente classificados (um almanaque) que atribui essa classificação quando os nomes se encontram no texto. E, de forma igualmente ingénua, se pode conceber que é esse o papel de um dicionário na análise sintáctica computacional.

De facto, dado o peso e relevância do contexto, não é preciso que as mesmas categorias se encontrem em ambos os lados da análise (ou seja, tanto no dicionário como no resultado da análise (sintáctica) de texto, ou tanto no almanaque como no resultado da análise semântica do texto), embora tenha de haver uma maneira de se fazer a ponte.

[6] Entidades semelhantes, tais como o marcador `<civ>`, são chamadas *híbridas* por Bick no capítulo 12. No HAREM são simplesmente codificadas através do operador |, ou seja `<LOCAL | ORGANIZACAO>`.

Concretizando: num almanaque, faz sentido que esteja armazenada a informação de que *França é um país*, mas na gramática de REM daria jeito que estivesse "um país pode ser um local, um povo, etc...". Tal qual como num dicionário pode estar que "*perfeito* é um adjectivo", mas na gramática terá de estar (ou seria desejável que estivesse) "um adjectivo pode ser usado como um substantivo, como um pós-nominal, como um pré-nominal, como uma parte de um composto, como uma exclamação, etc.", de forma a compreender ocorrências como, respectivamente, *o perfeito é inacessível, um perfeito disparate, um casal perfeito, amor-perfeito, Perfeito!*, etc. (vejam-se mais exemplos em Santos, 2006a).

Estamos pois a defender neste capítulo implicitamente um nível intermédio de processamento, ou melhor, uma forma de fazer PLN mais dirigida pelo contexto e menos pelo léxico. Em última análise, o desenvolvimento do sistema de REM fica ao critério do seu autor e dos seus objectivos, e muitos investigadores provavelmente quererão codificar "tudo" num dicionário ou num almanaque. No entanto, é importante salientar que estas duas abordagens são possíveis e que, a nível teórico pelo menos, uma não tem prioridade sobre a outra.

Note-se também que toda a discussão neste capítulo até agora tem sido sobre vagueza, ou seja a possibilidade de diversas interpretações simultaneamente. Outro assunto diferente é a ambiguidade, que talvez convenha também mencionar, para prevenir mal-entendidos em relação ao que até aqui se expôs.

Um caso claro de ambiguidade em REM é o seguinte: *Washington* representando o governo americano (ou o conjunto das pessoas correntemente fazendo parte do governo americano) e *Washington* representando o primeiro presidente dos Estados Unidos. Embora ambas sejam classificadas como PESSOA no HAREM: nome de uma cidade (capital) como menção a um grupo de pessoas pertencendo a entidade governativa[7] (<PESSOA TIPO="GRUPOIND">) e nome de uma pessoa que deu (por acaso) origem ao nome dessa mesma cidade (<PESSOA TIPO="INDIVIDUAL">), deve ser claro que qualquer texto em contexto ou se refere a um ou a outro.[8]

Igualmente no caso do adjectivo acima exemplificado, o facto de haver um substantivo ou adjectivo com o sentido de "categoria verbal" implica que *perfeito* é ambíguo, mas que, em qualquer contexto, ou significa uma ou outra das acepções.

É pois sempre preciso resolver a ambiguidade (isto é, escolher uma das várias opções mutuamente exclusivas), o que é uma tarefa completamente diferente de lidar como deve ser com a vagueza, que significa preservar vários sentidos relacionados.

[7] Veja-se o capítulo 16, para a distinção entre *Washington* como governo (<ORGANIZACAO TIPO="ADMINISTRATIVO">) e como grupo de pessoas pertencentes ao governo (<PESSOA TIPO="GRUPOIND">).

[8] Convém contudo notar que a medida de classificação semântica por categorias (capítulo 18) não permite entrar em conta com o facto de que as duas interpretações de *Washington* como PESSOA correspondem a dois sentidos diferentes. Um sistema que tivesse feito a escolha errada (entre as duas PESSOAs) não seria por isso penalizado no HAREM.

4.5 Alguma discussão em torno do modelo de REM do Primeiro HAREM

Um dos argumentos apresentado contra o modelo utilizado no HAREM, expressamente vocalizado durante a sessão plenária do Encontro do HAREM e também já presente em Bick (2006a,b), é a questão da relação entre o significado "intrínseco" (aquele que aparece num dicionário, sem contexto) de um nome próprio, e o papel que esse nome próprio desempenha em contexto. Segundo Eckhard Bick, ambos são necessários e devem ser marcados, mas a ligação com a realidade (se é rio, se é cidade, se é país) está no dicionário, e o resto provém da interpretação sintáctico-semântica, em termos mais gerais, na forma de papéis semânticos como os propostos originalmente por Fillmore (1968) (agente, paciente, direcção, instrumento, etc.). Nesta perspectiva, a conjugação dos dois tipos de informação permite inferir o que estamos a anotar no HAREM: `País` + `Agente` = `Governo` (ou `Povo`? ou `Equipa`); `País` + `Instrumento` = `Governo`; etc.[9]

É preciso, contudo, confirmar na prática se de facto se consegue: i) definir consensualmente os papéis semânticos necessários (algo que até agora não parece ter sido possível) e aplicá-los a texto real de forma satisfatória; e ii) definir uma álgebra que dê de facto as mesmas (ou mais satisfatórias) distinções do que as empregues no HAREM.

Admitindo que tal seja possível, ou seja, que usar uma classificação composta por um papel semântico genérico mais um conjunto de marcações específicas no léxico consegue produzir o mesmo que o HAREM procurou atingir, penso que tal funcionará mais como uma demonstração de que o nosso modelo de interpretação dos nomes próprios no HAREM é apropriado, do que como uma crítica ao nosso objectivo. Parece-me que esta posição — que é baseada num modelo de como fazer REM — acaba por redundar em mais um argumento a favor da anotação usada no HAREM para avaliar o REM.

4.6 Outros trabalhos

Uma distinção semelhante fora feita já por Johannessen et al. (2005) no âmbito de uma comparação entre vários sistemas para línguas nórdicas. Estes autores discutem a definição da tarefa de REM, identificando duas estratégias distintas, que baptizam como "forma com prioridade sobre a função" e "função com prioridade sobre a forma"[10]. A primeira estratégia pretende identificar formas com uma dada denotação, independentemente da sua função em contexto; a segunda pretende identificar um conjunto de funções com base no contexto. Talvez devido ao grande número de autores, a conclusão do artigo é de "indecisão quanto à estratégia preferível" (p. 97). Mais interessante é a afirmação de que os sistemas que dão prioridade à função são mais robustos em relação à diminuição drástica do tamanho dos almanaques. Não fica, contudo, claro como é que os autores podem

[9] Por `Governo` estamos aqui a abreviar a notação do HAREM correcta, que seria `<ORGANIZACAO TIPO="ADMINISTRACAO">`.
[10] Em inglês, *form over function* e *function over form*.

comparar os sistemas com uma mesma avaliação se de facto a tarefa a que se propõem é diferente nos dois casos.

4.7 Comentários finais

Este capítulo tentou apresentar o modelo semântico pressuposto pelo Primeiro HAREM, quer através da aplicação básica de conceitos semânticos genéricos ao REM, quer através de uma comparação detalhada com os modelos respectivos do MUC e do ACE.

Os dois pressupostos mais importantes dizem respeito à importância do contexto na interpretação, e à ubiquidade da vagueza na linguagem natural.

Contudo, o capítulo é profundamente teórico no sentido de não fornecer dados empíricos sobre a extensão das diferenças entre os modelos apresentados, e sugere imediatamente algumas tarefas que propomos também no capítulo 7.

Com efeito, seria muito interessante anotar a colecção dourada do HAREM com uma marcação estilo MUC (cujas directivas para o português ainda estão contudo por fazer a um nível de detalhe suficiente) e depois medir e analisar objectivamente os resultados: quantas vezes é que a diferença entre os modelos implicaria diferença a nível da classificação final?

Outra medição é a da dificuldade da tarefa proposta pelo HAREM, quer a nível de concordância entre anotadores, quer a nível de dispersão intracategorial e intercategorial dos nomes próprios em português. É preciso quantificar quantas EM são ambíguas e/ou vagas tanto em potência (no almanaque ideal) como na realidade (aproximada), em texto. Se fizermos uma nova anotação no estilo MUC, poderemos ter ainda outra medida da diferença entre a dificuldade das duas tarefas.

Um outro trabalho natural como continuação do HAREM é comparar os resultados obtidos por Markert e Nissim (2002) para o inglês analisando 2000 EM, com os resultados das colecções douradas do HAREM (mais de 9000 EM), investigados sob a perspectiva da metonímia.

Finalmente, talvez a questão mais interessante para uma teorização mais rigorosa do REM será investigar a redutibilidade de um problema ao outro. Será que do "tipo MUC" mais papel semântico se pode derivar o "tipo HAREM"? Será que do "tipo HAREM" mais o papel semântico de uma população poder-se-á inferir o "tipo MUC"?

Esperamos poder um dia vir a responder a estas perguntas, com a ajuda da comunidade reunida em torno do Segundo HAREM, visto que a criação de recursos dourados e a sua disponibilização não deve morrer com a comparação na primeira avaliação conjunta, mas sim produzir matéria prima para muitos estudos empíricos e mesmo futuras avaliações.

Agradecimentos

Este capítulo deve um número apreciável de agradecimentos: em primeiro lugar, a todos os presentes no Encontro do HAREM no Porto pelo interesse e entusiasmo dos debates, que tiveram uma influência decisiva na concepção do presente texto; em segundo lugar, aos meus colegas na organização do HAREM sem a qual não estaríamos aqui, muito em particular ao Nuno Cardoso com quem revi em conjunto toda a colecção dourada e, como tal, partilhei muitas horas dedicadas à compreensão das EM em texto em português.

Além de um agradecimento natural a todos os participantes no HAREM, é forçoso salientar, admirar e agradecer a postura do Eckhard Bick e da Cristina Mota, que participaram segundo as normas do HAREM apesar de, desde o início, terem discordado dessas normas no que se refere precisamente ao modelo semântico utilizado.

No que se refere ao presente texto, tenho de agradecer especialmente a revisão cuidada e as muitas sugestões de melhoria da Cristina Mota, do Nuno Cardoso e do Jorge Baptista em relação a versões anteriores, que levaram a uma reescrita quase completa do capítulo. Gostava também de mencionar o entusiasmo genuíno e sempre carente de mais fundamentação que foi exibido pelo Nuno Seco em relação ao modelo do HAREM, quando chegou, mais tardiamente, à organização do mesmo, como aliás é patente no capítulo anterior. Ele foi assim, embora inconscientemente, um dos inspiradores do presente texto.

Finalmente, este capítulo foi escrito integralmente no âmbito da Linguateca, financiada pela Fundação para a Ciência e Tecnologia através do projecto POSI/PLP/43931/2001, co-financiado pelo POSI, e pelo projecto POSC 339/1.3/C/NAC.