VARRA: um serviço para a Validação, Avaliação e Revisão de Relações semânticas no AC/DC

Cláudia Freitas (Linguateca-FCCN / PUC-Rio)
Diana Santos (Universidade de Oslo / Linguateca-FCCN)
Hugo Gonçalo Oliveira (CISUC-Universidade de Coimbra)
Violeta Quental (PUC-Rio)

Resumo:
A importância dos recursos lexicais – léxicos, ontologias, tesauros - para sistemas que lidam com o processamento computacional da língua é cada vez mais reconhecida, assim como as dificuldades inerentes a sua elaboração. Metodologias baseadas na extração automática de relações semânticas entre palavras a partir de corpos têm, na forma de avaliação/validação das relações, o seu ponto fraco.
Neste contexto, apresentamos o VARRA, um sistema desenvolvido com o objetivo principal de auxiliar a avaliação (ou validação) manual de relações semânticas entre pares de palavras. Com o VARRA, buscamos não apenas construir uma base confiável de julgamentos sobre uma dada relação, mas também tornar a tarefa de validação de relações entre palavras mais parecida com a interpretação humana (em oposição à validação de relações fora de contexto). Por isso, no VARRA, as palavras que participam de uma relação semântica são sempre consideradas em contextos autênticos, representados por frases de corpos do projeto AC/DC. No presente artigo, além de uma descrição do sistema, relatamos também os resultados iniciais obtidos com a validação de relações semânticas tendo em vista a avaliação e melhoria do PAPEL, uma ontologia lexical pública para o português, extraída de forma automática a partir de um dicionário.

Palavras-chave: relações semânticas; avaliação de relações semânticas; semântica computacional; ontologia lexical.

Abstract:
Lexical resources – ontologies, thesauri, lexicons - have been widely recognized as key components to natural language processing applications. However, such resources, especially when developed through automatic means, require a thorough evaluation to be accepted, and lack of evaluation is arguably a major weakness of many resources.
In the specific context of lexical ontologies for Portuguese, we present VARRA, a system designed to assist the manual evaluation of semantic relations between word pairs. This system presents the human evaluator with a set of real contexts from corpora, and s/he is required to assess whether those contexts confirm or not the putative semantic relation, instead of (only) evaluating the triple word RELATIONSHIP word. This way, we bring the validation task closer to the human process of understanding sentences.
In this paper, we present the preliminary results of the evaluation of PAPEL, a freely available lexical ontology for Portuguese, created semi-automatically from a general Portuguese dictionary, done in a pedagogical context with several different users. We present briefly the kinds of relationships present in PAPEL, the rationale for VARRA, and the measures put forward to take stock of the answers.

VARRA is available on the Web as a service for anyone interested in semantic relations in Portuguese, and we provide the raw data for others to use it as well.

Keywords: semantic relations; computational semantics; lexical ontology; evaluation of lexical resources.

1. Introdução

Cada vez mais é reconhecida a necessidade do uso de informação semântica para efetuar tarefas de processamento computacional da língua, e a construção e manutenção de léxicos computacionais têm assumido importância crescente. Para que a contribuição desses recursos seja efetiva, é preciso que cubram uma vasta porção da língua, como o fazem as ontologias lexicais ou, antes de esse termo ser cunhado, bases de dados – lexicais e de conhecimento sobre o mundo. (Para uma discussão da diferença e relação entre ontologias e bases de dados lexicais, veja-se Hirst (2004)).

A elaboração de tais recursos pode ser feita de duas maneiras: (i) por seres humanos, o que, se por um lado garante uma maior precisão, por outro, depende de uma vastíssima mão de obra, além de ser um trabalho moroso e custoso; ou (ii) automática ou semi-automaticamente, ancorando o conhecimento a ser obtido em vastas coleções de textos (corpos) ou dicionários.

Embora a metodologia automática ou semi-automática apresente a vantagem de minimizar o custo de elaboração (quanto ao tempo e à mão de obra), tem como principal ponto fraco a dificuldade de garantir a sua qualidade, e precisa, portanto, de um processo de avaliação muito mais rigoroso.

De maneira geral, podemos distinguir duas abordagens para a avaliação de recursos lexicais: a comparação com um “modelo ideal” (o que, naturalmente, pressupõe a existência de um “modelo ideal” nos moldes do que se quer construir, e que portanto só pode ser aplicada a poucos casos); e a avaliação humana. A avaliação humana é a maneira mais frequentemente escolhida para avaliar a qualidade de um recurso construído de forma automática, como demonstrado entre outros por Riloff e Shepherd (1997), Caraballo (1999) e Richardson, Vanderwende e Dolan (1993) — e não duvidamos de que é a forma mais confiável, ainda que, como tarefa subjetiva, possa variar entre os sujeitos da avaliação.

Nesse contexto, apresentamos o VARRA, um sistema desenvolvido com o objetivo principal de auxiliar a avaliação (ou validação) manual de relações semânticas entre pares de palavras (que, por sua vez, são o cerne de um recurso lexical como os já mencionados).

Com o VARRA, buscamos não apenas construir uma base confiável de julgamentos sobre uma dada relação, mas também tornar a tarefa de validação de relações entre palavras mais parecida com a interpretação humana (em oposição à validação de relações fora de contexto). Por isso, no VARRA, as palavras que participam de uma relação semântica são sempre consideradas em contextos autênticos, representados por frases de corpos do projeto AC/DC (Costa, Santos e Rocha,

2009, Santos, 2010), uma interface comum para acesso e disponibilização de corpos em português, que contém atualmente 22 corpos em português, das variantes brasileira e de Portugal, e cerca de 374 milhões de palavras. Assim, ao invés de um avaliador perguntar "à sua intuição" se, por exemplo, "mentira é sinônimo de ilusão", ele deve, ao olhar para um contexto/frase selecionado automaticamente de um corpo, julgar se aquela frase ilustra a relação de sinonímia entre "mentira" e "ilusão".

Paralelamente ao auxílio na validação de relações, o VARRA também foi criado para ajudar na descoberta e teste de expressões da língua capazes de veicular relações semânticas entre palavras. No entanto, como esta funcionalidade ainda está em desenvolvimento, nos deteremos aqui no que consideramos o aspecto principal do sistema — o auxílio na validação de relações semânticas — tomando como exemplo sua aplicação à melhoria do PAPEL (Gonçalo Oliveira, Santos & Gomes, 2010), uma rede lexical pública para o português cujas relações semânticas foram automaticamente obtidas a partir de um dicionário.

É importante mencionar que, embora tenha sido inicialmente desenvolvido para validar as relações semânticas do PAPEL, o VARRA não está diretamente vinculado a qualquer ontologia lexical, léxico, tesauro ou dicionário. O sistema pode – e deve – ser utilizado para verificar relações semânticas advindas de quaisquer recursos, como, por exemplo, os descritos em Santos et al. (2010).

Uma última nota quanto ao posicionamento teórico. Embora estejamos alinhados com uma perspectiva segundo a qual o significado das palavras é percebido quando estas estão inseridas em contextos de uso – e é justamente essa percepção que norteou o desenvolvimento do VARRA – buscamos não nos comprometer com modelos semânticos específicos, a fim de tornar, na medida do possível, o VARRA uma ferramenta "ateórica". Não nos furtamos, no entanto, ao uso de uma metalinguagem tradicional (usamos termos como hiperonímia, sinonímia), para tratarmos das relações semânticas extraídas.

Por fim, o VARRA é uma colaboração entre a Linguateca, o CISUC (Centro de Informática e Sistemas da Universidade de Coimbra) e o Departamento de Letras da PUC-Rio e, devido a esse perfil "multidisciplinar", o seu desenvolvimento teve em consideração, também, a possibilidade de utilização para o estudo da própria língua – tanto no aspecto de pesquisa (algumas explorações linguísticas com o VARRA estão na seção 5), quanto no aspecto didático (tratado na seção 6).

A seguir, antes de passarmos a um detalhamento do VARRA e dos resultados obtidos quanto à avaliação e melhoria do PAPEL, apresentamos a motivação para a elaboração deste recurso, bem como uma descrição do seu conteúdo.

2. Apresentação do PAPEL

O PAPEL – Palavras Associadas Porto Editora-Linguateca – é uma rede lexical pública para o português, extraída de forma automática a partir do Dicionário PRO da Língua Portuguesa da Porto Editora (DPLP), acessível a partir de http://www.linguateca.pt/PAPEL. A sua construção e desenho foram inspirados no MindNet (Richardson, 1997, Richardson et al., 1998), como explicado em pormenor em Gonçalo Oliveira, Santos & Gomes (2010).

2.1 Por que criar o PAPEL a partir de um dicionário de língua geral?

Há várias razões para partir de um dicionário, a primeira delas por serem os dicionários repositórios tradicionais de informação sobre o léxico de uma língua, fruto do trabalho acurado de vários lexicógrafos/linguistas, e que, portanto, pressupomos que se encontra a meio caminho entre texto livre (como seria o caso nos corpos) e uma formalização em termos de relações entre palavras. Será/seria, pois, mais fácil atingir uma ontologia lexical a partir de um dicionário do que a partir de textos.

Além disso, um dicionário esforça-se por cobrir a língua, donde a escolha dos verbetes é trabalho já feito, assim como se esforça por simplificar a sua apresentação (e é conhecido que a sintaxe de um dicionário é muito mais simples do que a de um texto incontrolado).

Há, no entanto, uma outra razão de peso, que não é tão óbvia como as anteriores, mas que, na nossa opinião, é talvez a mais importante de um ponto de vista linguístico: o fato de que os lexicógrafos usam a própria língua para esclarecer o sentido (em vez de um vocabulário formal). Isso implica que é possível descobrir generalidades sobre uma língua que se encontram implícitas num dicionário e que não pertencem ao conjunto de relações tradicionais ou canônicas propostas por lógicos ou por linguistas de outras línguas. Por outras palavras, explorar um dicionário de português pode levar à descoberta de generalizações importantes para o português, não só em termos de categorias lexicais/semânticas, mas também em termos de relações sistemáticas entre itens. Foi por isso que, sempre que nos pareceu observar uma regularidade em várias definições distintas e independentes, tentamos extraí-la, mesmo sem ter um nome ou uma garantia de já ter sido considerada antes por especialistas.

Assim se explica e, do nosso ponto de vista, se justifica, a quantidade de relações não canônicas que extraímos no PAPEL e pusemos à consideração da comunidade. Resta agora verificar ou confirmar que de fato descobrimos alguma coisa de interesse, o que tentamos fazer através do VARRA com uma primeira confirmação.

2.2. Introdução ao conteúdo do PAPEL

O processo de construção do PAPEL (ver também Gonçalo Oliveira et al., 2008) baseia-se num conjunto de gramáticas (regras) que faz uso de determinados padrões léxico-sintáticos para extrair relações semânticas entre o sentido de palavras que ocorrem numa definição (p) e o sentido da palavra definida (v), no formato de triplos (p RELACAO v). As relações que integram o PAPEL foram escolhidas com base na inspeção do conteúdo do dicionário e na revisão da literatura sobre relações entre palavras e forma de estruturar dicionários.

As gramáticas foram construídas tendo por base padrões frequentes no dicionário e que indicavam determinadas relações semânticas – desde as mais comuns, como sinonímia e hiperonímia, até relações menos comuns, que foram extraídas devido igualmente à existência de padrões frequentes.

Se, por um lado, no trabalho baseado em corpos, o mais usual é a extração de relações entre substantivos, no trabalho baseado em um dicionário é igualmente possível extrair relações entre palavras de outras categorias gramaticais, como verbos, adjetivos e advérbios. Assim, além de grandes grupos de relações, foram definidas sub-relações de acordo com a categoria gramatical dos seus argumentos.

A seguir apresentam-se as relações e sub-relações do PAPEL, acompanhadas de exemplos.

SINONÍMIA

Uma relação de sinonímia, p SINONIMO_DE v, indica que, em determinado contexto, p e v podem ter o mesmo significado.

Para que fosse possível identificar as categorias gramaticais das palavras relacionadas por sinonímia, foram definidas quatro sub-relações em função da categoria gramatical envolvida:

substantivo SINONIMO_N_DE substantivo
verbo SINONIMO_V_DE verbo
adjetivo SINONIMO_ADJ_DE adjetivo
advérbio SINONIMO_ADV_DE advérbio

A extração de sinonímia se baseou nas definições constituídas apenas por uma palavra ou por uma enumeração de palavras, como se pode verificar nos seguintes exemplos:

amabilidade, s. f. afabilidade
→ afabilidade SINONIMO_N_DE amabilidade

talhar, v. tr. gravar, cinzelar ou esculpir
→ esculpir SINONIMO_V_DE talhar

moldável, adj. 2 gén. adaptável, flexível
→ flexível SINONIMO_ADJ_DE moldável
→ adaptável SINONIMO_ADJ_DE moldável

sucessivamente, adv. seguidamente
→ seguidamente SINONIMO_ADV_DE sucessivamente

HIPERONÍMIA

Uma relação de hiperonímia, p HIPERONIMO_DE v, indica que um significado de v representa um tipo, gênero ou espécie de um significado de p. No contexto do PAPEL, a relação de hiperonímia foi extraída apenas entre substantivos, tendo como base as expressões *tipo/forma/género de*; palavras (eventualmente modificadas por adjetivos genéricos, tais como *comum*, *raro*) que ocorrem no início das definições (e que constituem o *genus* da definição), antes de padrões indicadores de outras relações, que constituem a *diferentia*; e definições iniciadas por palavras de uma lista de hiperônimos frequentes, tais como *pessoa*, *planta*, *instrumento*, *propriedade*.

Alguns exemplos de hiperonímia:

fardamento, s. m. tipo de farda
→ farda HIPERONIMO_DE fardamento

bioacústica, s. f. ciência que tem por objectivo o estudo dos sons produzidos por animais
→ ciência HIPERONIMO_DE bioacústica

fotojornalismo, s. m. género de jornalismo em que as fotografias constituem o principal material informativo
→ jornalismo HIPERONIMO_DE fotojornalismo

detonação, s. f. ruído causado por explosão
→ ruído HIPERONIMO_DE detonação

esfera armilar, s. f. dispositivo formado por armilas que representam círculos da esfera celeste
→ dispositivo HIPERONIMO_DE esfera_armilar

curvígrafo, s. m. instrumento que traça curvas
→instrumento HIPERONIMO_DE curvígrafo

PARTE

Uma relação PARTE_DE, p PARTE_DE v, indica que p é uma parte ou constituinte de v. No contexto do PAPEL esta relação foi extraída para os seguintes pares de categorias gramaticais:

substantivo PARTE_DE substantivo
substantivo PARTE_DE_ALGO_COM_PROPRIEDADE adjetivo
adjetivo PROPRIEDADE_DE_ALGO_PARTE_DE substantivo

A extração da relação PARTE_DE tomou por base os seguintes padrões, que, apesar de expressarem um refinamento da relação mais ampla "PARTE_DE", foram considerados, no PAPEL, padrões de uma mesma relação:

Indicadores de um todo: *parte de*
Indicadores de constituição: *constituído/formado/composto/provido/munido*
Indicadores de posse: *possui/contém*

Alguns exemplos da relação de PARTE:

citologia, s.f. parte da biologia que estuda as células
→citologia PARTE_DE biologia

cometa, s.m. astro geralmente constituído por núcleo, cabeleira e cauda, …
→núcleo PARTE_DE cometa
→ cabeleira PARTE_DE cometa
→ cauda PARTE_DE cometa

coberto, adj. que possui tampa ou qualquer cobertura
→ tampa PARTE_DE_ALGO_COM_PROPRIEDADE coberto
→ cobertura PARTE_DE_ALGO_COM_PROPRIEDADE coberto

avião, s.m. aparelho de locomoção aérea, munido de asas e de motores para propulsão
→asa PARTE_DE avião
→ motor PARTE_DE avião

deutolécito, s.m. parte do óvulo ou do ovo animal que contém as reservas nutritivas
→deutolécito PARTE_DE óvulo
→deutolécito PARTE_DE ovo
→ reservas PARTE_DE deutolécito

piloso, adj. provido de pêlos
→pêlo PARTE_DE_ALGO_COM_PROPRIEDADE piloso
→piloso PROPRIEDADE_DE_ALGO_PARTE_DE pêlos

MEMBRO

Podem normalmente definir-se vários de tipos da relação de parte de, ou meronímia. Por exemplo, na WordNet de Princeton (Fellbaum, 1998) existem três subtipos desta relação para o inglês: *part-of*, *member-of*, *substance-of*.

No entanto, tal como alguns trabalhos indicam (veja-se por exemplo Itoo & Bouma (2010)), não é simples (ou é mesmo impossível) identificar padrões específicos para cada um destes tipos. Ainda assim, pareceu-nos possível isolar alguns padrões que normalmente expressam a relação membro, p MEMBRO_DE v, que indica que v pode ser um conjunto constituído por uma ou várias instâncias de p. Apesar disto, continua a existir alguma sobreposição entre as relações PARTE_DE e MEMBRO_DE (o que não é de estranhar dada a vagueza inerente à linguagem natural, que milita fortemente contra a compartimentação estanque entre categorias relacionadas).

Tal como no caso de PARTE_DE, foram definidas as seguintes sub-relações para MEMBRO_DE:

substantivo MEMBRO_DE substantivo
substantivoMEMBRO_DE_ALGO_COM_PROPRIEDADE adjetivo
adjetivo PROPRIEDADE_DE_ALGO_MEMBRO_DE substantivo

Esta relação é extraída com base em padrões como os seguintes:

Indicadores de grupo (direto): *membro/elemento de*;
Indicadores de grupo (inverso): *grupo/conjunto/associação/família de*;
Indicadores de inclusão: *inclui/abrange*;
Indicadores de pertença: *pertence/pertencente*

Alguns exemplos da relação de MEMBRO:

director, s.m. membro de uma direcção ou de um directório →director MEMBRO_DE directório	cerimónia, s.f. conjunto de formalidades convencionais usadas na vida social →formalidade MEMBRO_DE cerimónia
celta, s.m. pessoa pertencente aos Celtas →celta MEMBRO_DE Celtas	fadistagem, s.m. grupo de fadistas →fadista MEMBRO_DE fadistagem
centáurea-maior, s.f. planta da família das Compostas, utilizada em Medicina →centáurea-maior MEMBRO_DE Compostas	centro-europeu, adj. pertencente ou relativo ao centro da Europa →centro-europeu PROPRIEDADE_DE_ALGO_MEMBRO_DE centro_da_Europa
multidisciplinar, adj. que abrange várias disciplinas →disciplina MEMBRO_DE_ALGO_COM_PROPRIEDADE multidisciplinar	decretório, adj que inclui decreto →decreto MEMBRO_DE_ALGO_COM_PROPRIEDADE decretório[i]

CAUSADOR

Uma relação de CAUSA, como p CAUSADOR_DE v, ocorre quando p pode causar, provocar ou dar origem a v. Assim, p pode ser, por exemplo, um agente, uma ação, um sintoma, um evento, ou um fenômeno.

Tendo em conta as categorias gramaticais dos seus argumentos, foram definidas as seguintes relações de CAUSA:

substantivo CAUSADOR_DE substantivo
substantivo CAUSADOR_DE_ALGO_COM_PROPRIEDADE adjetivo
adjetivo PROPRIEDADE_DE_ALGO_CAUSADOR_DE substantivo
verbo ACCAO_QUE_CAUSA substantivo
substantivo CAUSADOR_DA_ACCAO verbo

Seguem-se alguns dos padrões utilizados para extrair a relação de CAUSADOR:

Verbos que indicam causa:
No presente: *causa*, *provoca*, *origina*, *suscita.* No particípio passado: *causado por*, *provocado por*, *originado por*, *suscitado por.* No infinitivo: *causar*, *provocar*, *originar*, *suscitar*. Indicadores de causador: *causador*, *provocador*, *causa*, *origem* Indicadores de resultado: *resultado de*, *consequência de.* Outras expressões: *devido a*, *efeito de*.

Alguns exemplos de relações de CAUSA:

detonação, s. f. ruído causado por explosão
→explosão CAUSADOR_DE detonação

friagem, s. f. tempo frio, em geral por causa do vento
→vento CAUSADOR_DE friagem

assadura, s. f. irritação da pele devido a calor ou fricção
→fricção CAUSADOR_DE assadura
→calor CAUSADOR_DE assadura

ópio, s. m. o que causa adormecimento, entorpecimento
→ópio CAUSADOR_DE entorpecimento
→ópio CAUSADOR_DE adormecimento

fumigar, v. tr. desinfectar (local) ou exterminar (parasitas) por acção de fumo ou gases
→fumo CAUSADOR_DA_ACCAO fumigar
→gases CAUSADOR_DA_ACCAO fumigar

dardada, s. f ferimento provocado por golpe de dardo
→golpe_de_dardo CAUSADOR_DE dardada

renzilhar, v. intr. provocar quezílias
→renzilhar ACCAO_QUE_CAUSA quezílias

reactivo, adj. que suscita reacção
→reactivo PROPRIEDADE_DE_ALGO_CAUSADOR_DE reacção

purgação, s. f. acto ou efeito de purgar, limpar ou purificar
→purgar ACCAO_QUE_CAUSA purgação
→purificar ACCAO_QUE_CAUSA purgação
→limpar ACCAO_QUE_CAUSA purgação

prova, s. f. resultado de um ensaio ou teste
→ensaio CAUSADOR_DE prova
→teste CAUSADOR_DE prova

PRODUTOR

A relação PRODUTOR pode ser considerada um sub-tipo da relação de CAUSA mas, mais uma vez, pareceu-nos possível isolar alguns padrões indicadores deste sub-tipo.

Assim, uma relação PRODUTOR, p PRODUTOR_DE v, indica que p, um processo ou uma entidade, pode produzir ou gerar v.

Foram definidas as seguintes sub-relações de PRODUTOR_DE, de acordo com a categoria gramatical dos seus argumentos:

substantivo PRODUTOR_DE substantivo
substantivo PRODUTOR_DE_ALGO_COM_PROPRIEDADE adjetivo
adjetivo PROPRIEDADE_DE_ALGO_PRODUTOR_DE substantivo

Esta relação foi extraída com base nos seguintes indicadores textuais:

Verbos que indicam produção: *produzir*, *gerar*, *obter* (no presente, no infinitivo, ou no particípio passado)
Indicadores de produtor: *produtor de*, *gerador de*
Indicadores de produto: *produto de*, *fruto de*

Alguns exemplos de relações PRODUTOR:

borborigmo, s. m. ruído produzido por gases nos intestinos
→gases PRODUTOR_DE borborigmo

sublimado, adj. obtido por sublimação
→sublimação PRODUTOR_DE_ALGO_COM_PROPRIEDADE sublimado

fotógeno, adj. que gera ou emite luz
→fotógeno PROPRIEDADE_DE_ALGO_PRODUTOR_DE luz

FINALIDADE

Definimos que uma relação de FINALIDADE, como p FINALIDADE_DE v, ocorre quando v, um meio, tem ou pode ser utilizado com determinado objetivo ou finalidade, p. O meio v pode ser um instrumento (substantivo) ou um procedimento (descrito sob a forma de um substantivo ou verbo). Por seu lado, a finalidade p pode ser um estado (substantivo) ou uma ação (verbo).

Tendo em conta as categorias gramaticais dos seus argumentos, foram definidas as seguintes sub-relações de FINALIDADE:

substantivo FINALIDADE_DE substantivo
substantivo FINALIDADE_DE_ALGO_COM_PROPRIEDADE adjetivo
adjetivo PROPRIEDADE_DE_ALGO_FINALIDADE_DE substantivo
verbo ACCAO_FINALIDADE_DE substantivo
substantivo FINALIDADE_DA_ACCAO verbo
verbo ACCAO_FINALIDADE_DE_ALGO_COM_PROPRIEDADE adjetivo

Esta relação foi extraída tirando partido de padrões com as seguintes palavras chave:

Verbos relativos à utilização: *usar, utilizar;*
Verbos relativos à função: *servir;*
Outros verbos: *recorrer;*
Indicadores de finalidade: *finalidade, fim, objetivo, intuito;*
Expressões indicadores de meio para: *por meio de, com o auxílio de;*
Preposição indicadora de função: *para.*

Alguns exemplos de relações FINALIDADE:

pente, s.m. Instrumento de ferro usado para cardar a lã
→cardar_a_lã FINALIDADE_DE pente

comédia, s.f. obra de ficção cuja finalidade é fazer rir
→fazer_rir FINALIDADE_DE comédia

arrolo, s.m. toada para adormecer as crianças
→adormecer_as_crianças FINALIDADE_DE arrolo

cooperativa, s.f. associação que tem como objectivo a construção de habitações a custos controlados destinadas aos seus membros
→construção_de_habitações FINALIDADE_DE

cooperativa

simulação, s.f. diferença entre a vontade e a declaração, estabelecida por acordo entre as partes, com o intuito de enganar terceiros
→enganar FINALIDADE_DE simulação

enumerativo, adj. que serve para a enumeração
→enumeração
FINALIDADE_DE_ALGO_COM_PROPRIEDADE enumerativo

preventivo, adj. que tem por fim prevenir, acautelar ou impedir
→prevenir
FINALIDADE_DE_ALGO_COM_PROPRIEDADE preventivo
→acautelar
FINALIDADE_DE_ALGO_COM_PROPRIEDADE preventivo
→impedir
FINALIDADE_DE_ALGO_COM_PROPRIEDADE preventivo

festivaleiro, adj próprio para festival
→festival
FINALIDADE_DE_ALGO_COM_PROPRIEDADE festivaleiro

LOCAL

Uma relação de LOCAL, p LOCAL_DE v, ocorre entre dois substantivos, quando p é um local e v é natural de, habita ou pode encontrar-se em p. Os padrões utilizados para extrair esta relação levam em consideração as palavras-chave *natural*, *habitante* e *originário*.

Alguns exemplos de relações LOCAL:

coreano, s.m. natural ou habitante da Coreia do Norte ou da Coreia do Sul
→Coreia_do_Norte LOCAL_ORIGEM_DE coreano
→Coreia_do_Sul LOCAL_ORIGEM_DE coreano

fascólomo, s.m. mamífero marsupial semelhante ao texugo, originário da Austrália.
→Austrália LOCAL_ORIGEM_DE fascólomo

favelado, adj. e s.m. habitante ou designativo de habitante de favela
→favela LOCAL_ORIGEM_DE favelado

baiano, s.m. indivíduo natural da Bahia
→Bahia LOCAL_ORIGEM_DE baiano

2.3 Rebatismo de algumas relações para tornar o processo de validação mais intuitivo

É imediatamente observável que alguns nomes de relações no PAPEL são muito compridos e difíceis de compreender no contexto de um triplo. De forma a tornar o processo de validação no VARRA mais claro, em alguns casos mudamos o nome para algo mais fácil de ler e sobre o qual opinar.

A alteração é apenas uma maneira de tornar mais fácil a interpretação, uma questão cosmética, e resume-se simplesmente a traduzir, no contexto do serviço na Internet, as relações com nomes compridos por outras mais intuitivas, tais como

PROPRIEDADE_DE_ALGO_REFERENTE_A = DIZ-SE_SOBRE
viril DIZ_SE_SOBRE homem

PRODUTOR_DE_ALGO_COM_PROPRIEDADE = FAZ_O_QUE_É

sublimação FAZ_O_QUE_É sublimado

PROPRIEDADE_DE_ALGO_CAUSADOR_DE=O_QUE_CAUSA_É
reacção O_QUE_CAUSA_É reactivo

CAUSADOR_DE_ALGO_COM_PROPRIEDADE=CAUSA_O_QUE_É

PARTE_DE_ALGO_COM_PROPRIEDADE=É_TEM
coberto É_TEM tampa

PROPRIEDADE_DE_ALGO_MEMBRO_DE=É_PORQUE_PERTENCE
centro-europeu É_PORQUE_PERTENCE_A Centro da Europa

3. O VARRA

Como já mencionamos, o VARRA é um sistema criado para permitir uma validação detalhada de relações semânticas entre pares de palavras quando inseridas em um contexto, e, para o presente artigo, trataremos da validação das relações presentes no PAPEL.

Especificamente, pretendemos, com o VARRA, obter julgamentos precisos dos falantes com relação à seguinte questão: dado um triplo e uma frase, obtida automaticamente dos corpos do AC/DC, pedimos às pessoas que informem se a frase em questão ilustra – e, consequentemente, valida – a relação indicada no triplo. Assim, perguntamos, por exemplo, se uma dada frase ilustra a relação de sinonímia entre "mentira" e "ilusão". Nos trechos abaixo, as respostas seriam SIM e NÃO, respectivamente.

> *Mudança maior, porém, vem do novo presidente do Supremo Tribunal Federal, ministro Sepúlveda Pertence, que afirmou: `Desde que se superou a* ***mentira*** *de que um juiz, particularmente um juiz constitucional, é um puro técnico capaz de extrair uma norma supostamente de um único sentido válido de um fato, desde que essa* ***ilusão*** *foi desfeita, a verdade é que o juiz é um homem, enquanto cidadão, com crenças, convicções, tendências conscientes e inconscientes.*
>
> *Não era uma* ***mentira****, era* ***ilusão****, o cinema possui esse poder de criar ilusão.*

No entanto, a fim de obtermos respostas mais informativas do que simplesmente SIM / NÃO, e porque sabemos que, ao lidar com o significado das palavras, são frequentes os casos que se situam em uma zona cinzenta, o VARRA prevê cinco diferentes alternativas para dar conta das relações entre os pares de palavras e os textos. Estas alternativas correspondem a cinco possíveis respostas à pergunta "Os textos dos exemplos ilustram a relação entre as duas palavras apresentada na primeira coluna?"

- **1**: Sim.
- **2**: Não. É compatível com a relação, mas não a exemplifica.
- **3**: Não. O texto é completamente não relacionado.
- **4**: Não. Pelo contrário, invalida-a.
- **5**: Não sei.

A seguir explicamos brevemente cada uma das alternativas, usando frases autênticas encontradas nos corpos do AC/DC:

1. SIM.

Essa deve ser a alternativa escolhida quando o texto ilustrar a relação-alvo entre os pares de palavras.

1a - Relação a ser validada: `feijão PARTE_DE feijoada`
Texto*: Sábado é o dia de Tsholent, um tipo de* ***feijoada*** *judaica feita com* ***feijão*** *branco, batatas e carne de boi, no Cecília.*

2. NÃO. É compatível com a relação, mas não a exemplifica.

Com essa alternativa, procuramos cobrir os casos em a frase se encaixa na relação alvo, é compatível com ela, mas não a ilustra ou exemplifica.

2a - Relação a ser validada: `grana SINONIMO_DE dinheiro`
Texto: *Um cidadão cheio da* ***grana****, desesperado dos conselhos de seus assessores, ousou me consultar a respeito de um problema que eu nunca tive: em caso de reviravolta social, o que ele deveria fazer para perder* ***dinheiro*** *rápida e honestamente.*

No exemplo acima, embora os termos em negrito sejam usados como sinônimos, a frase não ilustra, necessariamente, a relação de sinonímia entre grana e dinheiro (ainda que seja compatível com ela).
Vale notar também que, em alguns casos, é possível que a decisão entre uma resposta do tipo 1 e uma do tipo 2 dependa da interpretação do avaliador.

3. NÃO. O texto é completamente não relacionado.

Com essa alternativa, buscamos cobrir os casos em que o texto exemplo não é capaz de fornecer qualquer pista sobre a relação em questão. Isso pode acontecer porque

a) alguma das palavras está sendo usada num sentido que não o indicado pelo par.

3a.Relação validada: `fruta HIPERONIMO_DE manga`
Texto: *Durante o banquete, Nasrudin ia jogando as* ***frutas*** *e a comida pela* ***manga*** *da túnica.*

3b.Relação validada: `roda PARTE_DE carro`
Texto*: A lei obriga a ter no* ***carro*** *chaves de* ***roda*** *e de fenda, macaco, triângulo e extintor de incêndio.*

b) a frase em si não é relevante para a validação de que se está à procura. Nos exemplos abaixo, não é possível estabelecer qualquer relação entre feijão e feijoada (3c) ou fruta e abacaxi (3d), por exemplo.

3c - Relação a ser validada: `feijão PARTE_DE feijoada`
Texto: *Mesmo assim, aí vai uma lista: vários bacalhaus (cozido, cru, assado e arroz , sopas (de alheiras, de tomate, de grão, de batata, de alho, de* ***feijão****, etc.), arroz de pato, caldeirada (de peixe e de cabrito) , moamba de galinha, cabrito assado nas brasas, perdiz (várias receitas) e* ***feijoada*** *de lebre .*

3d. Relação a ser validada: `fruta HIPERONIMO_DE abacaxi`
Texto*: Ensina a fazer fofinho de chocolate (rocambole),* ***frutas*** *com gelatina, friturinha de maçã, coroa de* ***abacaxi*** *e cajuzinho*

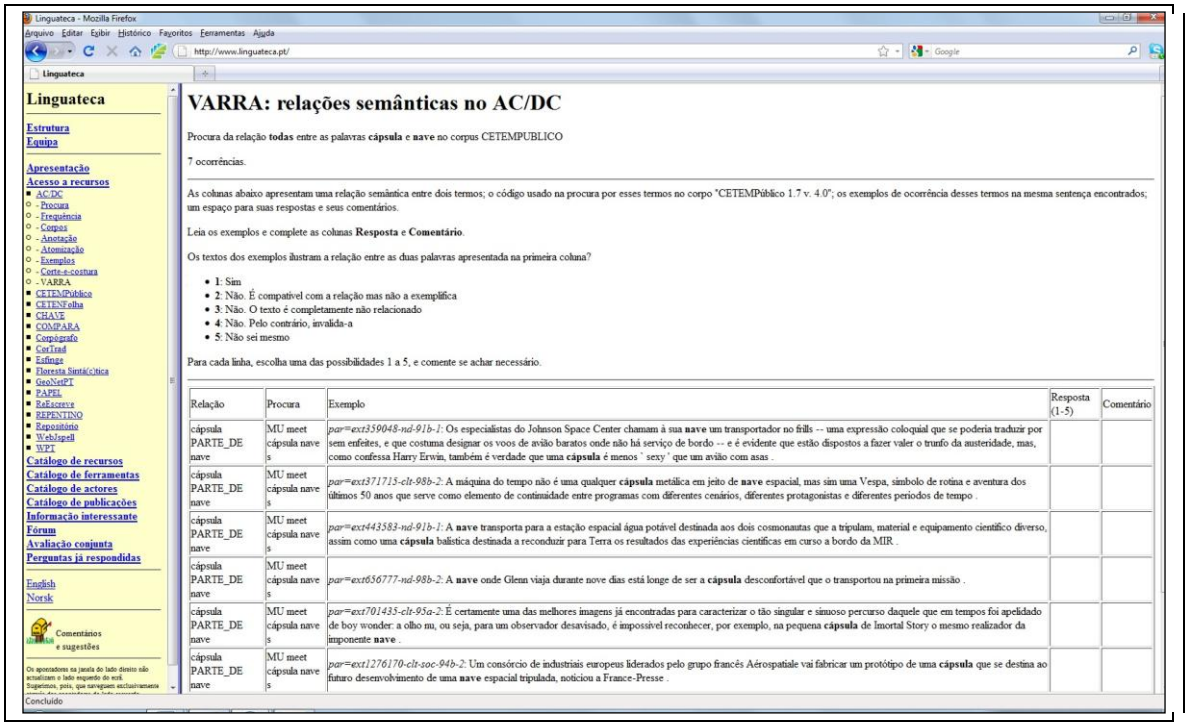


Figura 1: Interface do VARRA

4. NÃO. O texto não valida a relação; pelo contrário, invalida-a

A alternativa 4 corresponde aos casos em que não apenas é impossível validar a relação a partir do texto, mas o exemplo contraria, ou invalida, a relação.

4a. Relação a ser validada: `mentira SINONIMO_DE ilusão`
Texto: *Não era uma* ***mentira****, era* ***ilusão****, o cinema possui esse poder de criar ilusão.*

5: Não sei

Por fim, a alternativa 5 se aplica quando

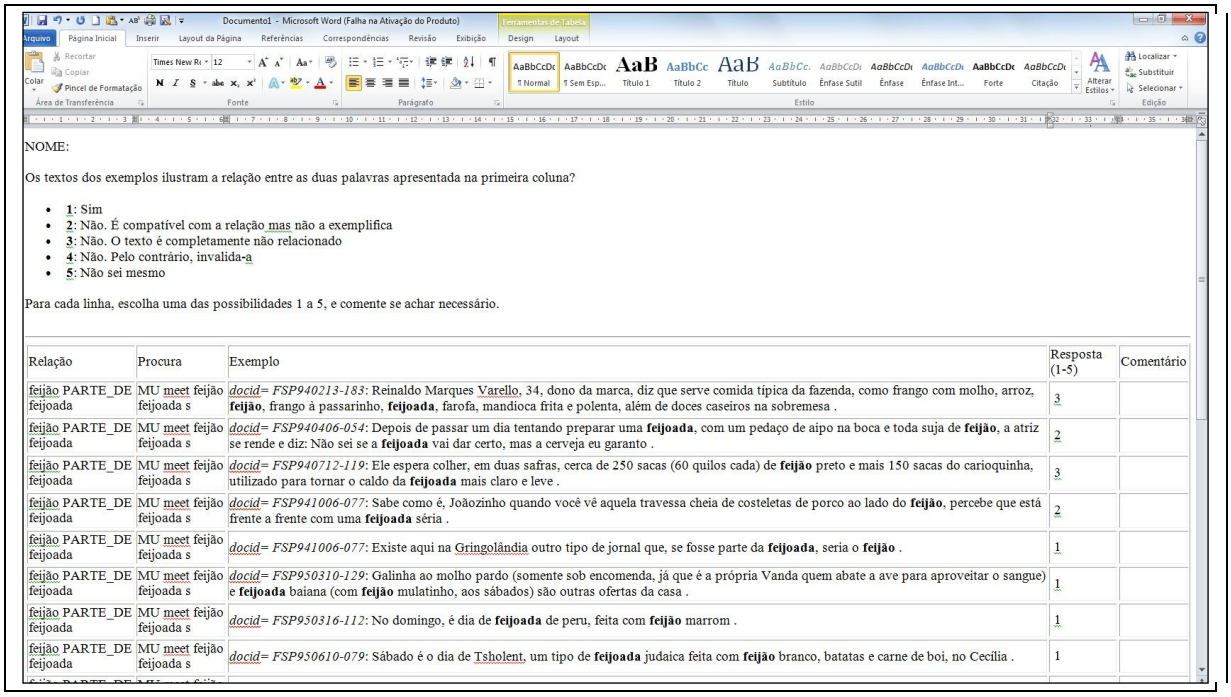
a) o avaliador não consegue perceber a relação expressa no texto

5a. Relação a ser validada: `mentira SINONIMO_DE ilusão`
Texto: *Distingue, nas promessas que lhe fazem, o que é* ***mentira*** *ou* ***ilusão*** **.**

b) há dúvida entre mais de uma opção
c) o texto é incompreensível.

Todas as relações e frases-exemplo estão acessíveis por meio de uma interface na rede, que contém ainda colunas para que os avaliadores façam seus julgamentos e,

eventualmente, incluam comentários (figura 1). Cada página com os dados pode ser levada para um editor de texto, e assim são criados o que chamamos de dossiês, que são em seguida distribuídos para os avaliadores (figura 2).



NOME:

Os textos dos exemplos ilustram a relação entre as duas palavras apresentada na primeira coluna?

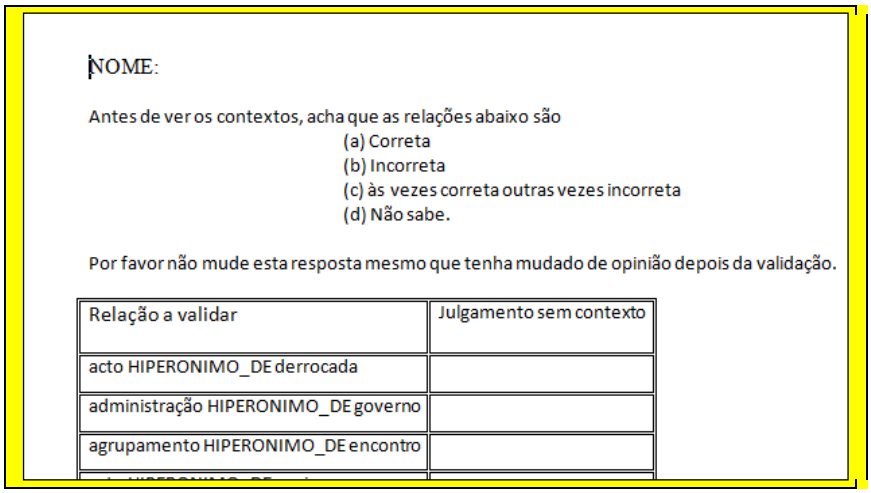
- **1**: Sim
- **2**: Não. É compatível com a relação mas não a exemplifica
- **3**: Não. O texto é completamente não relacionado
- **4**: Não. Pelo contrário, invalida-a
- **5**: Não sei mesmo

Para cada linha, escolha uma das possibilidades 1 a 5, e comente se achar necessário.

Relação	Procura	Exemplo	Resposta (1-5)	Comentário
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP940213-183*: Reinaldo Marques Varello, 34, dono da marca, diz que serve comida típica da fazenda, como frango com molho, arroz, **feijão**, frango à passarinho, **feijoada**, farofa, mandioca frita e polenta, além de doces caseiros na sobremesa .	3	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP940406-054*: Depois de passar um dia tentando preparar uma **feijoada**, com um pedaço de aipo na boca e toda suja de **feijão**, a atriz se rende e diz: Não sei se a **feijoada** vai dar certo, mas a cerveja eu garanto .	2	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP940712-119*: Ele espera colher, em duas safras, cerca de 250 sacas (60 quilos cada) de **feijão** preto e mais 150 sacas do carioquinha, utilizado para tornar o caldo da **feijoada** mais claro e leve .	3	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP941006-077*: Sabe como é, Joãozinho quando você vê aquela travessa cheia de costeletas de porco ao lado do **feijão**, percebe que está frente a frente com uma **feijoada** séria .	2	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP941006-077*: Existe aqui na Gringolândia outro tipo de jornal que, se fosse parte da **feijoada**, seria o **feijão** .	1	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP950310-129*: Galinha ao molho pardo (somente sob encomenda, já que é a própria Vanda quem abate a ave para aproveitar o sangue) e **feijoada** baiana (com **feijão** mulatinho, aos sábados) são outras ofertas da casa .	1	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP950316-112*: No domingo, é dia de **feijoada** de peru, feita com **feijão** marrom .	1	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP950610-079*: Sábado é o dia de Tsholent, um tipo de **feijoada** judaica feita com **feijão** branco, batatas e carne de boi, no Cecília .	1	

Figura 2: Dossiê criado pelo VARRA

Os validadores, também chamados de “varredores”, têm de preencher antes a sua opinião sobre a correção das relações fora do contexto (veja-se um exemplo na figura 3). Esta tarefa é para ser usada como um controle, quer para identificar casos em que as relações a validar são erradas (o que é perfeitamente possível dado terem sido extraídas automaticamente), quer para verificar se as intuições mudam com a exposição ao material.



NOME:

Antes de ver os contextos, acha que as relações abaixo são
(a) Correta
(b) Incorreta
(c) às vezes correta outras vezes incorreta
(d) Não sabe.

Por favor não mude esta resposta mesmo que tenha mudado de opinião depois da validação.

Relação a validar	Julgamento sem contexto
acto HIPERONIMO_DE derrocada	
administração HIPERONIMO_DE governo	
agrupamento HIPERONIMO_DE encontro	

Figura 3: Formulário com as relações sem contexto

Para facilitar o trabalho dos “alunos-varredores”, foi criado um manual (Freitas, 2010), em que são exemplificadas as várias alternativas de resposta, bem como fornecidas indicações de preenchimento do campo “Comentários”.

5. Relato da análise dos primeiros dossiês

Em um primeiro teste com o VARRA, foram distribuídos dossiês para 10 alunos de graduação do curso de Letras da PUC-Rio,[ii] resultando num total de 3271 instâncias de relações em contexto avaliadas (obtivemos 5243 julgamentos). Cada dossiê contém cerca de 200 instâncias de relações e, por questões operacionais, optamos por apresentar um tipo de relação por dossiê.

Mais pormenorizadamente, coligimos julgamentos relativos a 276 triplos diferentes, ilustrando as relações MEMBRO_DE (212), PARTE_DE (2156), HIPERONIMO_DE (1762) e PRODUTOR_DE (1113).

Os resultados obtidos, de uma forma sumária, foram os seguintes. A tabela 1 apresenta como, sem contexto, os 276 triplos foram julgados pelos varredores. Vários triplos (83) foram julgados por mais do que um varredor. Por outro lado, em 42 casos faltou esse julgamento no dossiê, por isso apresentamos 403 resultados:

Corretos	129
Incorretos	113
Às vezes uma, outras vezes outra	100
Não sabe	61

Tabela 1. Julgamento dos triplos sem contexto

Passando agora a relatar os julgamentos em contexto, tivemos 1365 casos em que a frase valida a relação, 966 em que é simplesmente compatível, 2148 em que não permite ao avaliador julgar, 536 em que a invalida, e 228 em que o varredor preferiu não se pronunciar. Mais detalhadamente, apresentamos na tabela 2 os casos por cada tipo de relação.

Relação	valida	compatível	não relacionada	invalida	não sabe
MEMBRO_DE	89	19	93	6	5
PARTE_DE	738	445	589	289	95
HIPERONIMO_DE	225	198	1144	114	81
PRODUTOR_DE	313	304	322	127	47

Tabela 2: os resultados obtidos por tipo de relação

Criamos também algumas medidas que nos permitem caracterizar quantitativamente o material, nomeadamente:

1) Grau de correta aplicação (GCA): esta medida busca refletir, para cada triplo, quantas vezes a relação semântica que ele ilustra descreve o que se passa no texto. Ou seja, quanto maior for o GCA, mais garantido que a relação é verdadeira em instâncias do uso das duas palavras em português. Temos dois tipos de GCA, parcial e total:

(a) GCA parcial (para cada triplo): (total(1) + total(2)) / total (julgamentos) (no que se segue, **n**: total de triplos, $\mathbf{m_i}$: total de julgamentos em contexto de um dado triplo i, $\mathbf{julgamento_i}$: julgamento em contexto do triplo em questão) Para calcular o GCA, usamos além disso a seguinte quantidade intermédia **ca**, que toma o valor 1 se só houver julgamentos do tipo "1" ou "2", e o valor 0 se houver julgamentos de outros tipos

$$GCA\ (triplo_i) = \frac{\sum_{j=1}^{m_i} ca(j)}{m_i}$$

(b) GCA total: a média dos GCA para todos os triplos considerados.

$$GCA\ (total) = \frac{\sum_{i=1}^{n} GCA(triplo_i)}{n}$$

Ainda que, de maneira mais imediata, o GCA por triplo seja a medida mais importante, o cálculo do GCA total pode ser uma medida indicadora para a comparação entre um conjunto de triplos (ou de varredores). Valor do GCA total no caso do experimento relatado acima: 0.4

2) Grau de previsibilidade (GP): indica a relação entre a competência linguística sem contexto e a realidade do uso na língua. Em triplos com alta previsbilidade não seria necessário usar o VARRA para a validação em contexto. Mas pode haver casos em que essa previsibilidade é baixa, e são esses os casos que queremos medir e averiguar.

(a) GP Parcial (para cada triplo): correspondências (sem_contexto --> com_contexto) / total(julgamentos_validos)

$$GP\ (triplo_i) = \frac{\sum_{j=1}^{m_i} corresp(s\_cont(triplo_i),\ julgamento_j)}{m_i}$$

(b) GP Total: soma(parciais) / total(triplos): Valor do GP total: 0.3

$$GP_{total} = \frac{\sum_{i=1}^{n} GP(triplo_i)}{n}$$

3) Dispersão: indica as diferenças de julgamentos entre diferentes varredores e/ou a diferença de julgamentos entre diferentes contextos. Triplos com maior dispersão deverão naturalmente indicar dificuldade de julgamento ou complexidade da relação a ser julgada.

(a) Dispersão Parcial: para calcular a dispersão parcial associada a cada triplo, ordenam-se os diferentes tipos de julgamento por frequência. Em primeiro lugar fica o julgamento que teve mais votos, em segundo o que teve o próximo número, etc. Calcula-se então a dispersão parcial como a soma das seguintes quantidades: a proporção de vezes que o triplo teve um dado julgamento, multiplicado pela sua ordem. Por exemplo, se houve

35 julgamentos para um dado triplo, divididos por 3 tipos, respectivamente 30, 4, e 1, a soma relevante é: (30/35)*1 + (4/35)*2 + (1/35)*3
Se todos os julgamentos fossem iguais, a dispersão seria 1.

(b) Dispersão Total: soma(parciais) / total(triplos): 9.3

4) Grau de confiança: é a medida inversa da dispersão. O grau de confiança corresponde ao veredito que os julgamentos humanos passam sobre a correção ou boa-formação do triplo em causa (ou sobre a sua incorreção ou malformação), e serve pois para julgar ou avaliar os triplos propostos. Casos em que o grau de confiança é baixo poderão exigir mais dados para avaliar com o VARRA, enquanto casos com elevado grau de confiança podem ser usados para avaliar o PAPEL (ou outro recurso usado).
GC = 1 / Dispersão: 0.1

Na tabela 3, referente a cada triplo, indicamos quantitativamente: os casos em que os triplos foram validados (GCA > 0,5), os casos inconclusivos (GCA entre 0 e 0,5), e os casos em que não houve validação, e que muito provavelmente correspondem a erros no PAPEL (GCA = 0).

Casos	Número de triplos
Validados	119
Inconclusivos	65
Invalidados	83

Tabela 3: Panorâmica em termos de triplos

O quadro 1 apresenta alguns exemplos de triplos validados, inconclusivos e invalidados.

Triplos validados	Triplos inconclusivos	Triplos invalidados
comunidade HIPERONIMO_DE falange	abelha HIPERONIMO_DE zângão	doença HIPERONIMO_DE esgana
laringe PRODUTOR_DE som	apresentação HIPERONIMO_DE antestreia	watt PRODUTOR_DE trabalho
edema PRODUTOR_DE inchação	cura HIPERONIMO_DE resolução	representação PRODUTOR_DE efeito
laringe PRODUTOR_DE som	dínamo PRODUTOR_DE corrente	copo HIPERONIMO_DE imperial
berma PARTE_DE plataforma	aviso HIPERONIMO_DE anúncio	saúde PRODUTOR_DE euforia
diapasão PRODUTOR_DE frequência	coxa PARTE_DE membro	substância PRODUTOR_DE luz

Quadro 1: Exemplos de triplos por tipo de classificação

Uma análise preliminar dos resultados GCA por triplo nos mostra alguns dados interessantes. Tomaremos apenas o exemplo do triplo invalidado copo HIPERONIMO imperial, cujo GCA = 0. Em Portugal, “imperial” é cerveja na pressão (que, no Brasil, chamamos “chope”), e “copo” é (informalmente) entendido como uma bebida alcoólica qualquer (“tomar um copo” ou “beber um copo”). Os varredores brasileiros nao apenas desconheciam este uso de “imperial”, como também, pelos contextos, por se tratar de um registro informal com poucas ocorrências no AC/DC, tal relação era impossível de ser inferida (um varredor que conhecesse o significado de “imperial” provavelmente atribuiria os julgamentos 2 e 3 às frases-exemplo.) Este

exemplo nos mostra que, para o futuro, será preciso levar em consideração o possível desajuste entre os corpos (e varredores) e um certo tipo de triplos. Além disso, vale mencionar que nos causou estranheza a definição de "imperial" presente no dicionário - "*copo de 33 cl, alto e mais estreito em baixo do que em cima, com cerveja tirada à pressão*" - em que o sentido de "copo" é o de "objeto", visto ser imperial um tipo de cerveja, e nao de copo (ou, antes, acreditamos que uma definição mais adequada seria algo como " *cerveja servida num copo de 33cl ...*").

Finalmente, na tabela 4 apresentamos uma panorâmica dos 1414 casos em contexto em que a validação foi feita por mais do que um varredor, e tabulamos os casos de concordância absoluta, parcial, ou discordância.

Casos	número
Concordância absoluta	752
Concordância parcial	558
Discordância	107

Tabela 4: Panorâmica em termos de concordância entre varredores

Apresentamos os valores obtidos como demonstração de que todo o enquadramento já está implementado, mas naturalmente o material não é ainda suficientemente vasto para estudos quantitativos, e a utilidade destas medidas terá de esperar confirmação do seu uso posterior.

Embora não tenhamos ainda dados suficientes para estabelecer conclusões ou tendências, um dado interessante é o grande número de ocorrências, na relação de HIPERONIMIA, em que a relação que se estabelece é a de sinonímia. Isto porque, frequentemente, usamos termos hiperônimos para dar coesão ao texto, em uma relação anafórica – e, nesses casos, os termos hiperônimos acabam por funcionar também como sinônimos, como ilustra o exemplo a seguir:

Relação a ser validada: `fruta HIPERONIMO_DE abacaxi`
Texto: *A região de Bauru (345 km a noroeste de São Paulo) é a maior produtora de* ***abacaxi*** *de sobremesa do Estado, com 600 hectares da* ***fruta*** *plantados.*

Esses casos, marcados como ocorrências do tipo 2, totalizam 4.7% das relações de hiperonímia, e 17.5% das relações de hiperonímia do tipo 2.

6. O VARRA no ensino

De um ponto de vista pedagógico, a utilização do VARRA tem como ponto positivo a capacidade de provocar, no aluno, a reflexão relativa ao comportamento semântico das palavras.

Os manuais de semântica costumam apresentar as relações entre palavras sem, no entanto, problematizá-las. Assim, acreditamos que o VARRA pode ser utilizado em aula de duas maneiras opostas e complementares.

Em um momento inicial, as relações do VARRA, ou mesmo dossiês previamente preparados, com relações selecionadas segundo o interesse do professor, podem funcionar como ilustração. Se assumirmos que (i) as relações semânticas têm seu

valor construído (principalmente?) no contexto; (ii) os exemplos presentes nos livros didáticos são escassos e quase sempre os mesmos, e raramente aparecem contextualizados, não é difícil perceber que a apresentação da relação entre pares de palavras em frases espontâneas da língua, isto é, em corpos, pode contribuir para uma melhor compreensão do fenômeno que se deseja ilustrar.

Por outro lado, ou mesmo em um momento posterior, é possível usar os exemplos em contexto justamente para problematizar as relações, apresentadas normalmente como pontos para os quais não há discussão. Por se apoiarem normalmente em exemplos prototípicos, dificilmente são apresentadas as dificuldades de se lidar com o estabelecimento de relações entre abstrações, por exemplo. Assim, com o VARRA, é possível problematizar, a partir de exemplos reais, os seguintes pontos:

a) Pares de palavras que se relacionam por meio de mais de uma relação;

b) Contextos que contradizem uma dada relação, normalmente estabelecida;

c) Dúvidas e incertezas durante o estabelecimento de relações;

d) Percepção de que nem sempre as relações propostas são capazes de dar conta das relações que se estabelecem na língua, explorando a motivação subjacente à criação dos tipos de relações semânticas normalmente propostos. Especificamente, é possível encaminhar a discussão para as seguintes questões: há inadequação? Quando? Há relações que não têm um nome explícito? Quais seriam?

Além da problematização das relações semânticas, outra possibilidade é explorar, com os alunos, os mecanismos utilizados pela língua para expressar relações semânticas. Isto é, trilhar o caminho inverso do VARRA, e refazer o caminho do PAPEL: a partir dos exemplos dos corpos, tentar identificar padrões linguísticos capazes de revelar determinadas relações semânticas.

Outro ponto interessante é examinar o contraste entre as intuições com relação às relações semânticas antes e depois da sua apresentação em contexto. Houve mudança de opinião entre as intuições e as relações em contexto? Houve casos em que essa mudança de opinião foi mais frequente? Por quê?

Por fim, mas não menos interessante, o VARRA pode ser uma ótima ferramenta para transformar a aula em uma oficina – de semântica e de pesquisa. Em um primeiro momento, o professor pode preparar dossiês em que apareçam os fenômenos que deseja tratar. Em seguida, distribui os dossiês para os alunos. Após explicar as alternativas, pede para que cada aluno preencha um dossiê, individualmente e sem consultar os colegas. Depois, contabilizam-se as respostas dos alunos. Houve mais similaridades ou discrepâncias? Que relações, em que contextos, deixaram pouca margem para variação nas respostas? Que relações, por outro lado, foram as que mais apresentaram variação? É possível alguma conclusão? (Idealmente, os próprios alunos podem ser os responsáveis pela tabulação e análise dos resultados, que por sua vez podem dar margem a um relatório de pesquisa.)

7. Considerações finais

Oferecemos o VARRA à comunidade num estágio ainda inicial, para que nos ajudem a definir aquilo que pode ser útil na descoberta e validação de relações semânticas em português, não só em termos conceituais, mas também usando, para validação na sala de aula (por todas as partes onde se fala e ensina português e linguística em português), vários dossiês que tenhamos criado ou que queiram criar e partilhar com a equipa do VARRA. Encorajamos por isso que nos contatem e que reutilizem os dossiês que já criámos[iii], ou que criem novos.

Desta forma, poderemos não só criar um repositório de relações validadas, como também de forma de validação e de diferenças de opinião entre os varredores, que também possam ser estudadas e, consequentemente, contribuir para a compreensão da semântica da língua portuguesa.

No processo de elaboração do VARRA, nossa escolha pelas alternativas 1-5 foi motivada não apenas pelas informações relevantes que podem ser oferecidas para avaliação e melhoria do PAPEL, mas também porque, de um ponto de vista linguístico, as alternativas nos pareceram suficientemente objetivas, mas sem perder as nuances que o julgamento humano é capaz de oferecer. Por essa razão estamos convencidos de que os resultados obtidos através do uso do VARRA também podem interessar à pesquisa em linguística, além de melhorar os recursos computacionais para a nossa língua.

De um ponto de vista linguístico, gostaríamos de aprofundar, por exemplo, a investigação relativa às contradições entre os julgamentos sem e com contexto; aos triplos que, simultaneamente, podem ser validados ou negados conforme o contexto; aos tipos de construções que mais frequentemente validam um dado triplo, por exemplo.

Quanto aos resultados apresentados neste trabalho, salientamos o caráter preliminar e experimental. Temos consciência de que, em termos quantitativos, ainda não é possível, a partir dos dados obtidos, produzir uma análise satisfatória sobre o tema, mas em nosso favor advogamos que esta foi uma maneira (i) de testarmos a forma de proceder com o VARRA e, como dito acima, (ii) divulgar todo o trabalho para a comunidade interessada, para que possamos, em conjunto, contribuir para estudos semânticos da língua e, também, se for o caso, aprimorar as funcionalidades do VARRA. Fica como compromisso e desejo para um futuro próximo a análise quantitativa dos resultados, assim que tivermos mais dossiês validados.

Sobre o tratamento estatístico e utilização dos resultados, devemos salientar que é perfeitamente possível, para diferentes objetivos, agregar as diferentes respostas em algo mais simples: um determinado usuário pode, por exemplo, estar interessado apenas em ter uma estimativa simples da qualidade dos seus resultados, e portanto pode estar interessado apenas em resultados do tipo SIM – NÃO (nesse caso, poderia e deveria agrupar as respostas do tipo 1-2 para SIM, e as restantes para NÃO).

Além disso, o VARRA pode ser útil na obtenção semi-automática (após validação humana), de exemplos de uso (glosas?) para um dado recurso. Por exemplo, podemos, no futuro, usar as frases do tipo 1 como exemplos das relações do PAPEL. A atividade de extração automática de exemplos para dicionários não é, de qualquer forma, trivial, como argumentam Kilgarriff et al. (2008).

Finalmente, quanto às frases do tipo 2 e 3, podem servir como frases-alvo para melhorar a busca por padrões, tanto do VARRA como de sistemas automáticos de deteção de relações em texto, por exemplo como material de treino para o aprendizado automático.

Agradecimentos

O VARRA está sendo desenvolvido no âmbito da Linguateca, co-financiada pelo governo português, pela União Européia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN.

Hugo Gonçalo Oliveira é apoiado pela bolsa FCT SFRH/BD/44955/2008, co-financiada pelo FSE.

Agradecemos à Andréa Barreto (bolsista PIBIC/CNPq) e aos alunos da PUC-Rio que participaram da validação preenchendo dossiês do VARRA: Vittorio Provenza, Jéssica Barcelos, Amanda Pacheco, Marcia Aleixo, Rafaella Hernandes, Alberta Barros, Marcela Lanius, Francisco Camelo, João Netto, Leticia Wendel e Isadora Guedes.

Referências bibliográficas

BICK, Eckhard. "*The Parsing System Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*". Dr.phil. tese. Aarhus University. Aarhus, Dinamarca: Aarhus University Press. Novembro de 2000.

CARABALLO, Sharon. "Automatic construction of a hypernym-labeled noun hierarchy from text". In *Proceedings of the 37th Annual Meeting of the Association For Computational Linguistics on Computational Linguistics* (College Park, Maryland, June 20 - 26, 1999). Association for Computational Linguistics, Morristown, NJ, pp. 120-126.

COSTA, Luis, SANTOS, Diana & ROCHA, Paulo Alexandre. "Estudando o português tal como é usado: o serviço AC/DC". In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)* (São Carlos, SP, Brasil, 8-11 de setembro 2009).

DIAS-DA-SILVA, Bento Carlos & MORAES, Helio Roberto de. "A construção de um thesaurus eletrônico para o português do Brasil ". *ALFA* **47**, 2, 2003, pp. 101-115.

DPLP. *Dicionário PRO da Língua Portuguesa*. Porto Editora, Porto, Portugal, 2005.

FELLBAUM, Christiane. “WordNet: An Electronic Lexical Database”, MIT Press, 1998.

FREITAS, Cláudia. “Instruções para a validação de relações semânticas entre palavras usando o VARRA - Validação, Avaliação e Revisão de Relações semânticas no AC/DC.” Versão 2, 29 de Novembro de 2010. http://www.linguateca.pt/acesso/InstrucoesVARRA.pdf

GONÇALO OLIVEIRA, Hugo, SANTOS, Diana & GOMES, Paulo. "Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação". *Linguamática* **2**, 1, 2010, pp. 77-93.

GONÇALO OLIVEIRA, Hugo, SANTOS, Diana, GOMES, Paulo & SECO, Nuno. "PAPEL: a dictionary-based lexical ontology for Portuguese". In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)* (Aveiro, Portugal, 8-10 de Setembro, 2008), Springer Verlag, pp. 31-40.

HIRST, Graeme. "Ontology and the lexicon". In Steffen Staab & Rudi Studer (eds.). *Handbook on ontologies*, Springer, 2004, pp. 209-229.

ITOO, Ashwin & BOUMA, Gosse. “On learning subtypes of the part-whole relation: do not mix your seeds”. In *Proceedings of the 48th Annual Meeting of the Association for Computacional Linguistics* (ACL 2010), pp. 1328-1336, Uppsala, Suécia, 2010.

KILGARRIFF, Adam, HUSÁK, Milos, MCADAM, Katy, RUNDELL, Michael & RYCHLÝ, Pavel. “GDEX: Automatically finding good dictionary examples in a corpus”. *Proc EURALEX 2008*, Barcelona, Espanha, 2008.

RICHARDSON, Stephen. "Determining Similarity and Inferring Relations in a Lexical Knowledge Base", Ph.D. thesis, The City University of New York, 1997, Microsoft Research Report MSR-TR-97-02.

RICHARDSON, Stephen, VANDERWENDE, Lucy & DOLAN, William. “Combining dictionary-based and example-based methods for natural language analysis. In *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, Japão, 1993, pp. 69–79.

RICHARDSON, Stephen, DOLAN, William & VANDERWENDE, Lucy. "MindNet: acquiring and structuring semantic information from text", *Proceedings of the 17th International Conference on Computational Linguistics, COLING-ACL'98* (August 10-14, Montréal, Québec, Canada), Vol. 2, pp. 1098-1102

RILOFF, Ellen & SHEPHERD, Jessica. "A corpus-based approach for building semantic lexicons". In Claire Cardie & Ralph Weischedel (eds.), *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing,* August 1-2, 1997, Brown University, Providence, Rhode Island, USA,1997.

SANTOS, Diana. "Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties". *OSLa: Oslo Studies in Language* **2** (2010).

SANTOS, Diana, BARREIRO, Anabela, FREITAS, Cláudia, GONÇALO OLIVEIRA, Hugo, MEDEIROS, José Carlos, COSTA, Luis, GOMES, Paulo & SILVA, Rosário. "Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL". In A. M. Brito, F. Silva, J. Veloso & A. Fiéis (eds.), *Textos selecionados. XXV Encontro Nacional da Associação Portuguesa de Linguística*. APL, 2010, pp. 681-700.

[i] Este exemplo não é ideal, visto que um ser humano preferiria extrair decretório PROPRIEDADE_DE_ALGO_COM_MEMBRO decreto mas não definimos tal relação n o PAPEL.

[ii] Cada um dos autores do presente artigo também preencheu um dossiê cada.

[iii] Em http://www.linguateca.pt/VARRA/ estão disponíveis todos os dossiês que preparamos.