

Guarda-fatos: notas sobre a anotação do campo semântico do vestuário em português

Diana Santos, Augusto Soares da Silva & Cristina Mota

Versão 2.3

Última actualização: 13 de Julho de 2010

(Primeira versão: 26 de Outubro de 2009)

1. Introdução

Este documento versa sobre a anotação do campo semântico do vestuário em todos os corpos AC/DC (e em geral em português), a partir da anotação seminal feita no corpo CONDIVport (Silva, 2006a,b,2008a,b) para estudar a convergência e divergência das variedades do português, e que foi integrado no projecto AC/DC (Santos & Sarmiento, 2003, Costa et al., 2009, Santos, 2010).

Estamos naturalmente conscientes do diferente objectivo da anotação que têm os dois projectos:

No primeiro caso, o da construção e anotação do CONDIV, foi tomada uma perspectiva de escolha criteriosa de campos/perfis com suficiente variação a nível lexical e sem demasiado peso de palavras raras, não padrão ou de nível de língua muito popular, e essa anotação foi feita exclusivamente em materiais cujo tema era a moda ou o vestuário, aliás na senda de Geeraerts & Grondelaers (1999) e Geeraerts et al. (1999).

As escolhas feitas obedeceram a três critérios. Em primeiro lugar, seleccionar termos alternativos para designar um mesmo referente, isto é, sinónimos denotacionais como meio de estudar convergência e divergência entre variedades linguísticas (neste caso, entre o português europeu e o português brasileiro). Os sinónimos denotacionais evidenciam tipos de diferenças sociolinguísticas e são estas diferenças regionais, sociais, estilísticas, pragmático-discursivas e históricas que definem a própria existência e a competição de variedades de uma língua. Em segundo lugar, estabelecer criteriosamente grupos de sinónimos na forma de perfis onomasiológicos: o *perfil onomasiológico* de um conceito numa determinada variedade linguística é o conjunto de termos sinónimos alternativos usados para designar esse conceito nessa variedade linguística, juntamente com as suas frequências relativas. Em terceiro lugar, foram excluídos os termos marcadamente populares para não inflacionar as diferenças entre as duas variedades nacionais do português. Com base nestes critérios, foram estabelecidos 22 perfis onomasiológicos (ver apêndice 1).

No segundo caso, que é o que nos interessa aqui, o de anotar completamente todos os corpos do AC/DC, pretendemos alargar o campo do vestuário a quaisquer palavras ou expressões com ele ligadas para permitir o seu varrimento/procura nos textos mais variados, e portanto não desejamos fazer limitações de vocabulário ou estilísticas. Como o objectivo do nosso trabalho não se resume ao estudo da variação sociolinguística, vamos estabelecer grupos de palavras quer os seus membros não apresentem diferenças conceptuais, como no caso anterior dos sinónimos denotacionais, quer os seus membros apresentem diferenças conceptuais de, por exemplo, classificação taxonómica.

Assim, enquanto a anotação do CONDIV moda se limitou, em consonância com os critérios acima referidos, a classificar sob, ou melhor, agrupar 225 termos diferentes em vinte e dois perfis onomasiológicos (correspondentes a uma peça de vestuário para homem, mulher ou unissexo, ver apêndice 1), nomeadamente BLUSAMULHER, BLUSÃO,

CALÇAS, CALÇASCURTAS, CALÇASJUSTAS, CAMISAHOMEM, CAMISOLA, CASACOMULHER, CASACOHOMEM, CASACOCURTO MULHER, CASACOCURTOHOMEM, CASACOCERIMONIA, CASACOMALHA, CASACOIMPERMEÁVEL, CASACOQUENTEINVERNO, CONJUNTOMULHER, FATOHOMEM, JAQUETA, JEANS, SAIA, T-SHIRT, VESTIDO, é claro que é possível, ao querer abranger todas as peças possíveis de roupa, produzir muitas mais classes (e sem que agora cada classe inclua apenas sinónimos denotacionais).

Uma das nossas preocupações nesta anotação é manter (e permitir o avanço de) o trabalho original feito ao longo dos anos pela equipa de Augusto Soares da Silva na Universidade Católica Portuguesa, permitindo portanto o acesso às classes originais, enquanto também obter uma maior cobertura do campo do vestuário em português.

2. Primeiros passos

A primeira tarefa a que nos dedicámos foi a de, a partir do conteúdo das classes iniciais do CONDIVport, anotar automaticamente o corpo CONDIV todo, o que permitiu naturalmente identificar casos flagrantes em que tal não era possível, assim como também detectar outros problemas, nomeadamente a dificuldade de em alguns casos identificar qual a acepção (das muitas que uma dada palavra de roupa pode ter) num dado contexto.

Um exemplo óbvio é *fato* (que sendo a grafia de *facto* no Brasil não representa nunca roupa em português brasileiro), outro é *camisola* que tem um sentido muito diferente no português de aquém e de além mar. Outros exemplos menos óbvios são ocorrências de *calças* que podem ser justas ou curtas ou compridas mas que pelo contexto limitado do texto nem sempre podemos decidir. Nesses casos (de uma mesma cadeia de caracteres poder pertencer a mais de um grupo) decidimos que todos os grupos seriam automaticamente adicionados, e desbastados numa fase posterior de revisão.¹

A segunda tarefa foi a de ir adicionando novos grupos à medida que os íamos encontrando no texto ou a partir do nosso conhecimento como falantes. E assim adicionámos grupos como ADEREÇO, MEIAS, ROUPADORMIR, SAPATOS assim como fomos populando com mais membros as classes já existentes. NÃOESPECIFICADA e OUTRAS pertencem à panóplia da anotação do AC/DC, para casos respectivamente "sem grupo", como são *roupa*, *trajo*, *trajes*, *indumentária*, etc., e para casos variados e raros em que não achámos necessário criar um grupo maior, como por agora *luvas*. Para uma lista exhaustiva de todas as classes (mas relembramos que este é trabalho em progresso), ver mais à frente.

Passámos depois a um tratamento mais exhaustivo do campo do vestuário seguindo a metodologia usada em Silva et al. (2008), Silva & Santos (em progresso) e Mota & Santos (2009), cujas premissas vamos documentar em seguida. Em poucas palavras, esse levantamento segue os seguintes passos, em que muitas das alíneas podem ter de ser feitas mais de uma vez, ou iterativamente:

1. inspecção das listas de frequências de todos os corpos, de forma a identificar palavras que podem ser roupa ou vestuário
2. classificação dessas palavras por grupo (ou conjunto de grupos)
3. classificação das mesmas por categoria gramatical, no caso de apenas serem roupa numa dada acepção

¹ Isto não significa que todos os casos de roupa tenham de ser reduzidos a um único grupo apenas. Os grupos de roupa não são necessariamente mutuamente exclusivos.

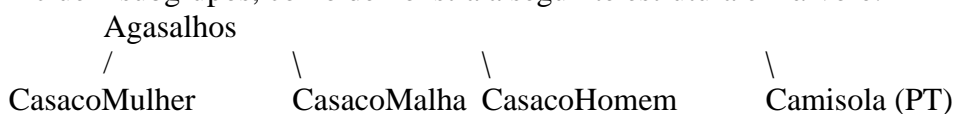
4. identificação dos casos em que são homónimas com palavras com sentidos diferentes, tal como *conjunto*, *sobretudo*, *laço*, *capa* (de livro) etc.
5. documentação, no presente documento, das variadas opções e dúvidas que foram surgindo
6. detecção, nos próprios corpos, de outras expressões, sobretudo multipalavra
7. criação de regras simples de correcção ou de anotação dos corpos com base no contexto lexical, morfossintáctico e semântico usando o programa corte-e-costura (Mota & Santos, 2009, Santos & Mota, 2010)

3. Critérios linguísticos

3.1 Opções iniciais de marcação da roupa

Para efectuar a marcação da roupa procedemos ao levantamento das palavras que indicavam roupa, usando como referência tudo o que nos parecesse ser indicativo de algo usado para ser vestido ou calçado pelo homem ou mulher, assim como por crianças ou bebés.

Tentámos usar os grupos já existentes, mas por vezes pareceu-nos necessário criar novos, e também agregar alguns dos grupos já existentes. Para não destruir a ordem e o trabalho já feito, esses grupos passam/passaram a ser supergrupos que incluem subgrupos, como demonstra a seguinte estrutura em árvore:



A motivação para isto foi variada, mas deve-se à nossa convicção de que existem palavras em português que se referem simplesmente aos níveis superiores e não apenas às folhas da árvore de classificação, e que por isso faz sentido incluir termos em níveis superiores da hierarquia. Exemplos são as próprias palavras *abafos* ou *agasalhos*.

Contamos assim com seis supergrupos, a negrito para mais fácil identificação: **AGASALHOS** (que inclui os grupos CASACOMULHER, CASACOMALHA, CASACOHOMEM, CAMISOLA, BLUSÃO, CASACOQUENTEINVERNO, CASACOIMPERMEÁVEL), **CALÇAS** (CALÇASJUSTASMULHER, CALÇASCURTAS, JEANS), **CASACOSCURTOS** (CASACOCURTOHOMEM, CASACOCURTOMULHER, JAQUETA), **EXTERIORTRONCO** (CAMISAHOMEM, BLUSAMULHER, T-Shirt), **VESTIDOOU SAIA** (VESTIDOMULHER, SAIAMULHER) e **MÚLTIPLA** (CONJUNTOMULHER, FATOHOMEM).

3.2 Novas marcações em torno da roupa

Ao executar este trabalho, um efeito secundário foi o de coligir dois novos grupos ligados a roupa: MATERIAIS e ELEMENTOSROUPA, com o possível intuito de poderem mais tarde ser objecto de estudo e/ou permitirem ajudar na desambiguação.

No grupo MATERIAIS incluíram-se os tecidos e materiais de que é feita a roupa (tal como *algodão*, *tafetá*, *seda*, *veludo*...), ao passo que o grupo ELEMENTOSROUPA engloba todos os elementos que integram a roupa mas que não chegam a sê-lo (tais como *colarinho*, *gola*, *presilha*, *zíper*,...).

3.3 Outra distinção: roupa para quem?

Também foi adicionada uma nova classificação que permitirá pesquisar a roupa consoante seja de mulher, de homem ou unissexo. Para isso foram criadas as classes

transversais: roupa:mulher, roupa:homem, roupa:criança, e roupa:unissexo.

De um ponto de vista de anotação isto significa que, em vez de uma árvore, temos um grafo/reticulado para o campo do vestuário.

3.4 Dúvidas que foram surgindo

O que fazer em relação a verbos associados ao campo do vestuário, tal como *vestir*, *calçar*, etc.? Por agora não os marcamos.

Também campos relativos a objectos pessoais associados ao vestir, exemplificados por palavras como *bengalas*, *guarda-chuvas*, *malas*, ou *relógios*, não foram marcados.

Excepção por agora feita ao campo jóias: *colares*, *pulseiras*, *coroas*, *tiaras* e *broches* foram marcados como ADEREÇOS, mas possivelmente isso será mudado se nos parecer inadequado.

Em muitos casos é através de metonímia que nos chegam novos termos, e isso é também operacional no campo do vestuário, veja-se os casacos e outras peles, com *raposa* e *arminho*. Um caso metonímico produtivo é o das marcas (designando peças de roupa ou conjuntos de peças): *Levi's*, *Chanel*, etc. Por agora não os marcamos, mas iremos investigar o assunto.

Finalmente, é possível que, tal como no campo semântico da cor, seja preciso marcar algumas designações de roupa como pertencendo a um grupo MÚLTIPLA; *conjunto-saia e casaco*, *saia e casaco*, *três-peças*, *conjunto gilet roxo e camisa amarela*, etc. Da mesma forma, teremos de resolver como marcar expressões metafóricas envolvendo roupa, tal como *de se lhe tirar o chapéu* e *pôr-se nas suas tamanquinhas*, que provavelmente merecerão o rótulo roupa:original. Eventualmente e numa segunda fase poderemos também investigar se faz sentido separar expressões fixas de expressões metafóricas que não sejam expressões idiomáticas, eventualmente com os traços idiomático vs. metafórico.

4. Questões técnicas e de utilização

Em conjunto com os atributos da análise sintáctica, podem ser utilizados nas pesquisas: o campo semântico *sema*² (com que podem ser pesquisadas as várias classes da roupa) e o campo com a informação sobre a roupa chamado *grupo* (com que podem ser pesquisados os diversos grupos de roupa).

4.1 Exemplos de procuras

Para explicar como ter acesso à informação de roupa presente no CONDIV, os utilizadores têm à disposição pesquisas tais como:

```
[sema="roupa.*"]  
[sema="roupa" & grupo="Saia.*"]3  
[sema="roupa:mulher"]  
[grupo="CamisaHomem.*"]
```

² Uma clarificação terminológica interna do AC/DC: O nome *sema* é apenas uma abreviatura de *semântica*, para distinguir do nome *sem* que indica o *semestre* (no caso dos corpos separados por este) e não tem qualquer relação com o termo *sema* da linguística estrutural.

³ Os nomes dos grupos têm a letra inicial maiúscula, para indicar que correspondem a uma classificação e não às formas ou lemas respectivos.

Para procurar casos de expressões com várias palavras que signifiquem roupa, uma forma possível de procurar será especificar:

```
<mwe> [sema="roupa.*"] []* </mwe>
```

visto que é sempre a primeira palavra da expressão que está marcada com os atributos sema e grupo.

Por outro lado, para procurar uma expressão específica, é possível estabelecer

```
<mwe> "casaco" "comprido" </mwe>
```

Para ver todas as expressões com mais de uma palavra que incluam roupa, em contexto, a expressão de procura ideal será:

```
[sema="roupa.*"] within mwe expand to mwe
```

Se apenas pretender expressões com mais de uma palavra que elas próprias sejam roupa, então a procura será

```
<mwe> [sema="roupa"] expand to mwe
```

5. Listagem das palavras de roupa em português e sua classificação

Todas as palavras e expressões de roupa identificadas como tal nos corpos do projecto AC/DC encontram-se no seguinte conjunto de ficheiros, disponíveis em:

- http://www.linguateca.pt/acesso/roupa_N.txt (palavras simples que são roupa quando substantivos)
- http://www.linguateca.pt/acesso/roupa_mwe.txt (expressões com várias palavras que denotam roupa)
- http://www.linguateca.pt/acesso/Grupos_roupa.txt (classificação das expressões anteriores em grupos ou classes)

Todos estes ficheiros vão sendo actualizados à medida que o trabalho prossegue. Prevê-se além disso a criação de novos ficheiros, por exemplo intitulados roupa_A.txt. para dar conta de adjectivos como *blusado*, *agasalhante* ou *evasée* quando não modificam palavras de roupa.

6. Resumo quantitativo

Aqui indicaremos quantas palavras referentes ao vestuário em cada corpo do AC/DC.

Agradecimentos

Este texto foi iniciado com a colaboração e co-autoria de Rosário Silva, que passou mais tarde a dedicar-se a outros campos semânticos e transferiu o trabalho da roupa para Cristina Mota.

O projeto AC/DC é uma das atividades da Linguateca, co-financiada pelo Governo português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, pela UMIC e pela FCCN.

Referências

[Costa et al. 2009]

Luís Costa, Diana Santos & Paulo Alexandre Rocha. "Estudando o português tal como é usado: o serviço AC/DC", *STIL 2009, The 7th Brazilian Symposium in Information and Human Language Technology* (São Carlos, Brasil, 8-11 de Setembro de 2009).

[Geeraerts & Grondelaers 1999]

- Geeraerts, Dirk & Stefan Grondelaers. "Purism and fashion. French influence on Belgian and Netherlandic Dutch". *Belgian Journal of Linguistics* **13** (1999), pp. 53-68.
- [Geeraerts, Grondelaers & Speelman 1999]
Geeraerts, Dirk, Stefan Grondelaers, and Dirk Speelman. *Convergentie en divergentie in de Nederlandse woordenschat*. Amsterdam: Meertens Instituut, 1999.
- [Mota & Santos 2009]
Cristina Mota & Diana Santos. "Corte e costura no AC/DC: auxiliando a melhoria da anotação nos corpos". Setembro de 2009. <http://www.linguateca.pt/acesso/corte-e-costura.pdf>
- [Santos 2010]
Diana Santos. "Linguatca's infrastructure for Portuguese and how it allows the detailed study of language varieties". In *OSLa 2*, 2010, no prelo.
- [Santos & Mota 2010]
Diana Santos & Cristina Mota. "Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora". In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)* (Valletta, Malta, 17-23 de Maio de 2010), European Language Resources Association, pp. 1437-1444.
- [Santos & Sarmiento 2003]
Diana Santos & Luís Sarmiento. "O projecto AC/DC: acesso a corpora/disponibilização de corpora". In Amália Mendes & Tiago Freitas (eds.), *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)* (Porto, 2-4 de Outubro de 2002), Lisboa : APL, pp. 705-717.
- [Silva 2006a]
Augusto Soares da Silva. "Convergência e divergência no léxico do Português Europeu e do Português Brasileiro: resultados do estudo sobre termos de futebol e de moda". In Joaquim Barbosa & Fátima Oliveira (eds.), *Textos seleccionados do XXI Encontro da Associação Portuguesa de Linguística, Setembro de 2005*, Lisboa: Colibri, pp. 633-646.
- [Silva 2006b]
Augusto Soares da Silva. "Sociolinguística cognitiva e o estudo da convergência/divergência entre o Português Europeu e o Português Brasileiro". *Veredas : Revista de Estudos Lingüísticos* **10** (2006). Universidade Federal de Juiz de Fora. <http://www.revistaveredas.ufjf.br>
- [Silva 2008a]
Augusto Soares da Silva "O corpus CONDIV e o estudo da convergência e divergência entre variedades do português". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguatca / Actas do encontro Linguatca: 10 anos*. Linguatca, 2008, pp. 25-28.
- [Silva 2008b]
Augusto Soares da Silva. "Integrando a variação social e métodos quantitativos na investigação sobre linguagem e cognição: para uma sociolinguística cognitiva do português europeu e brasileiro", *Revista de Estudos Lingüísticos* **16**, n. 1, Belo Horizonte, jan./jun. 2008, pp. 49-81.
- [Silva & Santos em edição permanente]

Rosário Silva & Diana Santos. "Arco-íris: notas sobre a anotação do campo semântico da cor em português". Primeira edição: 25 de Junho de 2009.
<http://www.linguateca.pt/acesso/ArcoIris.pdf>

[Silva et al. 2008]

Rosário Silva, Susana Inácio & Diana Santos. "Documentação da anotação relativa à cor no COMPARA". 31 de Dezembro de 2008.
<http://www.linguateca.pt/COMPARA/DocAnotacaoCorCOMPARA.pdf>

Anexo 1: Termos de vestuário – 22 perfis (CONDIVport)

- BLUSA F: *blouse, blusa, blusinha, bustier, camisa, camisa-body, camisão, camiseiro(inho), camiseta/e, (blusa) chemisier, (blusa) chemisiê*
- BLUSÃO M/F: *blazer, blêizer, blusão, bluson, camurça, camurcine, camisa esporte, casaco de pele, ganga, etc., colete, parka*
- CALÇAS M/F: *calça, calças, pantalone*
- CALÇAS CURTAS M/F: *bermuda(s), calças-capri, calça(s) corsário, calça(s) curta(s), calças 3/4, calções, cool pants, corsários, hot pants, knikers, pantacourt, pedal pusher, short(s), short cuts, short shorts, shortinho, slack(s)*
- CALÇAS JUSTAS F: *fuseau(x), fusô, legging(s)*
- CAMISA M: *blusão, camisa, camisa de gravata, camisa de manga curta, camisa desportiva, camisa esporte(iva), camisa jeans, camisa social, camiseta, camisete, camisette, camisinha*
- CAMISOLA M/F: *blusa, blusão, blusinha, body, cachemir, camisa, camisa-de-meia, camiseta, camisinha, camisola, camisolinha, canoutier, canoutiê, malha, malhinha, moleton, pull, pullover, pulôver, suéter, sweat, sweat shirt, sweater*
- CASACO F: *blazer, blêizer, casaco, casaquinho/a, manteau, mantô, paletó, paletot*
- CASACO M: *blazer, blêizer, casaco, paletó, paletot*
- CASACO CURTO F: *bolero, carmona, casa(i)b(v)equê, casaco curto, casaquilha, colete, colete camiseiro, corpete, corpinho, garibáldi, gilet, manguito, mini, minicasaco, roupinha, shortie, vasquinha*
- CASACO CURTO M: *casaco curto, colete, espartilho, gibão, gilet, jaleca, jaleco, jaqueta, véstia*
- CASACO DE CERIMÓNIA M/F: *black-tie, casaca, casaco cerimónia, fraque, manteau, mantô, paletó, paletot, pelerine, smo(c)king, sobrecasaca, tuxedo*
- CASACO DE MALHA M/F: *cardigã, cardigan, casaco/casaquinho de malha (de lã, de tricô), gilet, japona, malha, twin-set*
- CASACO IMPERMEÁVEL M/F: *ciré, ciré-maxi, anorak, canadiana, capa, capa de chuva, casaco impermeável, corta-vento, casaco-gabardina, gabardine/a, impermeável, kispo, parka, redingote*
- CASACO QUENTE (Inverno) M/F: *abafo, agasalho, balandrau, capote, casacão, casaco comprido, casaco de abafo/abafar, casaco de agasalho, casaco de/em pele, casaco-sobretudo, duffle-coat, gabão, gilet, manteau, mantô, manto, overcoat, paletó, pardessus, pelerine, samarra, sobrecasaca, sobretudo, sobreveste, trench (coat)*
- CONJUNTO F: *complet, completo, conjunto, costume, duas-peças, ensemble, fatinho, fato, saia-casaco, tailleur, toilette, toilette, vestido-casaco*
- FATO M: *beca, completo, costume, fato, terno*
- JAQUETA M/F: *casaca, casaco curto, jaleca, jaqueta, jaquette, jaquetinha, véstia*
- JEANS M/F: *calça(s) de ganga, calça(s) em denim, calça(s) em jeans, ganga, jeans*
- SAIA F: *kilt, maxi (máxi), maxissaia, micro-mini, micro-saia, míni (mini), mini-saia, minissaia, pareô, saia, saia-calça, saia-calção, saião, sainha, saiote*
- T-SHIRT M/F: *camisa, camiseta/e, camisette, camisola, licra, singlet, tee-shirt, t-shirt*
- VESTIDO F: *camiseiro, chemisier, chemisiê, shirt-dress, traje/o, veste, vestido(inho), vestido-camisa, vestido-camiseiro, vestido-camiseta, vestido-chemiser(ê), (vestido) cai-cai, (vestido) tomara-que-caia*

Índice

1. Introdução.....	1
2. Primeiros passos	2
3. Critérios linguísticos.....	3
3.1 Opções iniciais de marcação da roupa.....	3
3.2 Novas marcações em torno da roupa.....	3
3.3 Outra distinção: roupa para quem?.....	3
3.4 Dúvidas que foram surgindo	4
4. Questões técnicas e de utilização.....	4
4.1 Exemplos de procuras.....	4
5. Listagem das palavras de roupa em português e sua classificação.....	5
6. Resumo quantitativo	5
Agradecimentos	5
Referências	5
Anexo 1: Termos de vestuário – 22 perfis (CONDIVport).....	8
Índice	9