OSLa Simões, Barreiro, Santos, Sousa-Silva & Tagnin (eds.) *Linguística, Informática e Tradução: Mundos que se Cruzam*, Oslo Studies in Language 7(1), 2015. 425–438. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

SENTILEX-PT: PRINCIPAIS CARACTERÍSTICAS E POTENCIALIDADES

PAULA CARVALHO E MÁRIO J. SILVA

ABSTRACT

This paper describes the main characteristics of *SentiLex-PT*, a sentiment lexicon designed for the extraction of sentiment and opinion about human entities in Portuguese texts. The potential of this resource is illustrated on its application to two types of corpora, the *SentiCorpus-PT*, a social media corpus, consisting of user comments to news articles, and a literary piece of the early twentieth century, *The Poor* (Os Pobres), by Raul Brandão. The data were processed by UNITEX, a natural language processing system based on dictionaries and grammars.

[1] INTRODUÇÃO

A análise automática de sentimento (também designada na literatura como prospeção de opinião) dedica-se ao tratamento computacional de opiniões, sentimentos e atitudes, expressos em textos provenientes de origens diversas, em particular dos media sociais (Liu 2015). As aplicações que tiram partido desta análise baseiam-se, geralmente, em léxicos de sentimento, isto é, léxicos cujas entradas podem ser utilizadas para veicular um determinado sentimento ou emoção. Em geral, a informação de sentimento descrita nestes recursos corresponde à orientação semântica ou polaridade das palavras ou expressões. Neste âmbito, os traços mais comummente utilizados são os de *negativo*, *positivo* e *neutro*. A última categoria tem vindo a ser adotada para descrever os casos em que o sentimento associado a uma determinada expressão não é claramente positivo ou negativo, dependendo fundamentalmente do contexto (sintático, semântico e discursivo) em que é utilizado (e.g. *uma subida surpreendente* vs. *uma queda surpreendente*).

Uma das propriedades das línguas naturais é a ambiguidade ou, numa perspetiva mais abrangente, a vagueza (Santos 1998). Ao nível do sentimento, uma mesma palavra pode apresentar polaridades distintas em função do domínio em que ocorre, o que tem motivado a apresentação de abordagens para a construção de léxicos de domínios específicos (e.g. Zhang & Singh 2014). É, por exemplo, o caso de *quente*, que, na qualidade de modificador de um nome comestível como *sopa*, pode ser analisado como um predicador positivo (e.g. *A sopa ainda está quente*); porém, quando aplicado a um nome bebível como *champanhe*, veicula uma polaridade contrária (e.g. *O champanhe está quente*).

O *SentiLex-PT* é um léxico de sentimento especificamente concebido para a análise de sentimento e opinião sobre entidades humanas em textos redigidos em português. Trata-se de um recurso pioneiro para esta língua, sendo atualmente constituído por 7.014 lemas e 82.347 formas flexionadas.[1] As entradas adjetivais deste dicionário foram semiautomaticamente coligidas e classificadas, combinando uma abordagem linguística, para extrair candidatos a adjetivos humanos a partir de *corpora*, e uma abordagem de aprendizagem automática, para filtrar os adjetivos humanos a partir da lista de candidatos. A polaridade desses adjetivos foi atribuída com base num cálculo sobre as distâncias das palavras, com polaridade conhecida *a priori*, ligadas aos adjetivos por uma relação de sinonímia num grafo, inferido a partir de dicionários de sinónimos disponíveis para o português (Silva et al. 2012).

Embora existam atualmente alguns léxicos de sentimento para o processamento de texto em português (e.g. Balage et al. 2013; Freitas 2013; Souza et al. 2011), na altura em que o *SentiLex* foi concebido, não existiam ou não estavam disponíveis dicionários com estas características para esta língua, embora existissem para outras, em particular, o inglês (Hu et al. 2004; Wilson et al. 2005). De ressalvar, no entanto, que a análise linguística das emoções em português é uma temática que tem vindo a ser aprofundadamente investigada na literatura, destacando-se, entre outros, os trabalhos de Maia (1994/1996), Mendes (2004) e, mais recentemente, de Santos & Mota (2015).

Passados alguns anos desde a sua disponibilização, o *SentiLex* continua a ser um recurso inovador, distinguindo-se dos restantes léxicos por não ter a ambição de ser um dicionário geral, nem tão pouco um dicionário referente a um domínio específico. Trata-se, antes, de um léxico sintático-semântico, orientado não pelo domínio semântico em que as entradas podem ocorrer, mas pelas restrições sintáticas que os seus predicadores impõem.

As principais potencialidades de utilização deste léxico estão, pois, intimamente relacionadas com o número e a natureza dos atributos que descreve. Cada uma das suas entradas (adjetivos, verbos, nomes e expressões idiomáticas de natureza verbal) tem a propriedade de poder ser utilizada como predicador humano, isto é, exercer modificação sobre um nome de natureza humana, e é (apenas) esse o uso que está contemplado no dicionário. Em particular, cada entrada contém informação sobre:

- A natureza sintática do predicador (transitivo ou intransitivo);
- A natureza semântica dos argumentos, sobre os quais recai o sentimento (para já, apenas está contemplada a categoria de humano, mas, a qualquer momento, é possível incluir outras categorias semânticas);

[1] A primeira versão do léxico foi disponibilizada ainda em 2010 (SentiLex-PT01). A versão atualmente disponível pode ser obtida em: http://dmir.inesc-id.pt/project/SentiLex-PT_02.

- A polaridade do predicador, tendo em consideração o alvo que este modifica;
- O método de atribuição de polaridade (manual ou automático);
- A informação de lema e respetivas formas flexionadas.

A informação de polaridade associada às entradas foi, na maioria dos casos, manualmente atribuída. Certas entradas adjetivais foram, contudo, automaticamente classificadas por uma ferramenta (denominada JALC) desenvolvida para este fim, como anteriormente referido. As formas flexionadas dos verbos e das expressões idiomáticas, bem como os respetivos atributos morfológicos, foram extraídos semiautomaticamente do LABEL-LEX, um léxico de palavras simples desenvolvido pela equipa do LabEL para o português (Ranchhod et al. 1999).

[2] PROPRIEDADES DO SENTILEX-PT

As entradas do léxico correspondem a predicadores humanos, i.e. adjetivos, nomes, verbos e expressões idiomáticas de base verbal com a particularidade de se construir com nomes humanos, elementos nucleares de grupos nominais que, numa frase, podem desempenhar a função de sujeito ou de complemento. É, por exemplo, o caso de *frágil*, que, além de poder exercer modificação sobre um nome concreto (e.g. *cobertura frágil*) ou abstrato (e.g. *posição frágil*), também pode selecionar um nome de natureza humana, sobre o qual exerce modificação (e.g. *indivíduo frágil*). O adjetivo em análise veicula uma polaridade negativa, qualquer que seja a natureza do nome com que se combina.

Contudo, há outros casos em que a polaridade do predicador poderá diferir em função da especificação sintático-semântica dos argumentos com que este se constrói. Por exemplo, o adjetivo *gordo* veicula tipicamente um valor negativo, enquanto modificador de um nome de natureza humana (e.g. *indivíduo gordo*), mas pode assumir uma polaridade contrária, quando combinado com um nome como, por exemplo, *salário* (e.g. *salário gordo*).

Há ainda outros casos em que uma mesma forma poderá, em função da construção em que surge, ser, ou não, interpretado como um predicador de sentimento. Por exemplo, o adjetivo *distinto* deverá ser classificado como um predicador de sentimento, com polaridade positiva, quando combinado com sujeitos de natureza humana (e.g. *médico distinto*); contudo, em construções não humanas, a mesma forma poderá não veicular qualquer sentimento e/ou polaridade (e.g. *estratégias distintas*).

Assim, no desenvolvimento de qualquer léxico, em particular os de sentimento, deve ter-se em consideração os diferentes contextos sintático-semânticos em que as palavras podem ocorrer, para que a descrição das entradas seja o mais fiel possível, potenciando, desse modo, a sua aplicabilidade em tarefas de processamento.

Foi com base neste princípio que o *SentiLex* foi construído. Apenas as construções que selecionam como argumentos um nome de natureza humana foram consideradas no léxico. Há, portanto, termos de sentimento comuns na língua que, por serem meros modificadores de nomes não humanos, não estão contemplados neste léxico (e.g. *nítido*); pelo contrário, outros predicadores, como os ilustrados anteriormente, encontram-se atestados, apesar da sua ambiguidade inerente. De referir, contudo, que nesses casos, apenas a construção humana, objeto da nossa análise, se encontra atestada no *SentiLex-PT*.

Mesmo restringindo as entradas do *SentiLex* a predicadores humanos, é, ainda assim, possível registar entradas estruturalmente ambíguas. De facto, uma mesma forma pode ser encontrada em estruturas sintáticas distintas; isto é, predicadores homógrafos podem apresentar redes argumentais diferentes, distinguindo-se pelo número e tipo de argumentos que selecionam. É, por exemplo, o caso do adjetivo *responsável*, que pode ser simultaneamente encontrado em construções intransitivas e transitivas. No primeiro caso, o adjetivo, que se constrói com um sujeito de natureza humana, tem polaridade positiva (e.g. *Ele é uma pessoa responsável*). Na construção transitiva, o adjetivo seleciona, além do sujeito, um outro argumento, que ocupa a função de complemento, introduzido pela preposição *por* (e.g. *Ele é responsável por esse incidente*). Neste último caso, o adjetivo pode ser substituído por *culpado*, que detém um valor negativo.

[2.1] *O formato das entradas no SentiLex-lem-PT02*

O *SentiLex-PT* tem dois dicionários associados: um que descreve os lemas (ilustrado na Figura 1) e o correspondente de formas flexionadas (ilustrado na Figura 2). No dicionário de lemas, cada linha inclui informação sobre:

- Lema (convencionalmente a forma masculina do singular para os adjetivos, a forma singular para os nomes que flexionam em número e a forma infinitiva para os verbos e expressões idiomáticas);
- Categoria gramatical (ADJetivo, Nome, Verbo and IDIOMa);
- Atributos de sentimento:
  - Polaridade (POL), a qual pode ser positiva (1), negativa (-1) ou neutra (0);
  - Alvo da polaridade (TG), o qual corresponde a um nome de tipo humano (HUM), com função de sujeito (N0) e/ou complemento (N1);
  - Classificação de polaridade (ANOT), a qual pode ter sido manualmente (MAN) ou automaticamente atribuída, pela ferramenta JALC.

```
aberração.PoS=N;TG=HUM:N0;POL:N0=-1;ANOT=MAN
bonito.PoS=Adj;TG=HUM:N0;POL:N0=1;ANOT=MAN
castigado;PoS=Adj;TG=HUM:N0;POL:N0=-1;ANOT=JALC
estimado.PoS=Adj;TG=HUM:N0;POL:N0=1;ANOT=JALC;REV=AMB
enganar.PoS=V;TG=HUM:N0:N1;POL:N0=-1;POL:N1=0;ANOT=MAN
engolir em seco.PoS=IDIOM;TG=HUM:N0;POL:N0=-1;ANOT=MAN
```

FIGURA 1: Exemplos de entradas do *SentiLex-lem-PT02* (lemas).

```
aberração,aberração.PoS=N;FLEX=fs;TG=HUM:N0;POL:N0=-1;ANOT=MAN
bonita,bonito.PoS=Adj;FLEX=fs;TG=HUM:N0;POL:N0=1;ANOT=MAN
bonitas,bonito.PoS=Adj;FLEX=fp;TG=HUM:N0;POL:N0=1;ANOT=MAN
bonito,bonito.PoS=Adj;FLEX=ms;TG=HUM:N0;POL:N0=1;ANOT=MAN
bonitos,bonito.PoS=Adj;FLEX=mp;TG=HUM:N0;POL:N0=1;ANOT=MAN
engoliste em seco,engolir em seco.PoS=IDIOM;Flex=J2p|J2s;TG=HUM:N0;POL:N0=-1;ANOT=MAN
engolistes em seco,engolir em seco.PoS=IDIOM;Flex=J2p;TG=HUM:N0;POL:N0=-1;ANOT=MAN
engoliu em seco,engolir em seco.PoS=IDIOM;Flex=J4s|P3s;TG=HUM:N0;POL:N0=-1;ANOT=MAN
engulamos em seco,engolir em seco.PoS=IDIOM;Flex=Y1p|S1p;TG=HUM:N0;POL:N0=-1;ANOT=MAN
```

FIGURA 2: Exemplos de entradas do *SentiLex-lem-PT02* (formas flexionadas).

No dicionário de formas flexionadas, as entradas estão associadas ao respetivo lema. Neste formato, além das informações descritas no dicionário de lemas, as entradas adjetivais e nominais contêm informação sobre a flexão (FLEX) em género (masculino (m) ou feminino (f)) e número (singular (s) ou plural (p)). Os atributos morfológicos associados aos verbos e expressões idiomáticas incluem informação de tempo, pessoa e número, os quais foram automaticamente extraídos do dicionário LABEL-LEX.

[3] ALGUMAS ESTATÍSTICAS SOBRE O SENTILEX

A maioria das entradas do *SentiLex-PT* corresponde a predicadores intransitivos, contando atualmente com 6.627 construções intransitivas e 456 construções transitivas.

No que respeita à categoria gramatical das entradas, o léxico descreve maioritariamente adjetivos, mas também contempla nomes e verbos predicativos, assim como expressões idiomáticas de base verbal (cf. Figura 3).

Relativamente à distribuição de polaridade, observa-se que a maioria dos predicadores contemplados no léxico (67%) apresenta polaridade negativa (cf. Tabela 1). No caso dos predicadores transitivos, a classe mais representativa envolve a construção com um sujeito positivo e um complemento negativo (162 entradas), seguida da construção com um sujeito neutro e um complemento negativo (cf. Tabela 2).

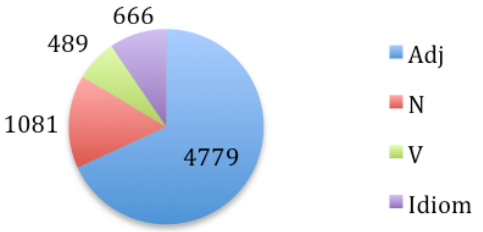

FIGURA 3: Distribuição dos lemas no *SentiLex* por categoria gramatical.

Polaridade	Nº de Predicadores Intransitivos	Exemplos
Negativo	4.453	arrogante; terror; morrer; não ter onde cair morto
Positivo	1.396	misericordioso; beleza; brilhar; levantar a cabeça
Neutro	1.396	misterioso; simples; humilde; ingénuo

TABELA 1: Distribuição da polaridade associada aos predicadores intransitivos no *SentiLex*

[4] EXPLORAÇÃO DAS INFORMAÇÕES DO SENTILEX EM CORPORA

Com o objetivo de ilustrar a utilidade das informações representadas no *SentiLex*, comparamos dois corpora distintos, tanto no que se refere ao género textual como à intenção comunicativa: o *SentiCorpus-PT*, um *corpus* proveniente dos media sociais, constituído por comentários de utilizadores a artigos noticiosos no âmbito da política[2], e *Os Pobres*, uma obra literária da autoria de Raul Brandão, datada do início do século XX.[3] Para criar o *SentiCorpus*, compilámos uma coleção de comentários, escritos por leitores da edição *online* do jornal Público, aos dez artigos que cobriram os debates políticos que antecederam as eleições legislativas portuguesas de 2009. A coleção é composta por 2.795 comentários (cerca de 8.000 frases), os quais se encontram associados aos respetivos artigos de notícia (Carvalho et al. 2011).

Ambos os textos foram processados pelo Unitex, um sistema concebido para o processamento de textos de grandes dimensões, em várias línguas (Paumier 2003). Nesta aplicação, todos os recursos linguísticos são internamente representados por transdutores de estados finitos, como é o caso dos dicionários e das gramáticas para análise morfológica ou sintática. A tabela 3 apresenta algumas estatísticas extraídas a partir dos referidos textos.

[2] O *SentiCorpus* está disponível em: http://dmir.inesc-id.pt/project/SentiCorpus-PT.

[3] Texto disponível em: http://www3.universia.com.br/conteudo/literatura/Os_pobres_de_raul_germano_brandao.pdf.

Polaridade		Nº de Predicadores Transitivos	Exemplos
N0_Pos	N1_Pos	1	estar à altura
N0_Pos	N1_Neg	162	calar, vencer
N0_Pos	N1_Neu	29	impressionar, salvar
N0_Neg	N1_Neg	22	encobrir, insultar
N0_Neg	N1_Pos	9	ceder, curvar-se
N0_Neg	N1_Neu	149	espezinhar, faltar ao respeito
N0_Neu	N1_Neg	55	desconfiar, ignorar
N0_Neu	N1_Pos	29	admirar, acreditar

TABELA 2: Distribuição da polaridade associada aos predicadores transitivos no *SentiLex*

	Os Pobres		SentiCorpus-PT	
Nº de tokens (tokens diferentes)	109.169	(7.451)	112.374	(8.591)
Nº palavras (palavras diferentes)	45.803	(7.423)	49.304	(8.544)
Nº palavras de sentimento (palavras diferentes)	4.575	(1.700)	4.548	(1.805)
Nº palavras positivas (palavras diferentes)	1.213	(426)	2.131	(650)
Nº palavras negativas (palavras diferentes)	3.140	(1.122)	2.338	(998)
Nº palavras neutras (palavras diferentes)	444	(159)	446	(181)

TABELA 3: Estatísticas extraídas dos *corpora*

As estatísticas da Tabela 3 permitem concluir que, apesar de diferentes, o número de palavras de sentimento reconhecidas em ambos os corpora, após a aplicação do *SentiLex-PT*, não chega a perfazer 10% do número total de palavras no texto. De referir, no entanto, que este valor é pouco informativo, uma vez que: (i) o léxico aplicado apenas compreende predicadores de natureza humana, (ii) não temos a garantia de que as palavras identificadas estejam a ser utilizadas como verdadeiros predicadores de sentimento nos corpora em questão, e (iii) não temos um ponto de referência sobre a distribuição das palavras de sentimento nos *corpora*.

Em ambos os textos, as palavras ou expressões identificadas como podendo veicular sentimento correspondem a cerca de 25% do número total de palavras registadas no *SentiLex-PT*. Porém, no que respeita à distribuição de polaridade, observa-se que no corpus literário, *Os Pobres*, predomina o sentimento negativo. De facto, cerca de 70% das palavras identificadas estão classificadas como negativas e a variedade (ou riqueza) lexical é também mais expressiva neste caso. Esta evidência vai ao encontro do caráter marcadamente negativo da obra, classificada como "uma meditação sobre a metafísica da dor e sobre o absurdo da condição hu-

mana, dentro da qual as coordenadas de tempo, espaço, intriga ou personagens, apenas esboçadas, servem de cenário universal e abstrato para o drama secular da luta do homem entre o sonho e a desgraça".[4] Pelo contrário, a distribuição da polaridade no *SentiCorpus-PT* parece ser mais equilibrada. No entanto, seria necessário analisar em profundidade a distribuição das palavras no texto, para aferir a validade desta observação. Por um lado, não estamos a ter em consideração o contexto sintático onde as potenciais expressões de sentimento ocorrem. Não estamos a prever, a título de exemplo, a possibilidade de estas estarem sob o escopo da negação. De facto, as concordâncias abaixo ilustradas, obtidas a partir da pesquisa de termos potencialmente positivos modificados pelo advérbio de negação *não*, confirmam esse uso.

ira , mesmo até á exaustão ,	não a torna verdade. DULCE FORTES NOVO
(os outros não vi) , onde	não apresenta nenhuma ideia concreta pa
o foi 1ª ministra mas também	não dá garantias de competência para ti
o , há 5 anos a investigar ,	não diz que o homem é inocente ou não a
ra cima do Louçã. O Sócrates	não é capaz de falar do futuro. O Sócra
cargo a que se candidata? Se	não é capaz de aguentar com ritmo um de
campanha eu vir que a Manela	não é capaz de vencer o Sócrates , voto
agamento Especial por Conta?	Não é crível que BE ou PCP governassem
fia tem uma força brutal mas	não é invencível. A maioria dos Portugu
asa gasta. Jerórimo de Sousa	não é perfeito nem o seu partido , como
de Estado e da Defesa , logo	não é responsável pelos erros de toda a
O Portadas já demonstrou que	não é transparente , muto menos sério!
debate destes pois partidos	não elucidaram o povo Português nas pol
da GALP , questão a que o PM	não esclareceu. Mas é preciso para se j
pessoa que me fascina. A MFL	não está à altura das exigências do gov
! ! ! ! Viva Sócrates! Você	não está a ser honesta, só não sei se p

Pelo outro lado, os termos classificados como positivos podem estar a ser utilizados de forma não literal, por exemplo, para expressar ironia, um fenómeno extremamente produtivo em textos provenientes das redes sociais (Carvalho et al. 2009).

A Tabela 4 apresenta a lista das cinco palavras de sentimento do *SentiLex*, com maior número de ocorrências em cada um dos *corpora*.

É interessante verificar que as palavras em questão, que remetem diretamente para as temáticas retratadas em cada um dos textos, são diferentes. No texto literário, a palavra mais frequente, *sonho*, é a única descrita no *SentiLex* como positiva. Pelo contrário, nos comentários aos debates políticos, o lugar de destaque é ocupado por predicadores transitivos, cuja polaridade é potencialmente positiva para

[4] Excerto de texto extraído do Dicionário de Língua Portuguesa com Acordo Ortográfico [em linha]. Porto: Porto Editora, 2003–2015. [Data da consulta: 2015-02-13]. Disponível em http://www.infopedia.pt/\protect\char"0024\relaxraul-brandao.

Os Pobres	Polaridade	Ocorr.	SentiCorpus-PT	Polaridade	Ocorr.
sonho	Pos	109	votar	Neu Pos	91
desgraça	Neg	89	voto	Neu Pos	76
pobres	Neg	61	verdade	Pos	50
só	Neg	56	votos	Neu Pos	37
triste	Neg	50	ganhou	Pos Neg	35

TABELA 4: Lista de palavras mais frequentes nos *corpora*

um dos grupos nominais que desempenham a função sujeito ou de complemento, como *votar* (*em alguém*).

Uma análise mais aprofundada dos *corpora* permite concluir que mesmo as palavras positivas são frequentemente utilizadas em contexto negativo. Por exemplo, as concordâncias a seguir ilustradas mostram, por exemplo, que o modificador adjetival do nome *sonho*, no texto literário, é, na maioria dos casos, negativo, alterando, pois, a polaridade da construção nominal.

```
cobre, a secura dos outros, o sonho calcado e por terra, lágrimas e en
co duro, impenetrável.{S} É o sonho cativo num ovo hermético de bronze
 de beleza. {S}O universo é o sonho dolorido de Deus. {S}Nada se perde
is visível a sua aspiração, o sonho escondido e inútil.{S} Só o Gebo n
onólogos cheios de gritos, de sonho espezinhado, todos lavados em lágr
rio, não esquecem esse fio de sonho espezinhado, que ainda sentem corr
noite traga, como farrapos de sonho espezinhado... {S}Todas as noites
ta;{S} a desgraça gasta até o sonho grotesco dos humildes.{S} E elas c
}E não podia.{S} Porque até o sonho mesquinho dos desgraçados se estan
s enfermarias corre também um sonho parecido com luar.{S}.. Será uma f
onhecido ou descobrindo outro sonho tão vivo, que, de vê-lo, caíra ful
raro se aqueciam ainda com um sonho vão.{S} Fixavam o olhar, perdidos,
o, da ambição, da vaidade, do sonho vão, para quê?{S} Para ser desgraç
 nele entram também, como nos sonhos grandiosos, como em todos os dram
las-íeis revolvidas, homens e sonhos misturados, um rio que tudo acarr
, ei-lo que enternecido conta sonhos rotos e tristes, o sonho dos pobr
 eles botavam realmente flor, sonhos tristes, mealhas, almas que nem s
s os que são apenas restos de sonhos vivos e despedaçados como eu, têm
```

De facto, para que possamos potenciar a utilização da informação descrita nos léxicos de sentimento, é fundamental criar gramáticas que permitam interpretar e contextualizar essa informação.

Relativamente à distribuição da polaridade nos textos por categoria gramatical, observa-se que as formas adjetivais e verbais são as categorias com maior representatividade em ambos os corpora, seguidas das formas nominais e, finalmente, das expressões idiomáticas.

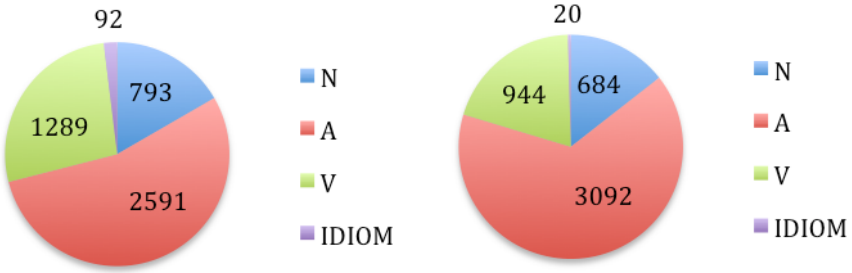

FIGURA 4: Distribuição da polaridade por categoria gramatical nos *SentiCorpus-PT* e *Os Pobres*, respetivamente

As expressões de sentimento multipalavra, embora menos expressivas em termos de representatividade no léxico, são menos ambíguas do ponto de vista lexical e, por isso, mais informativas, motivando um investimento no seu recenseamento e formalização. Este comportamento está ilustrado no extrato da concordância a seguir apresentado, extraído do *SentiCorpus-PT*.

SÓCRATES! Jovens do meu país abram os olhos e corram com esta cambada
que a imprensa de referência andou a dormir ou foi habilmente anestes
concebido que a pobre senhora andou aos bonés e a ver o comboio (TGV
undezas do inferno) . Tentou aplicar golpes baixos ao lider do BE , c
genharia como ele afirma ser. Batem no ceguinho , mas não têm proposta
governar. O BE vai sem dúvida captar eleitorado ao PS , os resultados
l. Votem neles e dpois venham chorar sobre o leite derramado. Votem no
u , questionou , barafustou e complicou a vida e o discurso de uma sen
ão , ... votar PS-socrático é dar um tiro no pé. O que eu não consigo
a carreira como os outros que davam o litro) é que não gostam porque
o mais comentários aparecem a denegrir a imagem de Manuela Ferreira Le
a sua primeira medida a tomar desviar dinheiro do SNS e segurança soci
.. E com razão , que Socrtaes deu cabo do partido E com razão. E como
a sempre. O operário jerónimo deu um baile ao falso engenheiro socas.
o acordo com o PP. O Socrates deu um baile à Manelita que está muito m
ldades é claro que o Sócrates deu um baile na Manuela. Para ser eleito
lo , a Manuela Ferreira Leite deu um banho ao Louçã que foi um espectá
em meu entender. Paulo Portas deu um banho de cultura e de inteligênci
chegou para ela ! .. . . Quem deu uma lição a essa senhora Manuela Mou
s ! ! ! Manuela com que então diz mal dos espanhois e quando eras deri
as muito verdadeiras e que só dizem a verdade , mas não passam de os m
a mentira. E quem anda aqui a dizer mal do Socrates , que vote noutro
COERÊNCIA DE DOIS LÍDERES. É dizer mal por dizer. E é patético o PM v
i muito melhor . . . Fica bem dizer mal ; alivia tensões , não é? Fico
istas a tomar a tomar chá e a dizerem mal do governo... uma peça de te
CÃ EM PAPEL SOLOFAN. Sócrates encostou à parede Louçã. Sócrates está a

urrículo , e ele cabisbaixo , engoliu em SECO! Toda a gente viu que o
s sem segurança. Este governo esbanjou dinheiro em áreas que não produ
e humanidade. Francisco Louçã está a anos luz de José Sócrates , com o
Pinocrates aldrabão Portugal está bem e recomenda-se. Para o Pinocrat
os ficar a saber que Portugal está de tanga , que os desfalques foram
ar até ao fim (6 , 7 % ?) , está nas mãos do meu amigo Zé do Vidoso.
la F. Leite complementam-se , estão bem um para o outro , podem-se cas
eitos e virtudes , demonstrou estar à altura do cargo de Primeiro Mini
e a Sócrates , acho que Louçã esteve bem ao ataque. Ao contrário do qu

Algumas das construções apresentadas nas concordâncias são transitivas, como é o caso da construção *encostar à parede*, apresentando dois valores de polaridade distintos: positivo para o argumento que desempenha a função de sujeito (no caso, *Sócrates*) e negativo para o argumento que desempenha a função de complemento direto (no caso, *Louçã*). De salientar que esta informação pode ser corretamente processada por aplicação do *SentiLex-PT* aos textos, dado que a informação de polaridade tem em conta as propriedades distribucionais dos predicadores, algo que é normalmente ignorado nos léxicos que têm vindo a ser construídos, tanto para o português como para outras línguas. Esta informação permite, por exemplo, tornar a extração de sentimento mais fina e rigorosa. Por exemplo, a concordância abaixo resulta do refinamento da pesquisa anterior, requerendo a presença de uma expressão idiomática, cuja polaridade é potencialmente positiva para o sujeito e negativa para o complemento do predicador.

u , questionou , barafustou e complicou a vida e o discurso de uma sen
.. E com razão , que Socrtaes deu cabo do partido E com razão. E como
a sempre. O operário jerónimo deu um baile ao falso engenheiro socas.
o acordo com o PP. O Socrates deu um baile à Manelita que está muito m
ldades é claro que o Sócrates deu um baile na Manuela. Para ser eleito
lo , a Manuela Ferreira Leite deu um banho ao Louçã que foi um espectá
em meu entender. Paulo Portas deu um banho de cultura e de inteligênci
chegou para ela ! .. . . Quem deu uma lição a essa senhora Manuela Mou
carro usado a Sócrates? Quem deu uma lição a essa senhora Manuela Mou
CÃ EM PAPEL SOLOFAN. Sócrates encostou à parede Louçã. Sócrates está a

[5] CONSIDERAÇÕES FINAIS

O *SentiLex-PT* é um recurso de acesso livre, que tem vindo a ser amplamente utilizado por equipas de investigação nacionais e internacionais, em diversas tarefas de expansão lexical (destacando-se, entre outros, o trabalho de Gonçalo Oliveira et al. 2014) e análise sentimento, por exemplo, no contexto político (Tumitan & Becker 2014).

No futuro, procuraremos disponibilizar uma nova versão, que incluirá informação estatística (extraída de *corpora* de grandes dimensões), que permita definir a probabilidade de um dado termo poder ser potencialmente utilizado como predicador de sentimento e, em particular, como predicador humano. Além disso, procuraremos refinar este dicionário, tirando partido de informações semânticas disponíveis noutros recursos, como é o caso do *Port4Nooj* (Barreiro 2008), e explorando redes de relações semânticas em bases de conhecimento baseadas na *WordNet* (cf. Rademaker et al. 2014).

AGRADECIMENTOS

Um agradecimento muito especial à Belinda Maia, corresponsável pelas duas Escolas de Verão organizadas pela Linguateca, onde tivemos a oportunidade de nos conhecer e de abraçar um projeto na área de análise de sentimento, de onde, entre outros recursos, nasceu o *SentiLex-PT*. Uma palavra de agradecimento também à Maria José Finnato e ao Hugo Gonçalo Oliveira, pela leitura do artigo e pertinentes sugestões.

O desenvolvimento deste trabalho foi parcialmente apoiado com financiamentos da Fundação para a Ciência e a Tecnologia (FCT), referências UID/CEC/50021/2013, EXCL/EEI- ESS/0257/2012 (DataStorm), PTDC/CPJ-CPO/116888/2010 (POPSTAR), UTA-Est/MAI/0006/2009 (REACTION) e SFRH/BPD/45416/2008.

REFERÊNCIAS

Balage, Pedro, Thiago Pardo & Sandra Aluísio. 2013. An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. Em *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 215–219.

Barreiro, Anabela. 2008. ParaMT: A paraphraser for machine translation. Em *Computational Processing of the Portuguese Language, 8th International Conference, PROPOR 2008, Aveiro, Portugal, September 8-10, 2008, Proceedings*, 202–211.

Carvalho, Paula, Luís Sarmento, Mário J. Silva & Eugénio de Oliveira. 2009. Clues for Detecting Irony in User-generated Contents: Oh...!! It's "So Easy";-). Em *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, 53–56. ACM.

Carvalho, Paula, Luís Sarmento, Jorge Teixeira & Mário J. Silva. 2011. Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates. Em *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, 564–568.

Freitas, Cláudia. 2013. Sobre a construção de um léxico da afetividade para o pro-

cessamento computacional do português. *Revista Brasileira de Linguística Aplicada* 13. 1031–1059.

Gonçalo Oliveira, Hugo, António Paulo Santos & Paulo Gomes. 2014. Assigning Polarity Automatically to the Synsets of a Wordnet-like Resource. Em Maria João Varanda Pereira, José Paulo Leal & Alberto Simões (eds.), *3rd Symposium on Languages, Applications and Technologies*, vol. 38, 169–184.

Hu, Minqing, & Bing Liu. 2004. Mining and summarizing customer reviews. Em *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177. ACM.

Liu, Bing. 2015. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge University Press.

Maia, Belinda. 1994/1996. *A Contribution to the Study of the Language of Emotion in English and Portuguese*: FLUP. Tese de Doutoramento. Versão revista: 1996.

Mendes, Amália. 2004. *Predicados Verbais Psicológicos do Português. Um contributo para o estudo da polissemia verbal*. FCT/Fundação Calouste Gulbenkian.

Paumier, Sébastian. 2003. Unitex 1.2. user manual. Relatório técnico. http://www-igm.univ-mlv.fr/~unitex/.

Rademaker, Alexandre, Valeria de Paiva, Gerard de Melo, Livy Maria Real Coelho & Maira Gatti. 2014. OpenWordNet-PT: A Project Report. Em Heili Orav, Christiane Fellbaum & Piek Vossen (eds.), *Proceedings of the 7th Global WordNet Conference*, 383–390.

Ranchhod, Elisabete, Cristina Mota & Jorge Baptista. 1999. A Computational Lexicon of Portuguese for Automatic Text Parsing. Em *SIGLEX99: Standardizing Lexical Resources*, s/pp. Association for Computational Linguistics.

Santos, Diana. 1998. A relevância da vagueza para a tradução, ilustrada com exemplos de inglês para português / The relevance of vagueness for translation: Examples from English to Portuguese. *TradTerm* 5. 41–78.

Santos, Diana & Cristina Mota. 2015. A admiração à luz dos corpos. OSLa: Oslo Studies in Language, Este volume.

Silva, Mário J., Paula Carvalho & Luís Sarmento. 2012. Building a sentiment lexicon for social judgement mining. Em Helena Caseli, Aline Villavicencio, António Teixeira & Fernando Perdigão (eds.), *Computational Processing of the Portuguese Language*, vol. 7243, 218–228. Springer.

Souza, Marlo, Renata Vieira, Débora Busetti, Rove Chishman & Isa Mara Alves. 2011. Construction of a Portuguese Opinion Lexicon from multiple resources. Em *In 8th Brazilian Symposium in Information and Human Language Technology*, 59–66.

Tumitan, Diego & Krin Becker. 2014. Sentiment-based features for predicting election polls: a case study on the brazilian scenario. Em *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 2, 126–133.

Wilson, Theresa, Janyce Wiebe & Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. Em *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347–354.

Zhang, Zhe & Munindar Singh. 2014. ReNew: A semi-supervised framework for generating domain-specific lexicons and sentiment analysis. Em *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics*, vol. 1, 542–551.

CONTACTOS

Paula Carvalho
Laureate International Universities & INESC-ID
pcc@inesc-id.pt

Mário J. Silva
Universidade de Lisboa, Instituto Superior Técnico & INESC-ID
mjs@inesc-id.pt