UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Relatório Técnico PPgSI-004/2016

Tratamento Lexical para Computação de Personalidade a partir de Textos

Ivandré Paraboni

Setembro - 2016

Série de Relatórios Técnicos

PPgSI-EACH-USP
Rua Arlindo Béttio, 1000 – Ermelino Matarazzo
03828-000 – São Paulo, SP.
TEL: (11) 3091-8197 http://www.each.usp.br/ppgsi

Tratamento Lexical para Computação de Personalidade a partir de Textos

Ivandré Paraboni[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo
São Paulo – SP, Brazil

{ivandre}@usp.br

Resumo. *Este documento descreve o pré-processamento de textos e a extração de atributos lexicais para aplicações de tratamento computacional da personalidade humana.*

1. Introdução

Este documento descreve o pré-processamento de textos e a extração de atributos lexicais para aplicações de tratamento computacional da personalidade humana. As funcionalidades aqui discutidas foram implementadas com uso de dicionários e bases lexicais auxiliares, algumas das quais provenientes de anotação manual. No decorrer deste relatório usamos o termo 'tratamento lexical' para designar estas operações, que englobam tanto a correção ortográfica como tratamento de itens não lexicais como termos compostos (e.g., 'hojeTemFesta') e diversas construções típicas da linguagem das redes sociais (e.g., 'kkkk', 'liiinda', 'blz').

2. Entradas e Saídas

A entrada para o tratamento lexical é uma coleção de arquivos-textos, cada qual representando os dados de um indivíduo (autor ou locutor do mesmo). Os textos em questão são tipicamente de dois tipos: atualizações de *status* da rede social Facebook, ou descrições textuais produzidas em resposta a determinados estímulos visuais de interesse para projetos de Geração de Língua Natural (GLN), especialmente para construção de córpus contendo pares estímulos-descrições (e.g., Gatt et al. [2007], Dale e Viethen [2009], Viethen e Dale [2011], Teixeira et al. [2014], Paraboni et al. [2016]).

As atualizações da rede social são textos de natureza pessoal e não estruturada. As descrições textuais, por outro lado, são tipicamente de natureza impessoal e consideravelmente mais estruturada. Do ponto de vista da tarefa aqui descrita, entretanto, não há distinção entre estas modalidades de entrada.

A saída do tratamento lexical realizado consiste de quatro tipos de arquivos *s1..s4* representando o texto original após o tratamento ortográfico e normalização, e arquivos de estatísticas lexicais e vocabulários. Estes arquivos objetivam apoiar a construção de modelos computacionais para tratamento computacional da personalidade a partir de texto de diversas formas, conforme será discutido nas seções seguintes.

- *s1* Os textos após correção ortográfica e normalização, com a substituição de certas expressões por símbolos genéricos. Por exemplo, nomes próprios são substituídos pelo símbolo $NAME$ e risadas como 'kkkk' são substituídas por $LAUGH$.
- *s2* Um arquivo com estatísticas lexicais de cada locutor.
- *s3* Arquivos representando o vocabulário de cada locutor.
- *s4* Um vocabulário global obtido pela união de todos vocabulários individuais.

Estes arquivos de saída *s1..s4* são detalhados individualmente nas subseções a seguir.

2.1. Texto normalizado (*s1*)

Dado um texto de entrada, o tratamento lexical produz uma versão 'normalizada' alterada pela segmentação dos termos e por várias operações de reescrita e correção ortográfica. Além disso, é estabelecido um certo grau de anonimidade pela substituição de nomes próprios e outros termos. A seguir é apresentado um exemplo de texto original e sua versão normalizada.

> *Texto original:*
> "Desafio alguem a me dizer que nunca fez nenhuma destas coisas tb, kkk acho que tiquei a lista do Brasil todo aki!!!!! #quemNunca"
>
> *Texto normalizado:*
> " Desafio alguém a me dizer que nunca fez nenhuma destas coisas também , $LAUGH$ acho que tiquei a lista do $NAME$ todo aqui ! ! ! ! # quem Nunca "

Neste exemplo é possível observar que houve segmentação (como os pontos de exclamação ao final), decomposição (quemNunca), correção ortográfica (alguem), reescrita (tb) e substituição de termos por símbolos gerais (kkk).

As substituições realizadas durante a normalização são relacionadas a seguir. As bases textuais consultadas para este fim são descritas na Seção 3.

> $NAME$: termos classificados como nomes próprios.
> $EMOTION+$: termos que denotam emoções positivas (e.g., 'yuppiiee').
> $EMOTION-$: termos que denotam emoções negativas (e.g., 'argh').
> $EMOTION*$: termos que denotam emoção ambivalente (e.g., 'hmmm').
> $LAUGH$: termos que denotam riso (e.g., 'huahuaha').
> $FOREIGN$: termos classificados como palavras estrangeiras (e.g., 'amore', 'yes').
> OOV: termos sem classificação definida.

2.2. Estatísticas lexicais (*s2*)

Durante o tratamento lexical, são computadas 155 estatísticas atualizadas a cada termo (palavra, símbolo etc.) processado. Estas estatísticas são divididas em 3 grupos de acordo com sua origem, a saber: 31 estatísticas provenientes da correção ortográfica, 64 estatísticas provenientes do dicionário psicolinguístico LIWC em Pennebaker et al. [2001] e 60 estatísticas provenientes do dicionário Português DELAF-Pt em Muniz [2004]. O conjunto completo forma um arquivo do tipo CSV em que linhas representam os sujeitos (ou textos) processados, e as colunas representam os atributos computados.

Exceto pelo atributo identificador *id* e pelos contadores que identificam o tamanho do texto (i.e., os contadores de sentenças, items e caracteres *sent* , *items*, *chars allTokens* e *wordTokens*), todos os valores das estatísticas são obtidas pela contagem do fenômeno representado dividido pelo total de *items*. Em certos casos, entretanto, estes valore podem assumir um valor maior do que 1 dependendo da natureza do atributo. Em especial, informações de gênero e número provenientes do dicionário tendem a ser superestimadas pois é comum que uma mesma entrada lexical possua múltiplas formas (e.g., masculino e feminino), caso em que todas são contabilizadas simultaneamente.

As 31 estatísticas computadas durante a correção ortográfica (após o identificador do autor *id* na ordem das colunas do arquivo) são sumarizadas na tabela ao final deste relatório. Várias destas estatísticas são provenientes de anotação manual que, tendo sido

realizada por apenas um anotador, devem assim ser consideradas apenas como estimativas, e não como indicadores precisos dos fenômenos que representam.

As 64 estatísticas obtidas a partir do dicionário LIWC são numeradas de 1 a 64 no arquivo de atributos lexicais seguidas do nome original utilizado no dicionário (de *1funct* a *64filler*). Estas estatísticas podem ser consultadas em Pennebaker et al. [2001], Tausczik e Pennebaker [2010], Filho et al. [2013].

Finalmente, o arquivo de estatísticas lexicais inclui ainda 60 estatísticas provenientes do dicionário DELAF-Pt. Estas estatísticas representam as informações de gênero, número, classe gramatical etc. considerados em Muniz [2004].

Além destas estatísticas, as colunas finais do arquivo CSV gerado (depois da coluna com o atributo lexical V3p do diconário DELAF-Pt) contém informações de personalidade e descritivas do autor de cada conjunto de textos (que corresponde a uma linha do arquivo), a saber:

> *inventory*: Indica se o questionário de personalidade foi preenchido via Facebook ou em experimento presencial.
> *exp*: Identificador do pesquisador responsável (somente para experimentos presenciais).
> gender: gênero do sujeito.
> *ti*: Indica se a área de atuação (profissional ou acadêmica) é ou não em TI.
> *reoligiousity*: Grau de religiosidade do sujeito, em uma escala de 0 (nem um pouco religioso) a 5 (muito religioso).
> *course*: Código identificador do curso de graduação do sujeito (somente para certos cursos selecionados).
> *extraversion, agreeableness, conscientiousness, neuroticism, openness*: Valor escalar do modelo Big-five.
> *isExtr, isAgre, isCons, isNeur, isOpen*: Verdadeiro se o valor está no quartil inferior da respectiva dimensão Big-Five, ou falso se estiver no quartil superior. Em outros casos, o valor será '?'.
> *arff*: Uma linha de texto completa contendo todas as colunas do arquivo (todas estatísticas lexicais exceto id e todas informações do sujeito), separadas por vírgula e usando ponto como separador decimal, para uso no ambiente WEKA Witten et al. [2011].

Estas informações podem ser utilizadas como calsses em métodos de aprendizagem de máquina para, por exemplo, aprender características do autor com base nas estatísticas lexicais. Em especial, a última coluna (arff) permite que este tipo de estudo seja conduzido facilmente no ambiente WEKA Witten et al. [2011].

2.3. Vocabulário individual e global (*s3* e *s4*)

As palavras (mas não os símbolos de pontuação ou caracteres especiais), sem distinção entre maiúsculas e minúsculas, formam um arquivo CSV de vocabulário de cada autor. Este arquivo contém a lista de todas palavras que ocorrem no texto (mais os termos especiais $LAUGH$, OOV etc., cf. seção anterior) e suas respectivas frequências.

Além dos arquivos de vocabulário individuais, é computado também um vocabulário global considerando os textos de todos os autores. De forma conjunta, estes arquivos podem ser usado como base para a construção de modelos do tipo bag-of-words, n-gramas (Pereira e Paraboni [2007]) e outros de maior complexidade (e.g., de Novais e Paraboni [2012]). Uma vez que os vocabulários gerais tendem a ser muito extensos para uso prático, a constrção de modelos deste tipo possivelmente terá que ser antecedida de

um procedimento de *stemming* (e.g., Orengo e Huyck [2001]). Além disso, observa-se que a construção de modelos do tipo bigrama ou de ordem superior requer a contagem direta dos termos a partir do arquivo texto normalizado (*s1*), já que a informação do vocabulário corresponde à contagem de unigramas.

3. Processamento

Nesta seção é descrita a sequência de procedimentos denominados tratamento lexical, que produz os quatro tipos de saída (*s2*..*s4*) descritos nas seções anteriores. Cada documento representando texto de um autor específico é processado individualmente para produção do texto normalizado (*s1*), estatísticas lexicais (*s2*) e vocabulário do autor (*s3*). As informações de vocabulário são no entanto acumuladas para formar o vocabulário global (*s4*) de todos os autores ao fim do processamento.

O arquivo texto de entrada é lido linha-a-linha, e estatísticas lexicais registradas no arquivo *s2* são atualizadas em diversos estágios do processamento. Os passos executados são descritos a seguir. Esta descrição faz referência às estatísticas relacionadas no Apêndice deste documento quando pertinente.

Para cada linha de texto lido do arquivo de entrada:

1. *Remoção de caracteres especiais*. Símbolos não pertencentes ao conjunto de caracteres ocidentais são substituídos por espaço.
2. *Remoção de URLs*. Linhas iniciadas por 'http' ou 'www' são desconsideradas, e o contador *links* é atualizado.
3. *Segmentação*. Espaços são inseridos antes e depois de cada palavra ou símbolos de pontuação, se necessário, e espaços consecutivos são eliminados. O resultado é uma série de termos (que podem ser palavras, pontuação ou termos compostos, como *hashtags*) a serem processados um a um.
4. *Contagem de sentenças*. Ao final de uma linha, o contador *sentences* é atualizado.

Para cada termo separado por espaço em uma linha de texto:

1. *Hashtags*. Os símbolos '#' são eliminados, e o contador *hashtag* é atualizado. O termo resultante é tratado normalmente.
2. *Linhas tracejadas*. Caracteres do tipo hífen no início do termo são eliminados. O termo resultante é tratado normalmente.
3. *Termos enfatizados por hífen*. Se metade ou mais dos caracteres do termo for um hífen (e.g., 'f-é-r-i-a-s'), os hifens são removidos, os contadores *emph* e *echars* são atualizados, e o termo corrigido é tratado normalmente.
4. *Termos compostos hifenizados*. Se, por outro lado, o termo possui mais de dois hifens, é considerado um termo composto (e.g.,'hoje-eu-vou-dormir-muito'). Neste caso o termo é desmembrado em termos individuais, a serem tratados normalmente, e o contador *compound* é atualizado.
5. *Termos compostos anotados*. Os termos que constam na lista de termos compostos anotados manualmente são desmembrados em termos individuais segundo regra

da própria lista (e.g., ‘nãovou’ é desmembrado em ‘não’ e ‘vou’) e o contador *compound* é atualizado. Os termos resultantes são tratados normalmente.

6. *Falsos homógrafos, Termos Reescritos e Erros ortográficos*. Os termos que constam nas listas de falsos homógrafos (Duran e Nunes [2015], Duran et al. [2015]), de formas de escrita alternativa (e.g., ‘véii’ como substituto de ‘velho’) e erros ortográficos anotados manualmente (e.g., ‘paralizacao’) são substituídos pelo termo correto, e o respectivo contador é atualizado: *mispell* para falsos homógrafos e erros, ou *rewrite* para escrita alternativa. O termo corrigido é tratado normalmente.
7. *Termos compostos delimitados por maiúsculas*. Termos na forma de sequência de caracteres minúsculos com separadores maiúsculos (e.g., ‘hojeTemFesta’) são desmembrados e o contador *compound* é atualizado. Os termos resultantes são tratados normalmente.
8. *Reiteração*. Os termos resultantes das operações acima podem eles próprios ser termos compostos, e por isso o processo é repetido até resultar em uma lista de termos atômicos (i.e., palavras ou sinais de pontuação) que não podem mais ser subdivididos.

Para cada termo atômico da linha de texto (sentença):

1. *Contagem de caracteres*. O contador *chars* é atualizado.
2. *Pontuação*. Se o termo for um sinal de pontuação, o contador *punct* é atualizado. Alem disso, se o sinal for de interrogação ou exclamação, os contadores *questions* ou *exclam* são atualizados de acordo. O processamento segue então analisando o próximo termo.
3. *Termos numéricos*. Se o termo inicia por um dígito, o contador *numbers* é atualizado. O processamento segue então analisando o próximo termo.

Tratamento de palavras:

1. *Contagem de itens*. O contador *item* é atualizado. Como a partir deste ponto apenas palavras são consideradas, este contador efetivamente representa a quantidade de palavras não-distintas (ou *tokens*) do texto.
2. *Contagem de maiúsculas/minusculas*. Os contadores *upcase*, *lowcase* e *firstup* são atualizados com base na distribuição de maiúsculas e minúsculas da palavra.
3. *Contagem de vocabulário amplo*. O contador *allTokens* é atualizado com base na palavra atual, que ainda não passou por normalização (e.g., substituindo-se termos como nomes próprios, indicadores de riso etc. por símbolos genéricos) e assim representa o total de palavras distintas empregada pelo sujeito.
4. *Classificação*. A palavra é consultada em uma série de dicionários e listas de anotação manual e, se identificada, o respectivo contador é atualizado. Primeiramente, é consultado o dicionário DELAF-Pt (contador *pt-lexicon*), seguido da lista de emoções positivas (*emo+*), negativas (*emo-*), ambivalentes *emo-*) ou de riso (*laugh*), palavras adicionadas manualmente (*added*), formas verbo-pronome (*verb-pro*), nomes próprios (*names*), palavras estrangeiras (*foreign*), desconhecidas (*unkn*), a serem ignoradas (*skip*) e o dicionário DELAF-En (*en-lexicon*).

5. *Contagem de vocabulário normalizado*. Se a palavra for encontrada em uma das bases consultadas, ela é normalizada (cf. Seção 2.1) e o contador de vocabulário normalizado *wordTokens* é atualizado.
6. *Ênfase por repetição de vogais*. Palavras que não foram reconhecidas até este ponto são examinadas para verificar se é um caso de ênfase por repetição de letras exceto R ou S (e.g., 'siiiiiiiiim'). Se uma versão sem letras repetidas existe no dicionário Português, a palavra é corrigida, os contadores *emph* e *echars* são atualizados, e a palavra corrigida é reclassificada repetindo-se o passo [Classificação].
7. *Acentuação automática*. Palavras que não foram reconhecidas até este ponto são examinadas para verificar se podem ser reescritas com certos tipos de acentuação. Por exemplo, palavras terminadas com 'cao' são reescritas com 'ção' e consultadas no dicionário. Se identificada, a palavra é convertida na sua forma normal, o contador *mispell* é atualizado, e a palavra corrigida é reclassificada repetindo-se o passo [Classificação].
8. *Dicionário DELAF-Pt*. Os atributos lexicais obtidos do dicionário DELAF-Pt atualizam as 60 estatísticas deste tipo, cf. Seção 2.2.
9. *Dicionário LIWC*. Os 64 atributos LIWC são atualizados.
10. *Palavras não classificadas*. Palavras que não puderam ser classificadas de nenhuma outra forma são transformadas em OOV para futura revisão manual.

4. Aplicações

Os textos normalizados (*s1*) são utilizados tipicamente quando é necessário realizar um tratamento subsequente ao já realizado, como tratamento de *part-of-speech* (e.g., Fonseca e Rosa [2013]) ou análise sintática (e.g., Bick [2000]). Estes textos são também a base para extração de atributos de interesse específico de aplicações de Geração de Língua Natural (e.g., Paraboni [2003]), já que estas normalmente exigem informações agrupadas por fenômeno ou estímulo de interesse, e não tirariam proveito das estatísticas lexicais (*s2*) agrupadas por locutor[1].

As estatísticas lexicais (*s2*) constituem a base para aplicações de interpretação de língua natural como reconhecimento de personalidade ou caracterização autoral a partir de texto. Abordagens que façam uso de modelos de língua podem complementar este conhecimento com uso dos arquivos de vocabulário (*s3* e *s4*) para construção de modelos de n-gramas e afins.

Agradecimentos

Este trabalho conta com apoio FAPESP nro. 2016/14223-0.

[1]Mas evidentemente aplicações que exploram a questão da variação humana em GLN (e.g., Ferreira e Paraboni [2014]) podem fazer uso das informações lexicais aqui produzidas.

Referências

Bick, E. (2000). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. Tese de Doutorado , Aarhus University.

Dale, R. e Viethen, J. (2009). Referring expression generation through attribute-based heuristics. Em *Proceedings of ENLG-2009*, páginas 58–65.

de Novais, E. M. e Paraboni, I. (2012). Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, 19(2):135–146.

Duran, M. S., Avanco, L., e Nunes, M. G. V. (2015). A normalizer for UGC in Brazilian Portuguese. Em *Proceedings of the Workshop on Noisy User-generated Text (WNUT) - 53rd Annual Meeting of the ACL*, páginas 38–47.

Duran, M. S. e Nunes, M. G. V. (2015). A importância dos falsos homógrafos para a correção automática de erros ortográficos em Português. Em *STIL-2015 IV Jornada de Descrição do Português -*.

Ferreira, T. C. e Paraboni, I. (2014). Referring expression generation: taking speakers' preferences into account. *Lecture Notes in Artificial Intelligence*, 8655:539–546.

Filho, P. P. B., Aluísio, S. M., e Pardo, T. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. Em *9th Brazilian Symposium in Information and Human Language Technology - STIL*, páginas 215–219.

Fonseca, E. R. e Rosa, J. L. G. (2013). Mac-Morpho revisited: Towards robust part-of-speech tagging. Em *9th Brazilian Symposium in Information and Human Language Technology*, páginas 98–107.

Gatt, A., van der Sluis, I., e van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. Em *Proc. of ENLG-07*.

Muniz, M. C. M. (2004). A construção de recursos linguístico-computacionais para o Português do brasil: o projeto de Unitex-PB. Master's thesis, ICMC / USP São Carlos.

Orengo, V. e Huyck, C. (2001). A stemming algorithm for the Portuguese language. Em *8th Symposium on String Processing and Information Retrieval*.

Paraboni, I. (2003). *Generating references in hierarchical domains: the case of Document Deixis*. Tese de Doutorado , University of Brighton.

Paraboni, I., Galindo, M., e Iacovelli, D. (2016). Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*.

Pennebaker, J. W., Francis, M. E., e Booth, R. J. (2001). *Inquiry and Word Count: LIWC*. Lawrence Erlbaum, Mahwah, NJ.

Pereira, D. B. e Paraboni, I. (2007). A language modelling tool for statistical NLP. Em *5th Workshop on Information and Human Language Technology (TIL-2007). Anais do XXVII Congresso da SBC*, páginas 1679–1688, Rio de Janeiro.

Tausczik, Y. R. e Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Teixeira, C. V. M., Paraboni, I., da Silva, A. S. R., e Yamasaki, A. K. (2014). Generating relational descriptions involving mutual disambiguation. *Lecture Notes in Computer Science*, 8403:492–502.

Viethen, J. e Dale, R. (2011). GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. Em *Proceedings of UCNLG+Eval-2011*, páginas 12–22.

Witten, I. H., Frank, E., e Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufman Publ., Burlington, 3 edição.

Apêndice A. Atributos lexicais

Tabela 1. Atributos de tratamento ortográfico

Atributo	Descrição
id	identificador numérico do autor.
sentences	nro. de linhas de texto processadas.
items	nro. de palavras lidas. Representa o tamanho do texto produzido. É o fator usado para cálculo das estatísticas exceto *echars*, que é baseado em *chars*.
chars	nro. total de caracteres lidos.
allTokens	nro. de palavras distintas, incluindo nomes próprios, expressões de riso etc. Não *case-sensitive*.
wordTokens	nro. de palavras distintas após normalização (i.e., nomes próprios transformados em $NAME$ etc.). Representa o tamanho do vocabulário 'real' do sujeito. Não *case-sensitive*.
skip	itens desprezados na análise por não constituírem texto válido.
compound	itens compostos de palavras concatenadas em Português (e.g., "jáVoltei"). Itens contendo palavras estrangeiras são desconsiderados (que são computados como *foreign*).
hashtags	itens iniciados por "#".
links	itens iniciados por "www" ou "http".
punct	caracteres de pontuação.
questions	pontos de interrogação (?).
exclam	pontos de exclamação (!).
numbers	itens iniciado por um dígito, como "123" ou "646-SP".
upcase	itens totalmente escritos em maiúsculas, como "BOM".
lowcase	itens totalmente escritos em minúsculas, como "bom"
firstup	itens escritos com apenas a primeira letra maiúscula, como "Bom".
pt-lexicon	itens encontrados no dicionário DELAF-Pt. São as palavras padrão do Português.
added	itens que correspondem a palavras de uso comum não encontradas no léxico, como estrangeirismos de uso cotidiano ("spam"), palavras que perderam o trema, neologismos etc.
verb-pro	itens representando formas verbo-pronome como "leia-se" etc.
names	nomes próprios, em qualquer idioma (e.g., "Maria", "Athens").
en-lexicon	itens encontrados no dicionário DELAF-En. São as palavras padrão do Inglês.
rewrite	itens em Português propositalmente grafados de forma incorreta. Podem ser reescritos como uma sequência de palavras em Português como "ozóio" (os olhos).
mispell	itens grafados de forma incorreta por descuido ou erro (e.g., "nao", "atrazo" etc.). Inclui falsos homógrafos corrigidos automaticamente (e.g., "bufes" corrigido para "bufês").
foreign	itens contendo palavras estrangeiras, mas que não encontrados no dicionário inglês (e.g., "day5", "IloveSP", "buenas noches", "Monday" etc.).
emo+	itens não-padrão em qualquer idioma que possuem uma clara carga de emoção positiva, exceto risadas. Incluem emoticons como "o/" e expressões como "yay".
emo-	itens não-padrão em qualquer idioma que possuem uma clara carga de emoção negativa, como "argh", "affe", "uó" "zzz" etc.
*emo**	itens em qualquer idioma indicando emoção ambivalente (e.g., "ahh", "hmm" etc.).
laugh	itens indicativos de riso, como "kkkkk", "huahuahua" etc.
emph	itens não computados como emoção, e que foram enfatizados com uso repetido de certas letras como "hoooooojeeee" ou hifens ("a-d-o-r-o").
echars	caracteres expressos em palavras enfatizadas e computadas em *emph*. Calculado sobre o total de caracteres *echar*. Representa uma estimativa do grau de ênfase (e.g., "siim" vs. "siiiiim").
unkn	itens de significado indeterminado, e não contabilizados nas categorias anteriores.