

Linguística de corpus e estudos fraseoparemiológicos

Éric Laporte

eric.laporte@univ-paris-est.fr

Université Paris-Est Marne-la-Vallée

LIGM

<http://infolingu.univ-paris-est.fr>

Universidade federal do Espírito Santo

Departamento de Línguas e Letras

Sumário

A análise de corpus, um critério de qualidade
Especialização em linguística de corpus
Interação com outros linguistas

A análise de corpus

Uma prática imprescindível

Observação de fatos

Avaliação de resultados, exemplo:

make/provide/produce contribution como
paráfrases de *contribute* (Laso, 2009)

Com todo respeito à liberdade metodológica
do pesquisador

A análise de corpus

Um critério de qualidade

Garantia de que são enfrentados fatos em
contexto de uso

Um progresso histórico

Empiricidade

A análise de corpus

Para noções difíceis de formalizar

Conotações *quebrar um galho*

Ênfase *de mão cheia*

Preferências estilísticas, exemplo: *in terms of, in regard to, in/with respect to/of, in/with regard to, in/with reference to, concerning, regarding* em função de tipos de documentos e contextos sintáticos (Rankin & Schiftner, 2009)

“Pontes” intermetodológicas

Linguística de corpus combinada com métodos complementares

- observação **introspectiva**, exemplo: seleção inicial das paráfrases (Laso, 2009)
- **formalização**, exemplo: integração de colocações num léxico para tradução automática (Seretan, 2009)

“Pontes” intermetodológicas

As vantagens de um método compensam os limites de outros

Todos os linguistas deveriam ser linguistas de corpus

Não precisam abandonar as outras atividades

“Pontes” intermetodológicas

Pouco mencionadas por linguistas de corpus
Especialização em linguística de corpus como
metodologia **exclusiva**

Consequências científicas da exclusividade

Variações de expressões

Limitando-se às formas atestadas em um
corpus

As portas da companhia estão abertas

“Pontes” intermetodológicas

Exclusividade, porquê?

- prestígio, imagem moderna
- a linguística de corpus teve que lutar nos anos 1960 e 1970 (rivalidade)

Uma escolha geralmente **implícita**

Especialização implícita em linguística de corpus

- Colocações de linguistas de corpus que sugerem um **questionamento** sobre a validade de outros métodos: é possível que
- sejam empíricos?
 - levem em conta contextos de uso?
 - sejam baseados em fatos reais?
 - alcancem abrangência?

Empiricidade

*O critério estatístico a partir de corpus (...),
como alternativa a uma abordagem
fortemente baseada na intuição do
pesquisador, [é] altamente promissor
(Garrão, 2011)*

**Abordagens usando a intuição do pesquisador
são válidas?**

Introspecção e empiricidade

Más práticas e boas práticas na introspecção
(Laporte, 2008)

- falante nativo treinado e com atividade regular
- controle coletivo por pares
- uso de critérios formais
- avaliação da reprodutibilidade dos critérios

Introspecção e empiricidade

Precauções típicas de uma investigação
empírica

A linguística de corpus fornece uma disciplina
de observação empírica

Não necessariamente a única

Consideração do contexto de uso

Oposição tradicional entre os *contextos de uso* e a *introspecção chomskyana*, que é *gramatical*

A introspecção é capaz de levar em conta contextos de uso?

A introspecção é só chomskyana?

Consideração do contexto de uso

Gramaticalidade (Chomsky)

Colorless green ideas sleep furiously

Pressupõe uma gramática conhecida pelo linguista

Aceitabilidade (Harris)

**Colorless green ideas sleep furiously*

O falante nativo pode avaliar quanto uma forma existe no uso

Gramática: alvo de investigação empírica

Realidade dos fatos

[Com] recursos estatísticos (...), contamos com um valioso aliado (...): dados reais da língua (Garrão, 2011)

Estudos sem estatísticas contam com dados reais?

Realidade dos fatos

**Dicionários construídos sem estatísticas
contêm erros**

Métodos sem estatísticas são válidos?

Os erros invalidam a forma de aplicação do
método, não necessariamente o método

Realidade dos fatos

Erros em gramáticas

Formas permitidas pela gramática mas que não ocorrem (Matuda, 2011)

A gramática é válida se não prediz quais formas ocorrem?

É incompleta

É possível completar a gramática de tal forma que prediga quais formas ocorrem?

Realidade dos fatos

- “*A gramática*” nessas perguntas se refere
- a um conceito que preexiste à pesquisa linguística? mas qual conceito?
 - a um resultado da investigação? é melhor dizer “*uma gramática*”

Realidade dos fatos

É possível completar uma gramática existente de tal forma que prediga quais formas ocorrem?

Existem caminhos para isso

Levar em conta o léxico e as unidades fraseológicas

A linguística de corpus é um desses caminhos

Abrangência

A linguística de corpus contribui a ampliar o leque de formas contempladas

Métodos sem estatísticas conseguem abranger todo vocabulário?

Abrangência

A linguística de corpus é mais eficaz para formas frequentes

Tem limites inerentes para formas raras

Verfahren einleiten “entrar com processo”

eingeleitetes Verfahren “processo iniciado”

Einleitung des Verfahrens “entrada em processo”

Verfahrenseinleitung “entrada em processo”

Dispersão de dados solucionada com ajuda de dados introspectivos (Fritzinger, Heid, 2009)

Uma retórica de rivalidade

A integração já está a caminho entre

- linguística de corpus
- observação introspectiva
- formalização

A linguística de corpus é respeitada

Não precisa mais de retórica de rivalidade

Especialização em linguística de corpus

A linguística de corpus (LC) pode continuar separada?

Linguistas de corpus já usam dados introspectivos (todas ref. bibliográficas)

A força principal da LC é a observação de fatos

Outros linguistas podem “terceirizar” atividades de observação para a LC?

Alvos de investigação

Construção de léxicos para processamento de línguas

Imensas necessidades de observação

Para cada unidade fraseológica:

- quais sentidos?

O empreendedor abre as portas de sua loja

Aquilo me abriu as portas para a carreira de atriz

(Xatara, comunicação pessoal)

Alvos de investigação

- quais variantes sintáticas?

O empreendedor abre as portas de sua loja

As portas da sua loja vão se abrir

Sua loja abre as portas

Dunga abaixou a bola do adversário

**A bola do adversário se abaixou*

O adversário abaixou a bola

(Xatara, com. pess.)

Alvos de investigação

- quais variantes sintáticas? (Garrão, 2011)

*O Munícipe fez **questão** de parabenizar publicamente a cidade*

Depois da **questão do Munícipe de parabenizar publicamente a cidade... [não verbo suporte]*

*A Igreja Adventista do Sétimo Dia fez **acordo** teológico com a Igreja Católica*

*Depois do **acordo** teológico da Igreja Adventista do Sétimo Dia com a Igreja Católica... [verbo suporte]*

Alvos de investigação

Essas perguntas

- são inúmeras
- são muito específicas
- não são imediatamente respondidas por estatísticas

Motivação

Aplicações informáticas: tradução, motores de pesquisa, gestão das informações...

Alvos de investigação

A complexidade das informações necessárias evoca uma estreita interação, mas do que uma “terceirização”

Formalização

Tradução automática pelo método “simbólico”

Tradução de uma palavra

bike = bicicleta

Tradução de uma sequência de palavras

lip service = aprovação fingida

Tradução de uma palavra em presença de
outras

ride [a bike] = andar [de bicicleta]

Formalização

Um número finito de modelos (Seretan, 2009)
Distribuir listas de colocações em um número
finito de modelos
Passagem de fatos brutos, estatísticos, a um
modelo formal simplificador

Conclusão

A interação necessita conhecimento mútuo

A fraseologia necessita que os linguistas de corpus e os outros aceitam essa integração

Bibliografia

- Fritzinger, Heid, 2009. “Automatic grouping of morphologically related collocations”, *Proceedings of the Corpus Linguistics Conference*, University of Liverpool.
- Garrão, 2011. “A identificação de expressões fixas verbais com base em *corpora*”, *Congresso internacional de Fraseologia e Paremiologia*, Universidade de Brasília.
- Laporte, 2008. “Exemplos atestados e exemplos construídos na prática do léxico-gramática”, *Revista (Con)textos Lingüísticos* 2, pp. 26-51. <http://halshs.archives-ouvertes.fr/halshs-00325926>

Bibliografia

- Laso, 2009. “A corpus-based study of the phraseological behaviour of abstract nouns in medical English”, *Proceedings of the Corpus Linguistics Conference*, University of Liverpool.
- Matuda, 2011. “Extração de unidades fraseológicas especializadas em *corpora* comparáveis”, *Congresso internacional de Fraseologia e Paremiologia*, Universidade de Brasília.
- Rankin, Schiftner, 2009. “The use of marginal and complex prepositions in learner English”, *Proceedings of the Corpus Linguistics Conference*, University of Liverpool.
- Seretan, 2009. “An integrated environment for extracting and translating collocations”, *Proceedings of the Corpus Linguistics Conference*, University of Liverpool.