ARQUITETURA E CRIAÇÃO DE UM *CORPUS* PARA O ESTUDO DA EROSÃO LINGUÍSTICA

Lúcia de Almeida Ferrari – UFMG
Orientador: Tommaso Raso - UFMG

INTRODUÇÃO

O trabalho mostra a aplicação da linguística de *corpus* (LC) ao estudo da erosão linguística (ou atrito linguístico) da L1.

Os trabalhos sobre erosão linguística que se utilizam da metodologia de *corpus* são escassos, geralmente são empregados:

- testes de tradução
- julgamento sobre a gramaticalidade de enunciados
- questionários sociolinguísticos
- *C-Tests* com completamento de lacunas
- narrativas dirigidas ou narração de trechos de um filme mudo
- *Wug Tests*
- *Can-do Scales*.

Keijzer (2007) a utiliza somente sob forma de narrativas guiadas, entre muitos outros testes que considera mais relevantes.
Hutz (2004) faz um estudo longitudinal sobre um *corpus* de cartas, mas trata-se de um *corpus* escrito.

ENTÃO PORQUE NESTE ESTUDO FOI UTILIZADA A LINGUÍSTICA DE *CORPUS*?

Permite:

- **obter informações o mais espontâneas possíveis;**
- **o acesso à língua realmente usada pelos informantes;**
- **a comparação com o uso dos informantes não sujeitos a erosão.**

EROSÃO LINGUÍSTICA ⟶ re-estruturação gradual, convergência ou perda das estruturas fonológicas, morfossintáticas, lexicais e pragmáticas na produção de falantes de L1, em contato prolongado com uma L2, devido à interferência desta, ou por falta de insumo (Schmidt *et alii*, 2004). Neste trabalho a L1 estudada é o italiano, enquanto a L2 é o português brasileiro (PB).

Os estudos mais influentes encontram-se em:

- Seliger e Vago (1991);
- Köpke e Schmid (2004);
- *International Journal of Bilingualism* (Vol 8:3, 2004)
- várias edições da revista *Bilingualism: Language and Cognition*.

OBJETIVOS DO ESTUDO

O objetivo de nosso estudo foi **verificar os resultados** das pesquisas de Raso e Vale (2007 e 2009) que analisaram um *corpus* de língua falada:

série de pesquisas sobre a erosão linguística de italianos cultos em contato prolongado com o PB nas quais a metodologia da LC é parte essencial.Os estudos concentraram-se sobre um grupo de clíticos.

Pronome *ci* nos valores atualizante, lexicalizante e locativo; pronome *ne* em função partitiva, argumental e locativa; pronomes acusativos de terceira pessoa: *lo*, *la*, *li*, *le*, *l'*.

METODOLOGIA, DADOS E RESULTADOS

Para verificar os resultados das pesquisa anteriores foi **criado um novo *corpus*** cuja arquitetura buscou seguir as indicações de Biber (1993), McEnery; Wilson (1996) e Meyer (2004), entre outros.

Total de palavras: 21298 (corpus pequeno)

Passos seguidos para a compilação:

►**Escolha dos informantes** adequados: italianos nativos crescidos na Itália até a idade adulta e que ali tivessem completado pelo menos até o segundo grau, que é de cinco anos, e possivelmente possuíssem formação universitária. Isto para evitar a ambiguidade entre erosão e aquisição incompleta e para contar com informantes que tivessem uma adequada capacidade de reflexão metalinguística. Estes deveriam residir no Brasil há pelo menos oito anos, período tradicionalmente considerado suficiente para a detecção dos fortes sinais da erosão linguística.

Tabela 1: Dados sobre os informantes contatados para a pesquisa.

Informante	Sexo	Idade	Proveniência	Título de estudo	Tempo de residência	Profissão	Retorna à Itália
MAS	M	faixa 2	Brescia	segundo grau e faculdade na Itália	8 anos	professor e tradutor	a cada 1 ou 2 anos
UCR	M	faixa 3	Varese	segundo grau na Itália, graduação no Brasil	12 anos	professor	raramente
GIC	M	faixa 3	Torino	segundo grau e faculdade (incompleta) na Itália	15 anos	tradutor e gráfico	raramente
MON	F	faixa 3	Oristano	segundo grau e faculdade na Itália	11 anos	missionária	a cada 2 ou 3 anos
ANG	M	faixa 4	Bergamo	segundo grau e faculdade (incompleta) na Itália	20 anos	contador e missionário	quase todos os anos
MRC	M	faixa 3	Milano	segundo grau e faculdade (incompleta na Itália); graduação no Brasil	19 anos	professor	raramente
LIV	F	faixa 3	Chiasso	segundo grau na Itália e graduação no Brasil	33 anos	professora aposentada	de vez em quando
PAT	F	faixa 3	Chiasso	segundo grau na Itália, graduação e pós-graduação no Brasil	33 anos	professora	de vez em quando

Faixas etárias : de 18 a 25 anos **faixa 1**; de 26 a 39 anos **faixa 2**; de 40 a 60 anos **faixa 3**; mais de 60 anos **faixa 4**.
A identidade do informante é mantida em sigilo através das siglas que os identificam e do *Termo de Consentimento* por eles assinado e aprovado pelo *Comitê de Ética em Pesquisa* (COEP) da UFMG.

LIMITAÇÕES DO *CORPUS* RASO-FERRARI

Serão enumeradas algumas dificuldades encontradas na compilação do *corpus* que ajudam a explicar suas limitações e a **necessidade de continuar sua compilação para que seja atingida uma maior representatividade e um melhor balanceamento**.

- dificuldades em se encontrar os **informantes adequados** ou que aceitassem participar da pesquisa;
- dificuldade em **criar situações comunicativas variadas** com uma variedade tão pequena de informantes que deveriam interagir entre si;
- dificuldade em encontrar algum **software** adequado que facilitasse a varredura do *corpus*;
- o **tamanho do *corpus* é reduzido**, isto porque foi necessário operar vários **cortes** em textos muito longos ou excluir textos muito curtos: conseguimos chegar assim a 8 texto que perfazem entre 1500 e 3000 palavras, comparáveis em tamanho com o *C-ORAL-ROM italiano* e com as indicações de Biber (1993) de que trechos de 1000 palavras fornecem uma amostragem suficientemente confiável para itens frequentes como os pronomes;
- mesmo após os cortes operados nos diálogos e interações, das **21.298 palavras** que compõem o *corpus*, **8.495 pertencem a um único informante**, o que pode **comprometer o balanceamento e a representatividade**.

Estes fatores foram levados em conta no momento da análise e da interpretação de dados e serviram para compreender melhor as dificuldades em se trabalhar com *corpora.*

►**Gravação** com equipamentos sofisticados (Gravador digital Marantz PDD660 com cartão de memória Compact Flash de 2 *gigabytes*;*Kits wireless* Sennheiser Evolution EW100 G2 - *receiver*, *transmitter*, microfone de lapela- com dois *kits* /carregador adaptados para o *receiver*, ou solução alternativa com bateria própria e seis microfones completos; microfone omnidirecionail Sennheiser MD 421 com pedestal Hunter PMP103 e cabos RCL303569 de 6 metros, ou sistema *wireless)* para garantir uma qualidade acústica suficiente para a análise prosódica.

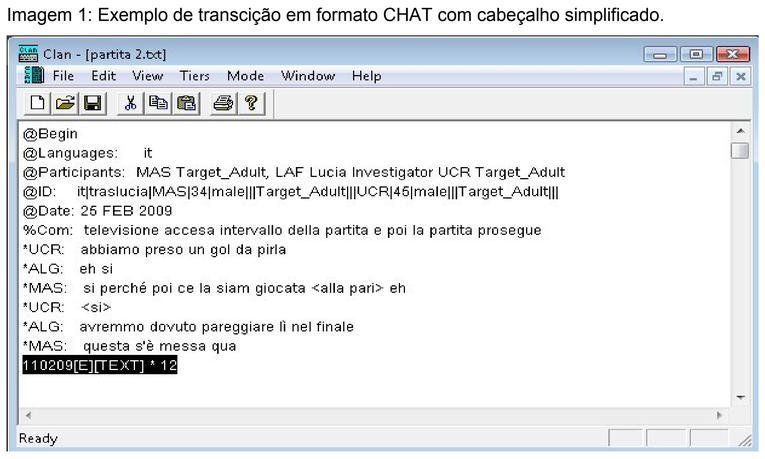
►**Escolha de interações** diafasicamente muito variadas para cobrir a maior variedade possível de atos de fala.

Tabela 2: informações sobre o *corpus* e cada um dos textos.

Título	número total de palavras	número de palavras por informantes	Tipologia e descrição da interação
Pranzo	4408	MRC: 1843 MAS: 2565	**Diálogo** entre os dois participantes durante o almoço em casa de MAS: os assuntos versam opiniões sobre futebol, o vício do cigarro e uma longa discussão sobre vinhos. (LAF está presente mas fala pouco). Gravado com microfone omnidirecional.
Partita 1	1284	UCR: 533 MAS: 751	**Diálogo** entre UCR e MAS e depois entre UCR e LAF enquanto aguardam que inicie um jogo de futebol na televisão, na casa de MAS. O primeiro diálogo trata de comentários sobre pessoas conhecidas e futebol; o segundo diálogo sobre o que é feito nos finais de semana. Gravado com microfone omnidirecional.
Partita 2	3578	UCR: 1348 MAS: 2230	**Conversação** entre MAS, UCR, ALG (informante nativo não adequado à pesquisa) e LAF: são comentários feitos durante um jogo de futebol na televisão, na casa de MAS. Gravado com microfone omnidirecional.
Sorelle	3691	PAT: 1194 LIV: 2497	**Diálogo**, dividido em duas partes, entre duas irmãs. O primeiro diálogo versa sobre livros e aulas de italiano. O segundo sobre a mudança de apartamento de LIV e a nova decoração da casa. Gravado com gravador de fita cassete.
Missionari	1764	MON: 1764	**Monólogo**: é uma entrevista a uma missionária sobre sua experiência no Brasil (LAF, a entrevistadora, intervém muito pouco). A entrevista foi feita na sede da paróquia. Gravado com microfone omnidirecional.
Genitori	1480	ANG: 1480	**Monólogo**, com algumas esporádicas intervenções de outros participantes que não são informantes adequados à pesquisa. ANG relata sua mudança com a família para o Brasil, as motivações que os troxeram e as dificuldades iniciais. Gravado na paróquia onde ele trabalha. Gravado com microfone omnidirecional.
Medici	2144	GIC: 2144	**Diálogo** entre GIC e LAF (a interação de GIC é muito superior à de LAF), na residência de GIC, em que GIC comenta sobre suas experiências negativas com os médicos e relata vários episódios sobre o assunto. Gravado com microfones de lapela.
Cena	2949	MAS: 2949	**Diálogo** entre MAS e LAF (a presença de LAF é bastante grande) em sua residência. Enquanto preparam o jantar comentam sobre os ingredientes e a forma de cozimento, o resultado da receita e vários assuntos. Gravado com microfones de lapela.

►**Transcrição** em formato **CHAT** (MacWhinney, 1994, 2000) das gravações e seleção dos trechos a serem utilizados. Para cada texto foi criado um **cabeçalho** que contém os dados essenciais como participantes, faixa etária, situação comunicativa, data e local, transcritor.

Imagem 1: Exemplo de transcição em formato CHAT com cabeçalho simplificado.



►**Varredura do *corpus***. Optou-se por não fornecer o *corpus* de nenhum tipo de anotação morfossintática pois, após um teste piloto, percebemos que o que melhor se adaptaria ao italiano, o *Tree Tagger*, não conseguia distinguir de forma suficientemente confiável os pronomes, que eram nosso objeto de estudo, de outras partes do discurso homófonas e homógrafas como, por exemplo, os artigos.

Uma varredura preliminar, após a limpeza dos metadados, foi feita com o software *Texstat 2,* que realiza buscas de índices de frequência e de itens específicos, com opção de visualização do contexto. Para a conferência das ocorrências foi também utilizado o *Notepad* ++. Contudo, visto que a análise incluia também verificar se os clíticos estudados estariam ausentes, a varredura teve que ser feita de forma manual.

►**Análise dos resultados**. Foi montado um *sub-corpus* de comparação de monolíngues extraindo 14 textos do *C-ORAL-ROM italiano* (Cresti-Moneglia, 2005) entre conversações, monólogos e diálogos que fossem o mais parecidos possível com aqueles do *corpus* de erosão analisado, para um total de **21.224 palavras**.

Entre os textos escolhidos, tentou-se buscar aqueles com uma maior participação de informantes de localidades diferentes da Itália, buscando assim obter uma maior variação diatópica, além de optar por aqueles em que os participantes tivessem um nível escolar mais alto, similar portanto àquele dos sujeitos pesquisados no *corpus* de erosão.

Também deste *corpus* foi feita uma varredura de forma manual, como aquela descrita para o *corpus* de erosão. Os dados de ambos os *corpora* foram então **normalizados e comparados**. Em seguida os resultados foram comparados com aqueles dos estudos de Raso e Vale, inclusive com cruzamento de dados dos *corpora* de erosão e daqueles de comparação.

CONCLUSÃO

Os resultados **confirmaram** em boa medida o que foi encontrado nos estudos anteriores, ou seja sinais de erosão para todos os clíticos analisados. Todavia são as divergências que mostraram-se extremamente úteis para a reflexão metodológica.

O *corpus* analisado por Raso e Vale é composto principalmente por entrevistas com um assunto definido; o novo *corpus* privilegia interações diafasicamente variadas. Isto favoreceu a produção de estruturas diferentes nos dois *corpora*. A variação nas ocorrências de alguns pronomes parece portanto ser devida à arquitetura dos *corpora* analisados e às diferentes tipologias textuais, mais ainda do que ao tempo de contato com o português brasileiro, que é maior no *corpus* analisado por Raso e Vale e menor naquele por nós elaborado.

A suposição de que a arquitetura e as diferenças diafásicas determinem resultados muitas vezes contrastantes foi confirmada também na análise dos *corpora* de comparação: aquele utilizado por Raso e Vale, o BADIP (De Mauro, 1993) composto por interações mais planejadas e o *C-ORAL-ROM italiano*, muito espontâneo, evidenciaram diferenças às vezes muito evidentes no número de ocorrência dos clíticos.

REFERÊNCIAS

BIBER, D.. Representativeness in Corpus Design. In: *Literary and Linguistic Computing*, Vol. 8, n. 4, Oxford University Press, 1993.

CRESTI, E.; MONEGLIA, M. (a cura di) *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Language*. Amsterdam-Philadelphia: John Benjamin, 2005

DE MAURO, T. *et alii* . *Lessico di frequenza dell'italiano parlato.* Milano: EtasLibri, 1993

FERRARI, L.A.; *A erosão linguística de italianos cultos em contato com o português brasileiro: aspectos do sistema pronominal.* Dissertação de Mestrado ApresentadA junto ao Poslin do Curso de Letras da UFMG para a obtençaõ do título de Mestre. Belo Horizonte, 2010

HUTZ, M. Is there a natural process of decay? A longitudinal study of language attrition. In: KOPKE, B.; SCHMID, M.S. *First Language Attrition. Interdisciplinary perspectives on methodological issues.* Amsterdam/Philadelphia: John Benjamin, 2004, pp. 189-206

KEIJZER, M. *Last in first out? An investigation of the regression hypothesis in Dutch emigrants in Anglophone Canada*. Vrije Universiteit, 2007

KÖPKE, B.; SCHMID, M.S. *First Language Attrition. Interdisciplinary perspectives on methodological issues.* Amsterdam/Philadelphia: John Benjamin Publishing Company, 2004

MACWHINNEY, B. *The CHILDES project: tools for analysing talk.* Hillsdale: Lawrence Erlbaum, 1994
_____ *The CHILDES Project: Tools for Analyzing Talk. Volume 1: Transcription format and programs. Volume 2: The Database.* Mahwah, NJ: Lawrence Erlbaum Associates, 2000

McENERY, T.; WILSON, A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 1996

MEYER C.F. *English Corpus Linguistics. An introduction*. Cambdridge: Cambridge University Press, 2004

RASO, T. Erosione dei clitici e strutture tematizzanti in italiani colti in contatto prolungato col portoghese brasiliano. In: *Sintassi storica e sincronica dell'italiano. Subordinazione, coordinazione, giustapposizione*. Atti del X Congresso della Società Internazionale di Linguistica e Filologia Italiana, 2009. p. 384-399.

RASO, T. ; VALE, H. P. A erosão lingüística em italianos cultos em contato prolongado com o português do Brasil: os clíticos e alguns efeitos na estrutura do enunciado. In: *Revista de Italianística*, v. 16, 2009, p. 1-22

SCHMID M.S. "*A new blueprint for language attrition research*". In KÖPKE B. e SCHMID M.S., *First Language Attrition. Interdisciplinary perspectives on methodological issues.* Amsterdam/Philadelphia: John Benjamin Publishing Company, 2004

SELIGER H.W. & VAGO R.M. (Eds.), *First Language Attrition* (pp. 175-188). Cambridge: CUP, 1991

VALE, H. P. . A erosão lingüística dos italianos cultos em contato prolongado com o português do Brasil: os clíticos. Belo Horizonte: Monografia apresentada no Curso de Graduação em Letras da Faculdade de Letras da Universidade Federal de Minas, 2007

UFMG

1941 - 2011 70 anos do Curso de Letras na UFMG

PosLin

ELC - 2011 X Encontro de Linguística de Corpus
EBRALC - 2011 - V Escola Brasileira de Linguística Computacional