O que é a Linguística de *Corpus*?

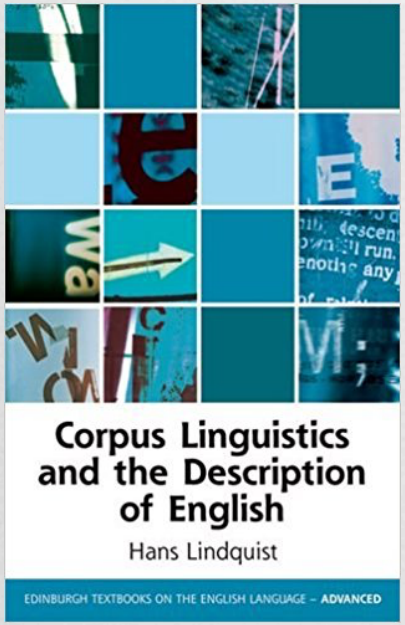
Rodrigo Garcia Rosa (Doutorando)
FFLCH/USP

O que é a **Linguística de *Corpus***?

- **Socio**linguística
- **Psico**linguística
- **Neuro**linguística
- Linguística **histórica**
- Linguística **cognitiva**
- Linguística **computacional**

Linguística de *Corpus*

- O que é o "*corpus*" da Linguística de *Corpus*?



"A posição do presente livro é a de que linguística de *corpus* não é um ramo da linguística similar aos demais ramos da linguística, uma vez que o "*corpus*" não lhe diz o que é estudado, mas diz que uma metodologia particular é utilizada. Portanto, linguística de *corpus* é uma **metodologia** que abarca uma série de diferentes métodos que podem ser usados por pesquisadores de muitas orientações teóricas distintas"

Lindquist (2009:01)

Linguística de *Corpus*

"…é uma área focada em uma **série de procedimentos**, ou **métodos**, para o estudo da língua…"
(McEnery & Hardie, 2012:01)

"A Linguística de Corpus (LC) oferece uma **metodologia** que veio facilitar muito a identificação das unidades convencionais da língua"

(Tagnin, 2013:29)

Corpus-based ou corpus-driven?

Corpus-based ou corpus-driven?

- Corpus-driven:

  Teoria (investigação explanatória) *vs*
  Dados (suporte empírico)

"Isto seria uma insensatez conceitual. Dados são dados e teorias são teorias: observações, apesar de importantes como suporte necessário para explicações, claramente não constituem explicações em si"

McEnery & Hardie (2012:148)

Abordagens *Corpus-based*

"A linguística de *corpus* tem o potencial de reorientar todo o nosso modo de estudar línguas. Ela é capaz de refinar e redefinir uma gama de teorias linguísticas"

McEnery & Hardie (2012:01)

"Portanto, linguística de *corpus* é uma metodologia que abarca uma série de diferentes métodos que podem ser usados por pesquisadores de muitas orientações teóricas distintas"

Lindquist (2009:01)

Linguístia de *Corpus*: possibilidades ou probabilidades?

Chomsky: "For example, you said you want to know about possibility and impossibility. A corpus never tells you what is impossible. In fact, it doesn't even tell you what is possible".

Aarts (2000)

"Só o impossível acontece. O possível apenas se repete, se repete, se repete" (Chacal)

Possibilidades vs Probabilidades

- Nem tudo que é **possível** é **provável**, mas tudo que é verdadeiramente **provável** é necessariamente **possível**.

Linguística de *Corpus*: empirismo probabilístico

- A linguística de *Corpus* focará nas probabilidades da língua em vários níveis de análise:
  - **Morfologia:** a produtividade dos sufixos nominais – *ção* e –*mento*
  - **Léxico:** padrões lexicais como colocações (*atender a porta* vs *answer the door*)
  - **Sintaxe:** relação entre verbos e estruturas argumentais (Rosa, 2014)
  - **Semântica:** investigação de itens polissêmicos (Gries, 2006)
  - **Pragmática:** verificação da intenção pragmática e sua saliência no discurso.
  - **Discurso:** verificação de escolhas lexicais que caracterizam certos discursos (machismo vs feminismo)

Linguística de *Corpus*: dados vs introspecção

"…qualquer *corpus* que tive a chance de examinar, por menor que fosse, ensinou-me fatos que eu nunca poderia ter imaginado encontrar de qualquer outro modo"

"Eu não creio que haja quaisquer corpora, por maiores que sejam, que contenham informações sobre todas as áreas do léxico e da gramática do inglês que eu queira explorar"

Fillmore (1992:35)

Investigação Corpus-based

Case study 1: Existential Constructions

1. *I went to Berkeley. I know what it's like. There are a lot of women like that. She compared Ann Romney to Hitler.* **[SPOK/2014]**

2. *Chi-square test analysis revealed that there were no statistically significant differences between male and female students' health promotion behaviors* **[ACAD/2013]**

3. *So, you know, there 's a lot of things we've got to clean up in our own neighborhoods* **[SPOK/2015]**

4. *Right, and you know, there are a number of issues with these older reactors* **[SPOK/2011]**

5. *When someone like Kim dies, there 's a lot of questions you ask yourself.* **[NEWS/1999]**

Existential constructions

FREQ	CONTEXT	FREQ
1	THERE ARE A LOT OF THINGS	486
2	THERE 'S A LOT OF THINGS	285
3	THERE ARE A NUMBER OF THINGS	121
4	THERE ARE A COUPLE OF THINGS	117
5	THERE ARE A LOT OF QUESTIONS	107
6	THERE WERE A LOT OF THINGS	101
7	THERE ARE A NUMBER OF REASONS	93
8	THERE ARE A LOT OF REASONS	90
9	THERE ARE A LOT OF WOMEN	89
10	THERE ARE A NUMBER OF WAYS	87
11	THERE WERE NO STATISTICALLY SIGNIFICANT DIFFERENCES	81
12	THERE ARE A LOT OF WAYS	73
13	THERE ARE A LOT OF FOLKS	62
14	THERE 'S A COUPLE OF THINGS	58
15	THERE ARE A LOT OF ISSUES	54
16	THERE ARE A LOT OF PROBLEMS	47
17	THERE ARE A LOT OF PLACES	44
18	THERE ARE A NUMBER OF FACTORS	44
19	THERE 'S A LOT OF QUESTIONS	42
20	THERE WERE TEARS IN HER EYES	42

Investigação Corpus-based

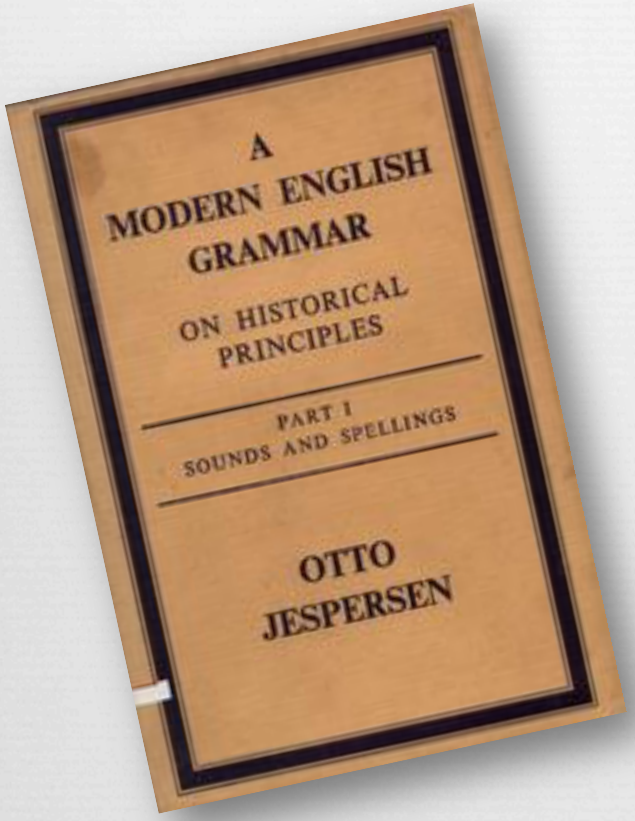
- Como verificamos as restrições (*constraints*) estruturais?
- **There is a lot of/a couple of* = nenhuma ocorrência
- "Não ocorre no *corpus*, logo é agramatical" (=incorreto)

6. **There was no significant main effects for list or presence of background noise.*

7. **Look. There is a lot of problems with the system. It needs to be fixed.*

Muito depende da forma como o *corpus* foi compilado

O que é um *Corpus*?



- A Modern English Grammar on Historical Principles (1909 – 1949) de Otto Jesperson
- "To do" or "to make" that's the question (1981) de Stella Tagnin

Como se define um *corpus*?

"…um conjunto de textos que podem ser lidos eletronicamente e que são considerados uma boa base sobre a qual se pode estudar um conjunto específico de perguntas de pesquisa"

McEnery & Hardie (2012:01)

Como se define um *corpus*?

"…um conjunto de textos que podem ser lidos eletronicamente e que são considerados uma boa base **sobre a qual se pode estudar um conjunto específico de perguntas de pesquisa**"

McEnery & Hardie (2012:01)

Como se define um *corpus*?

- Segundo Rastier (2001 apud Neveu, 2008), "um 'dado' é dado não *ao* pesquisador, mas *pelo* pesquisador"
- Um *corpus* deve ser compilado segundo critérios específicos e deve ser suficientemente representativo da parte da língua que se pretende estudar (Tagnin, 2008).

Critérios fundamentais

- **Representatividade**
- **Perguntas de Pesquisa**

Corpus: a representatividade e as perguntas de pesquisa

"…não podemos (ou podemos com alguma cautela) fazer afirmações gerais sobre a natureza de uma língua baseados em um *corpus* que contém somente um tipo de texto ou um número limitado de tipos de textos"

McEnery & Hardie (2012:02)

Corpus: a representatividade e as perguntas de pesquisa

Fraseologia do verbo *get* na língua inglesa: uma abordagem da Linguística de Corpus e da Gramática de Construções

Corpus: a representatividade e as perguntas de pesquisa

Motivação para o estudo

Linhas de Concordância retiradas do *Corpus of Contemporary American English* (COCA)			
(19)	It can't	**get**	any better, and it hasn't gotten much better.
(20)	They	**get**	access to health care, and even just with emergency rooms
(21)	we need to	**get**	away from the national perspective on this issue.
(22)	We wanted to	**get**	rid of a dictator, but by the time we realized that there
(23)	when we fall in love, we	**get**	high on dopamine, much the way addicts do on cocaine.
(24)	Carter snaps back, "	**get**	out of my way, will you?
(25)	The United States can	**get**	around its own legal restrictions on technology transfer
(26)	Students	**get**	into the habit of reflecting on what they have done
(27)	he would " just have to	**get**	used to and learn to live with.
(28)	We need to	**get**	better at supporting people to manage their health

Corpus: a representatividade e as perguntas de pesquisa

Significância do objeto

Frequência geral de *get* e sua distribuição em diferentes gêneros (wordandphrase.info)

RANK	PoS	WORD	TOTAL	SPOKEN	FICTION	MAGAZINE	NEWS	ACADEMIC
39	Verb	GET	1080432	389689	275602	188656	187778	38707

Tabela 1 - Distribuição dos 10 verbos mais frequentes em cinco gêneros discursivos (COCA)

	Língua Falada		Ficção		Revista		Jornal		Acadêmico	
Rank	Verbo	Frequência	Verbo	Frequência	Verbo	Frequência	Verbo	Frequência	Verbo	Frequência
1	KNOW	301147	KNOW	77553	GET	38619	SAY	53201	SEE	48070
2	THINK	296322	THINK	48592	WANT	38234	THINK	43837	NEED	18903
3	WANT	102209	WANT	44958	SEE	36850	WANT	37549	INCLUDE	16025
4	MEAN	96888	GET	39086	THINK	35928	GET	29760	MAKE	14573
5	SAY	79153	SEE	36525	MAKE	34081	KNOW	28390	PROVIDE	13747
6	THANK	72499	GO	32764	KNOW	33789	NEED	21907	USE	13509
7	GET	62100	COME	32563	NEED	30219	GO	21291	KNOW	13337
8	GO	57680	LOOK	31067	SAY	27882	SEE	20608	THINK	13234
9	LOOK	47869	SAY	29105	TAKE	26670	MAKE	19854	SUGGEST	11113
10	LET	46068	LET	27384	GO	25667	TAKE	17266	WANT	11035

Corpus: a representatividade e as perguntas de pesquisa

- Dada a versatilidade sintática e semântica de *get*, quais são as principais construções gramaticais em que esse verbo pode ocorrer?
- Quais construções licenciam quais sentidos em termos de frequência de uso?
- Quais são os fraseologismos que instanciam mais frequentemente as construções com *get*?

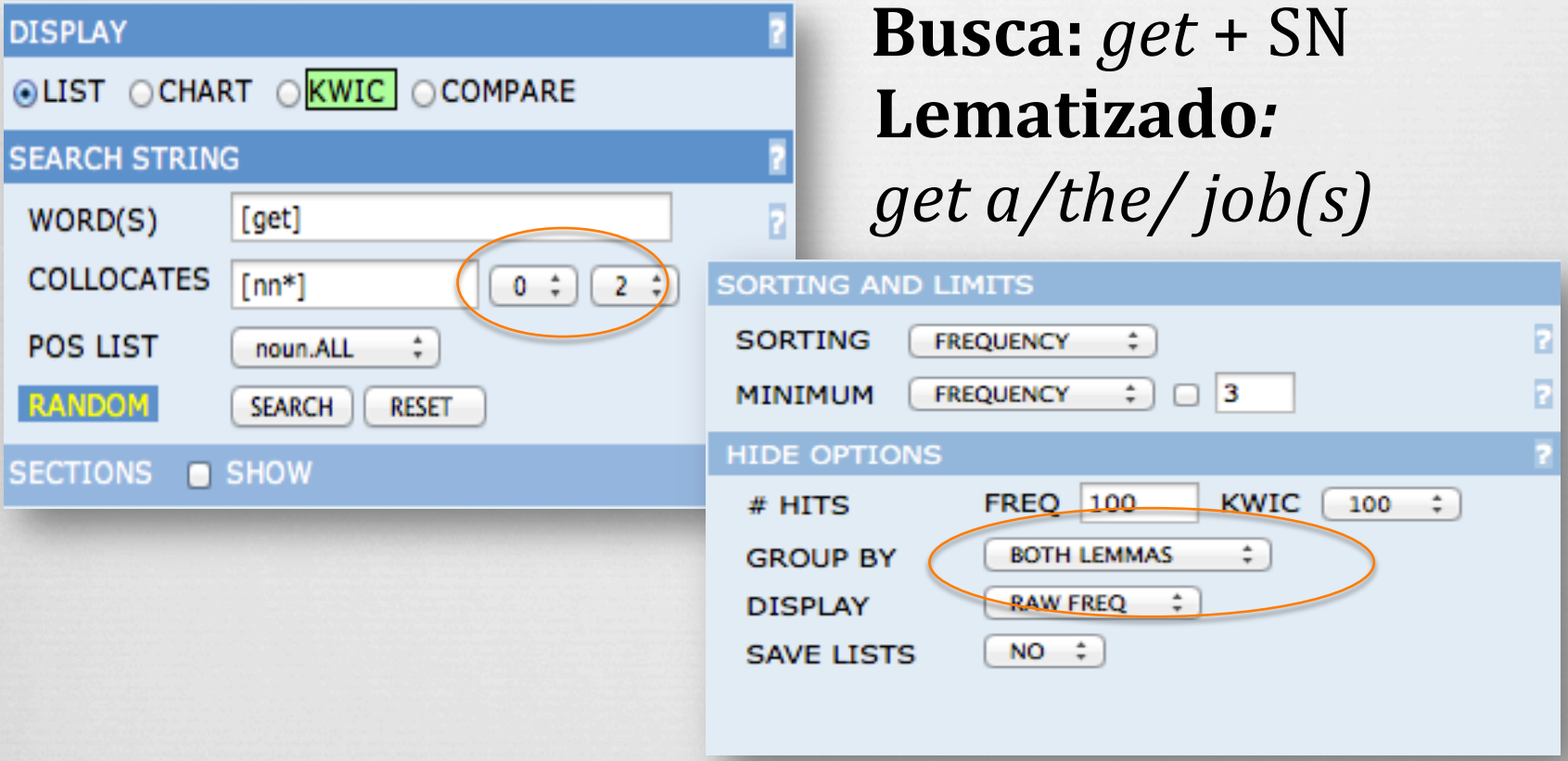
Corpus: a representatividade e as perguntas de pesquisa

Construções com *get*

1. Uso de posse: *He got a book*
   (GET + SN)
2. Uso incoativo: *He got poor*
   (GET + Sadj)
3. Uso de movimento: *He got safe to land*
   (GET + SPrep)
4. Uso causativo: *They got my father to contribute*
   (GET + SN + inf.)

Tagnin (1977, p. 254)

Corpus: a representatividade e as perguntas de pesquisa



Busca: *get* + SN
Lematizado:
get a/the/ job(s)

Evitar: [*get* [$_1$ *away* [$_2$ *with* [$_3$ *the* [$_4$ *problem*]]]]]

Extração dos dados

Tabela 10 – Lista de palavras resultante da busca por *get* + *noun* com até duas palavras à direita

Rank	Colocado	Frequência absoluta	*Rank*	Colocado	Frequência absoluta
1	[GET] || [JOB]	14030	26	[GET] || [WAY]	2608
2	[GET] || [PEOPLE]	11322	27	[GET] || [PICTURE]	2600
3	[GET] || [MONEY]	9596	28	[GET] || [POINT]	2560
4	[GET] || [CHANCE]	8478	29	[GET] || [NEWS]	2526
5	[GET] || [THING]	7216	30	[GET] || [PHONE]	2470
6	[GET] || [CALL]	6138	31	[GET] || [LETTER]	2442
7	[GET] || [INFORMATION]	5912	32	[GET] || [VOTE]	2414
8	[GET] || [ATTENTION]	5622	33	[GET] || [GUY]	2322
9	[GET] || [IDEA]	5584	34	[GET] || [FOOD]	2204
10	[GET] || [HELP]	5548	35	[GET] || [LIFE]	2194
11	[GET] || [TROUBLE]	4880	36	[GET] || [RESULT]	2092
12	[GET] || [MESSAGE]	4816	37	[GET] || [LOOK]	2064
13	[GET] || [SHOT]	4614	38	[GET] || [BALL]	2006
14	[GET] || [HAND]	4462	39	[GET] || [QUESTION]	1994
15	[GET] || [SENSE]	4062	40	[GET] || [SUPPORT]	1954
16	[GET] || [KID]	3934	41	[GET] || [CHILD]	1952
17	[GET] || [WORD]	3468	42	[GET] || [SLEEP]	1902
18	[GET] || [ANSWER]	3312	43	[GET] || [GUN]	1880
19	[GET] || [WORK]	2864	44	[GET] || [STORY]	1762
20	[GET] || [PROBLEM]	2830	45	[GET] || [BREAK]	1720
21	[GET] || [NAME]	2750	46	[GET] || [CAR]	1696
22	[GET] || [FEELING]	2738	47	[GET] || [TREATMENT]	1696
23	[GET] || [CREDIT]	2736	48	[GET] || [FOOT]	1680
24	[GET] || [TIME]	2672	49	[GET] || [WATER]	1644
25	[GET] || [HOLD]	2610	50	[GET] || [HELL]	1622

184.198 ocorrências

Corpus: a representatividade e as perguntas de pesquisa

Quantidade de dados por cada palavra

vs.

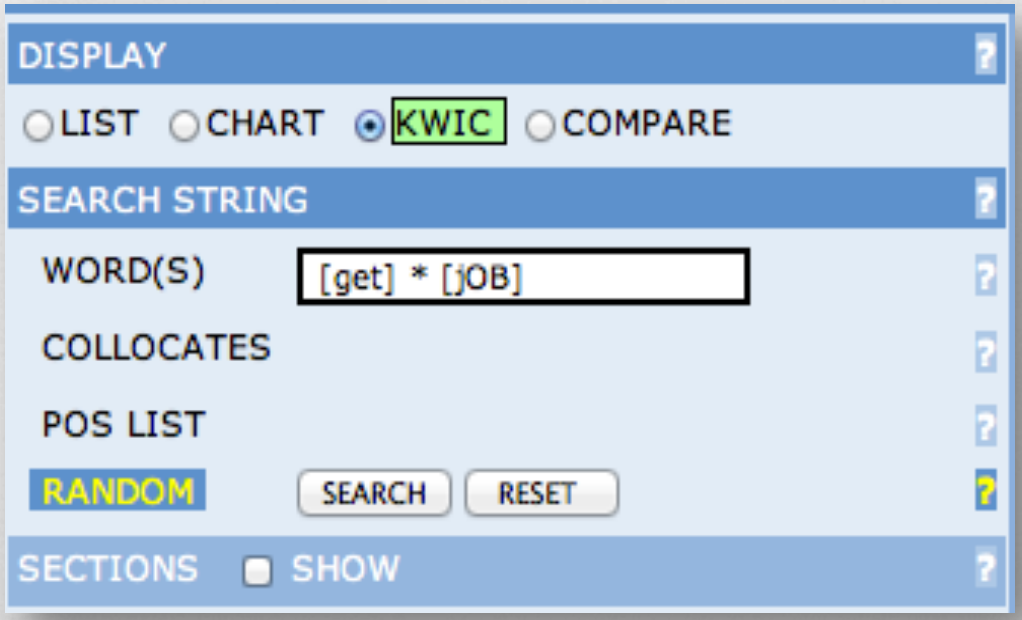
Número de palavras selecionadas

Tabela 11 – Dados que compõem nosso *corpus* de estudo

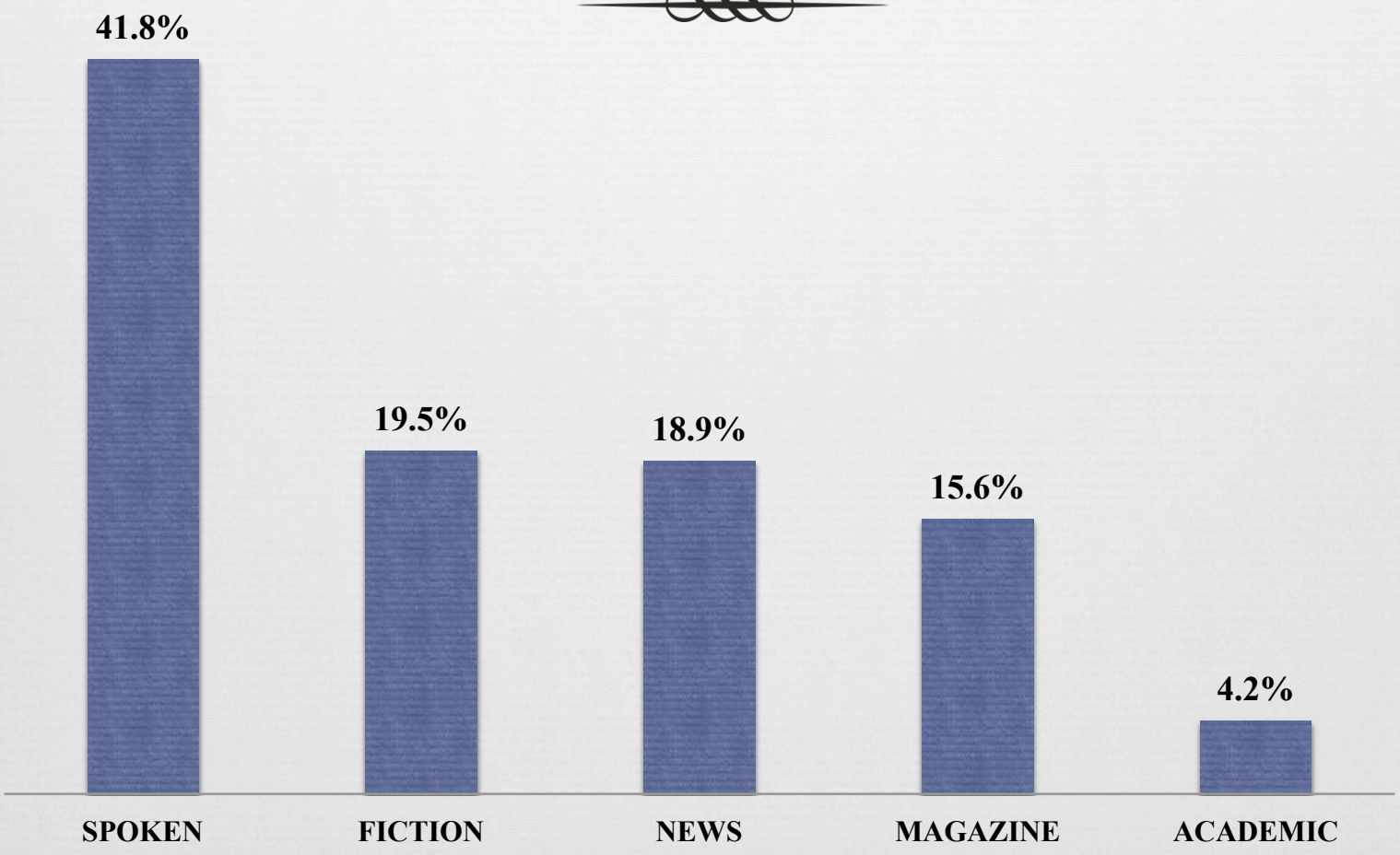
Número	Colocado	Freq. absoluta	5%	Soma
1	[JOB]	14030	702	702
2	[PEOPLE]	11322	566	1268
3	[MONEY]	9596	480	1747
4	[CHANCE]	8478	424	2171
5	[THING]	7216	361	2532
6	[CALL]	6138	307	2839
7	[INFORMATION]	5912	296	3135
8	[ATTENTION]	5622	281	3416
9	[IDEA]	5584	279	3695
10	[HELP]	5548	277	3972
11	[TROUBLE]	4880	244	4216
12	[MESSAGE]	4816	241	4457
13	[SHOT]	4614	231	4688
14	[HAND]	4462	223	4911
15	[SENSE]	4062	203	5114
16	[KID]	3934	197	5311
17	[WORD]	3468	173	5484
18	[ANSWER]	3312	166	5650
19	[WORK]	2864	143	5793
20	[PROBLEM]	2830	142	5934

Corpus: a representatividade e as perguntas de pesquisa

- 9.210 enunciados com uma variabilidade de 50 substantivos
- Busca por KWIC dos 50 colocados
- Busca por formas lematizadas
- Geração e seleção de linhas de maneira randômica



Corpus: a representatividade e as perguntas de pesquisa



Distribuição dos enunciados do *corpus* de estudo por gênero discursivo

Corpus: a representatividade e as perguntas de pesquisa

Categorização semântica das construções

Construção	Sentido Central	Ocorrências
Transitiva		6220
	Obtenção	4192
	Recebimento	874
	Posse	565
	Existência	428
	Entendimento	94
	Propósito	67
Movimento Causado	*Mover*	1284
Resultativa	*Fazer*	293
Causativa	*Fazer*	872
Intransitiva de Movimento	*Mover*	476
Bitransitiva	*Transferir*	66
Total		9210

Categorização sintática das construções

Construção	Estrutura Sintática	Sentido	Paráfrase
Transitiva de *obtenção*	$Suj_{[agente]}$ V Obj *I got good results in the test*	*obtenção*	*obtain, grab, collect*
Transitiva de *recebimento*	$Suj_{[paciente]}$ V Obj *She got a letter from her*	*recebimento*	*receive*
Transitiva de *posse*	$Suj_{[agente]}$ GOT Obj *She's got two kids*	*posse*	*have, own, possess*
Transitiva *existencial*	Suj GOT Obj (Obl) *We got a problem at home*	*existencial*	*there to be*
Transitiva de *entendimento*	Suj V Obj *I didn't get the message*	*entendimento*	*understand*
Transitiva de *propósito*	Suj V Obj $V_{infinitivo}$ *I'll get the money to pay the rent*	*obtenção*	*obtain, grab, collect*
Causativa Analítica com part. pass.	Suj V Obj $V_{part.pass.}$ *He got his job done in time.*	*causar a realização de algo*	*have, make*
Causativa Analítica com *-ing*	Suj V Obj V_{-ing} *We have to get this thing going*	*causar a movimentação de algo*	*have, make*
Causativa Analítica com infinitivo	Suj V Obj $V_{inf.}$ *She got her kids to eat healthily*	*fazer alguém agir*	*have, make*
Movimento Causado	Suj V Obj Obl *... the internet will get more people into the stores*	*fazer mover para algum lugar*	*make*
Resultativa	Suj V Obj $Obl_{SAdj/SPrep.}$ *I'll get the kids ready for school*	*fazer mudar de estado*	*make*

Corpus: a representatividade e as perguntas de pesquisa

Construção	Sentido Central	Freq. Abs.	Freq. Relativa
Transitiva com sentido de *obtenção*	*Obtain*	4192	45,51%
Movimento Causado	*Move*	1284	13,94%
Transitiva com sentido de *recebimento*	*Receive*	874	9,48%
Causativa Analítica	*Make*	872	9,46%
Transitiva com sentido de *posse*	*Have*	565	6,13%
Intransitiva de Movimento	*Move*	476	5,16%
Transitiva com sentido *existencial*	*Existential*	428	4,64%
Resultativa	*Make*	293	3,18%
Transitiva com sentido de *entendimento*	*Understand*	94	1,02%
Transitiva com sentido de *propósito*	*Obtain*	67	0,72%
Bitransitiva	*Transfer*	66	0,71%
Total		9210	

Linguística de *Corpus*: vantagens da utilização de dados reais

- Dados de *corpora* são mais objetivos do que dados baseados na introspecção.
- Dados de *corpora* podem ser facilmente verificados por outros pesquisadores.
- Dados de *corpora* são necessários para o estudo da variação de dialetos, registros e estilos.
- Dados de *corpora* podem fornecer a frequência de ocorrência de itens linguísticos.
- Dados de *corpora* são essenciais para uma série de campos de aplicação tais como ensino e tecnologia (tradução automática, síntese de fala, plataformas de busca por palavras-chave etc.)
- Dados de *corpora* também se prestam à generalizações linguísticas (para os *corpora* em questão)
- Dados de *corpora* democratizam o acesso à informação tanto entre falantes nativos como não nativos da língua.

Startvik (1992: 08-10)

Muito obrigado!

Rodrigo_21usp@yahoo.com.br