Corpora as Digital Humanities Tools for Learning Foreign Languages

Iryna Dilay
irynadilayi@mail.montclair.edu

Language teaching practice is becoming more learner-centered

As an assistant tool in language learning, corpus makes an easy and quick analysis of the greatest amounts of linguistic data possible, and the learners are provided with a new approach to learn a language independently.

What is a corpus?

CORPUS is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair 1996, 2004).

Corpora are the main language data base.

Pedagogic utility

Sinclair: ***"One does not study all of botany by making artificial flowers"*** (1991, 6)

1987 - the publication of Collins COBUILD English Language Dictionary, the first corpus-based dictionary for learners.

1988 – Johns' paper on the use of corpus-derived and corpus-based materials in the language classroom

1990 – Sinclair's Collins COBUILD English Grammar

Biber et al., 1998; Hunston, 2002; Kennedy, 1998; McEnery & Wilson, 2001; McEnery et al., 2005, in press; Meyer, 2002; Partington, 1996; Stubbs, 1996, 2001; Tognini-Bonelli, 2001;Tomlinson ,1998, etc.

Useful links

- http://corpus.byu.edu - Mark Davies' site: COCA, COHA, BNC, GloWbE, CORE, Now, etc
- http://www.wordandphrase.info/ - COCA based
- http://corpus.byu.edu/coca/files/Teaching_Vocabulary_Through_DDL.pdf
- http://www.lextutor.ca/ Compleat Lexical Tutor
- http://www.linguee.com/ Linguee – a bilingual multidictionary supported by parallel translations.
- http://projects.ael.uni-tuebingen.de/backbone/moodle/ Backbone - Pedagogic Corpora for Content & Language Integrated Learning
- http://www.corpustool.com/ UAM Corpus Tool (to analyze your own corpus)
- http://www.papyr.com/applets/concordancer/

BENEFITS OF CORPORA FOR LANGUAGE TEACHING

- More accurate descriptions of language than textbooks/intuitions
- Exposure to contextualized, meaningful language in 'real' usages
- Examples of specific registers/genres of language
- Reference tool for independent/autonomous language investigation and learning
- Cited from: Jonathan Smart, Northern Arizona University

Pedagogical use

- **Corpus-based direct ESL/EFL teaching**
- **Corpus-based indirect ESL/EFL teaching**
- **Self-study**
- **Syllabus design**
- **Materials development**
- **Language testing**
- **Teaching LSP**
- **Teaching academic English**
- **L1-focused corpora – developmental (CHILDES)**
- **L2-focused corpora – learner corpora**

Corpus-based vs Corpus-driven approaches

***Elena Tognini-Bonelli (2001*) distinguishes between a corpus-based and corpus-driven approaches.**

- **The corpus-based approach (CBA) is a method that uses an underlying corpus as an inventory of language data. From this repository, appropriate material is extracted to support intuitive knowledge, to verify expectations, to allow linguistic phenomena to be quantified, and to find proof for existing theories or to retrieve illustrative samples. It is a method where the corpus is interrogated and data is used to confirm linguistic pre-set explanations and assumptions. It acts, therefore, as additional supporting material.**

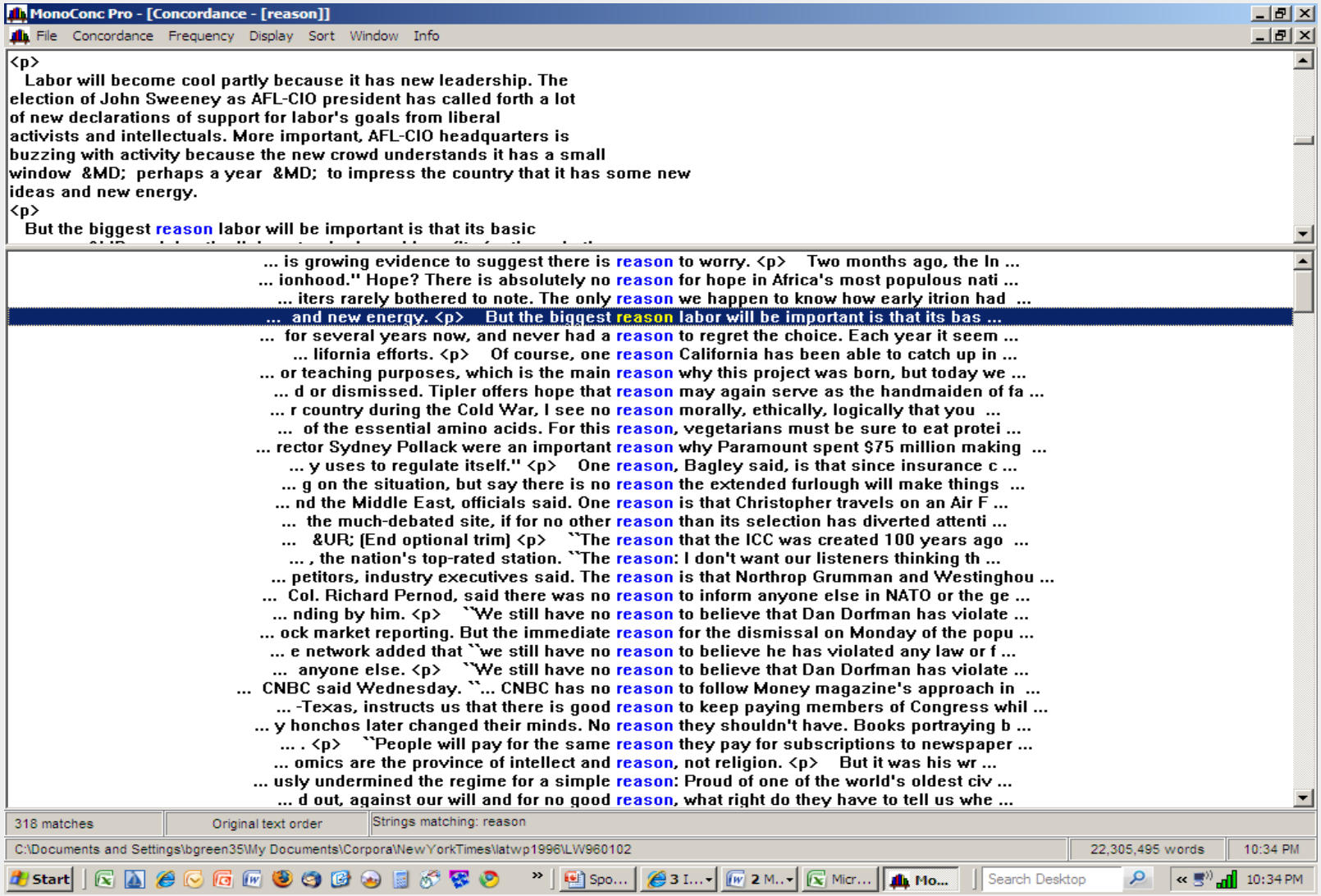
- **The corpus-driven approach (CDA) is a methodology**

Corpus-based (corpus-informed) learning

- **Deductive Approach**

Through a deductive approach, the definition and explanations of a word can be tested by running a concordancer. The teacher and the students are then exposed to a large number of contexts of the same word to confirm what is learned so as to refine the original generalizations. This, therefore, assists the student in exploring the language in great detail and thereby gaining further insights into its structure.

Key Word in Context (KWIC)



Corpus-based learning

- **Establishing Word Frequency**

All the students are motivated to find out the frequency of some word in the unit, then the figures are gathered and the frequency list of each unit takes shape. The list helps the students have a sense of the distribution of the word in vocabulary and acts as an important guide in their vocabulary learning.

The frequency of *know* across varieties

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	3077622	753024	214497	622175	149328	243370	121207	137269	62090	75692	51523	72454	70083	78759	49656	66918	85018	55546	59812	45022	64179
PERMIL	1,633.06	1,946.76	1,591.63	1,605.14	1,478.07	1,642.08	1,489.20	1,423.50	1,332.89	1,473.55	1,304.86	1,685.97	1,682.92	1,821.08	1,227.58	1,475.12	1,993.57	1,432.77	1,456.65	1,280.49	1,621.92

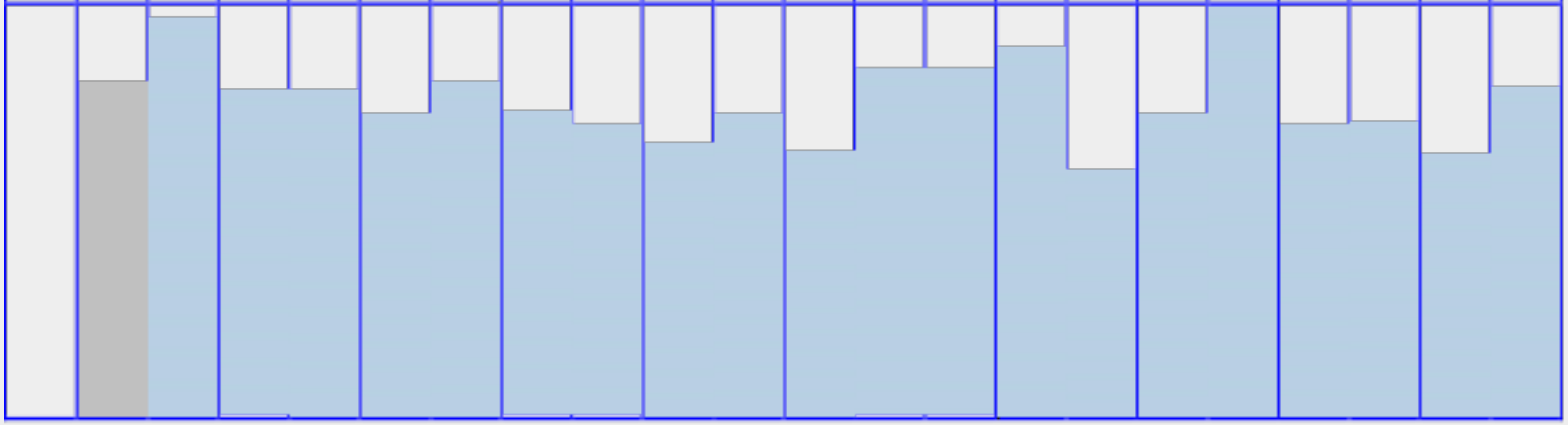


Table 1. Distribution of *know* across GloWbE varieties

The frequency of *think* across varieties

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	2824266	745325	190665	684665	131165	247302	116530	103226	44359	50701	36026	68018	55212	54165	42854	53744	53950	37256	39506	28554	41043
PERMIL	1,498.63	1,926.85	1,414.79	1,766.35	1,298.29	1,668.61	1,431.74	1,070.47	952.25	987.03	912.38	1,582.75	1,325.82	1,252.42	1,059.42	1,184.71	1,265.06	960.99	962.12	812.12	1,037.23

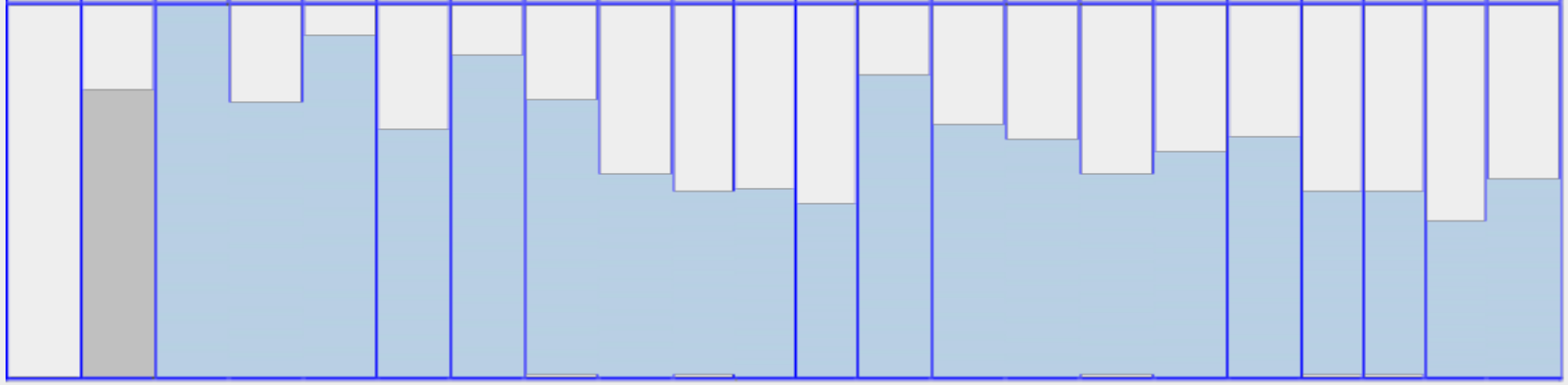


Table 1. Distribution of *think* across GloWbE varieties

Democracy over time (COHA)

Corpus of Historical Amer | Full-Text Corpus Data: ba | How to take a screensh

corpus.byu.edu/coha/

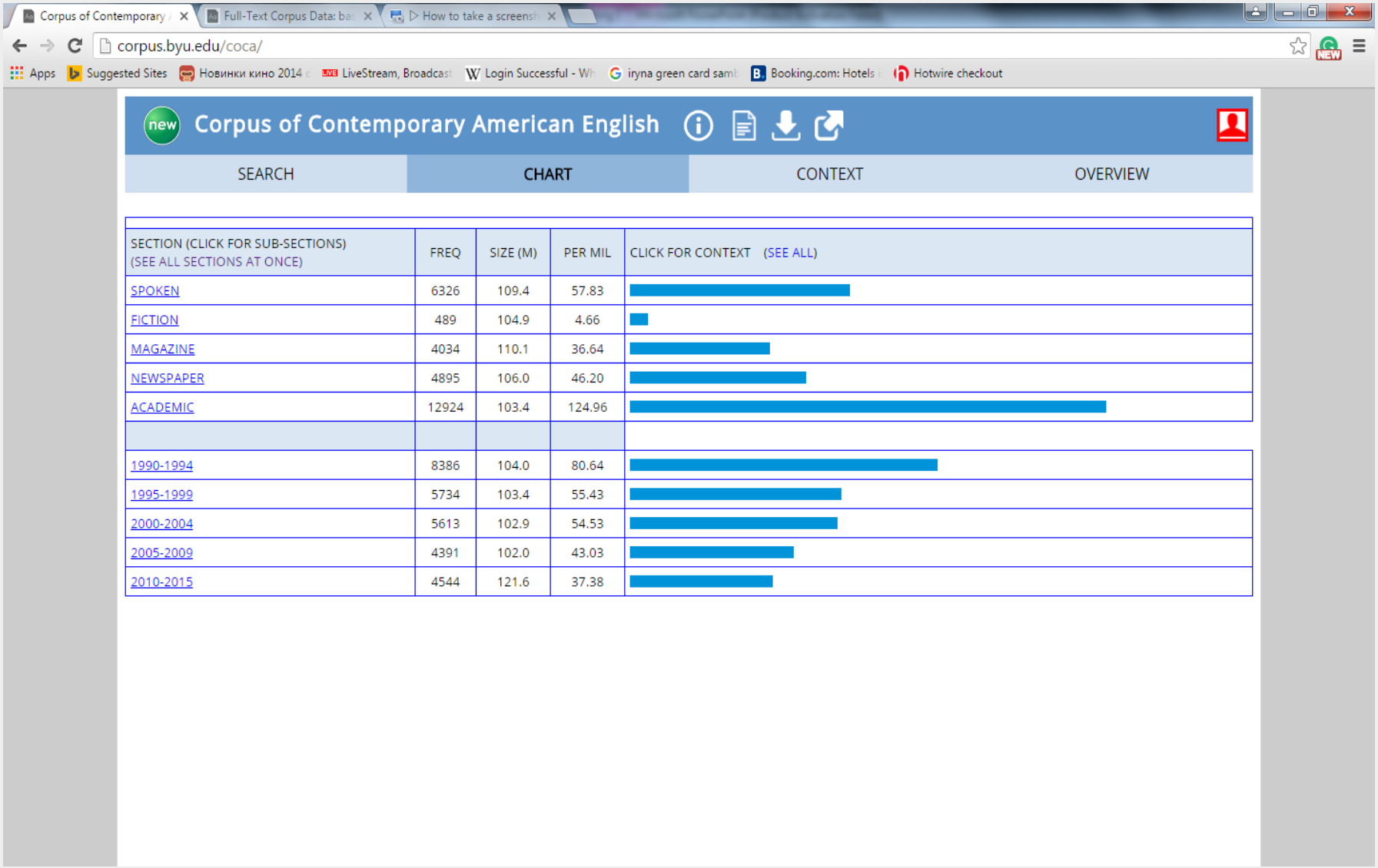
Apps | Suggested Sites | Новинки кино 2014 | LiveStream, Broadcast | Login Successful - Wh | iryna green card sam | Booking.com: Hotels | Hotwire checkout

new **Corpus of Historical American English**

SEARCH | CHART | CONTEXT | INTERFACE

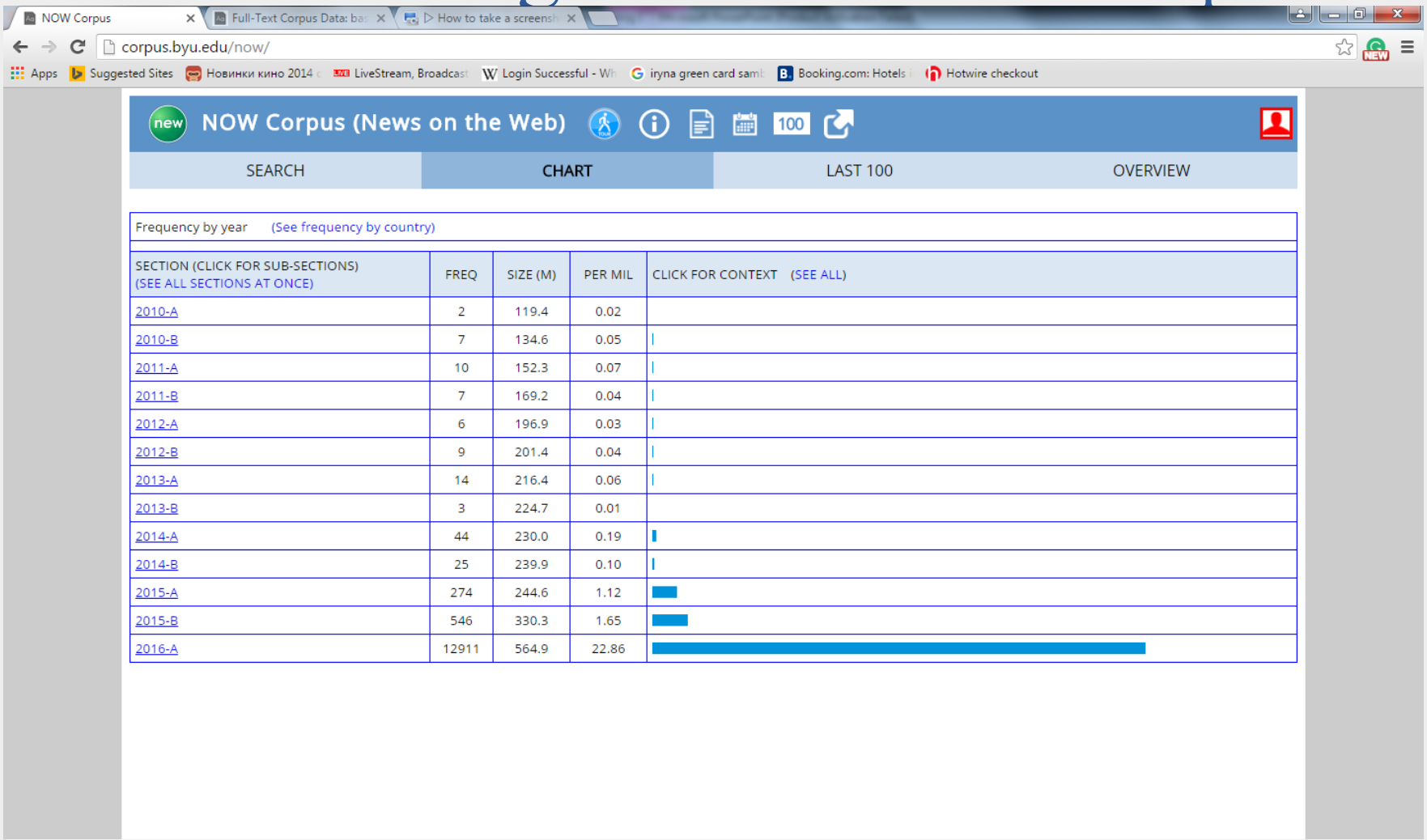
SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
1810	10	1.2	8.47	
1820	30	6.9	4.33	
1830	332	13.8	24.10	
1840	429	16.0	26.73	
1850	493	16.5	29.93	
1860	506	17.1	29.67	
1870	453	18.6	24.40	
1880	512	20.3	25.20	
1890	643	20.6	31.21	
1900	613	22.1	27.74	
1910	1098	22.7	48.37	
1920	1207	25.7	47.05	
1930	1280	24.6	52.03	
1940	1851	24.3	76.02	
1950	1046	24.5	42.62	
1960	828	24.0	34.53	
1970	618	23.8	25.95	
1980	1088	25.3	42.98	
1990	1111	27.9	39.76	
2000	913	29.6	30.88	

Democracy in COCA



SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
SPOKEN	6326	109.4	57.83	
FICTION	489	104.9	4.66	
MAGAZINE	4034	110.1	36.64	
NEWSPAPER	4895	106.0	46.20	
ACADEMIC	12924	103.4	124.96	
1990-1994	8386	104.0	80.64	
1995-1999	5734	103.4	55.43	
2000-2004	5613	102.9	54.53	
2005-2009	4391	102.0	43.03	
2010-2015	4544	121.6	37.38	

New coinages: *Brexit* in Now Corpus



NOW Corpus | Full-Text Corpus Data: ba | How to take a screensh

corpus.byu.edu/now/

Apps | Suggested Sites | Новинки кино 2014 | LiveStream, Broadcast | Login Successful - Wh | iryna green card samb | Booking.com: Hotels | Hotwire checkout

new NOW Corpus (News on the Web) 100

SEARCH | CHART | LAST 100 | OVERVIEW

Frequency by year (See frequency by country)

SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
2010-A	2	119.4	0.02	
2010-B	7	134.6	0.05	
2011-A	10	152.3	0.07	
2011-B	7	169.2	0.04	
2012-A	6	196.9	0.03	
2012-B	9	201.4	0.04	
2013-A	14	216.4	0.06	
2013-B	3	224.7	0.01	
2014-A	44	230.0	0.19	
2014-B	25	239.9	0.10	
2015-A	274	244.6	1.12	
2015-B	546	330.3	1.65	
2016-A	12911	564.9	22.86	

Collocations: adj + *rain* in COCA

			FREQ		ALL	%	MI	
1		HEAVY	1388		45597	3.04	5.72	
2		COLD	464		66639	0.70	3.59	
3		TROPICAL	287		8466	3.39	5.87	
4		POURING	260		5519	4.71	6.35	
5		TORRENTIAL	238		416	57.21	9.95	
6		FREEZING	205		5009	4.09	6.14	
7		DRIVING	185		36030	0.51	3.15	
8		STEADY	182		15670	1.16	4.33	
9		WET	140		20525	0.68	3.56	
10		PURPLE	100		10476	0.95	4.04	
11		GENTLE	77		10032	0.77	3.73	
12		TEMPERATE	67		1329	5.04	6.44	
13		HEAVIER	64		4331	1.48	4.67	
14		ICY	55		4373	1.26	4.44	

From ‘three Ps’ to ‘three Is’

Presentation – practice – production =corpus-based

Illustration – interaction – induction (Carter and McCarthy (1995) **=corpus-driven**:

a) looking at real data;
b) discussing and sharing opinions and observations;
c) making one’s own rules.

Data-driven learning – a learner is a researcher! (Johns (1991), Leech (1997), Flowerdew (1993), etc.)

Corpus-driven (data-driven) learning

- **Inductive Approach**

The study of some key words can be conducted through this approach. The teacher can ask the students to observe and discuss the concordance before explaining the word. Working with corpora does not necessarily require students sitting at computer terminals. The students are motivated in their learning of English and at the same time, learn some important skills that would make their learning less intimidating and keep them in regular contact with English.

- **Establishing Colligation, Collocation and Semantic Prosody**

The computerized corpus, with sufficient, authentic and typical data, enables users to conduct comprehensive and overall research into collocational patterns. Once patterns of non-native deviance have been discovered, students can be explicitly made aware of these patterns, and they will eventually be able to modify their linguistic behavior into a more native-like direction.

Example: to bear _____ resemblance

CONCORDANCE:

- All of this **bears no resemblance** to the Europe of today. …
- Queuing up at the polls at election time **bears a strong resemblance** to a gathering of the herd at a...
- Most **bear little resemblance** to their wild ancestors.
- ….which is under fire for a tea kettle ad that **bears an uncanny resemblance** to Adolf Hitler.
- Their stories **bear a striking resemblance**.

Leaning vocabulary

In isolation:

- through naming objects;
- through translation;
- through definitions;
- through associations: synonyms, antonyms, word families, etc.

In context:

“You shall know a word by the company it keeps” (Firth).

Lexical Approach by Lewis (1993) – learning a word in lexical chunks (collocations).

Concordance

"Learning a word from either a short definition or a single sentence context tends to produce inert lexical knowledge that does not facilitate the word's comprehension in a novel context, while learning a word from several contexts, with or without a definition, tends to produce rich, transferable knowledge" (Mezynski, 1983).

What is the missing word?

The prime minister didn't offer any (…) alternatives.

What is the word now?

The prime minister didn't offer any (…) alternatives.

Plans can not be expected to continue losing money year after year and remain (...) .

Mr. Jeffries could be a (…) candidate for mayor as early as 2017.

They are totally not economically (…) in rural areas.

Only in rare cases is the embryo (…), and the condition can be lethal for the mother.

Stevens' Experiment (1991)

The task was to have students recall a known word to fill a gap in a text, which was either a single gapped sentence or a set of gapped concordance lines. Learners retrieved words from memory more successfully when cued by the concordance lines in spite of their chopped-off nature.

Learning Synonyms

The use and differentiation of synonyms is usually a problem for many learners of English. The best way to help them to deal with this problem is exposing them to ample examples of real English,

which can be obtained easily via a large English corpus. In this aspect,

concordances can play a unique role here.

Establishing connotations:

- *Childlike, youthful, childish, young, immature, juvenile*
- *Chubby, fat, plump, overweight*
- *Confident, secure, proud, egotistical*

Semantic prosody of synonyms

'Whichever way you look at it, it is a fact that "the meaning of a word can often be illuminated by the other words which it tends to co-occur with" (Wierzbicka 1987, 21), so that e.g.

comparing the adverbs which the verbs *rebuke, reprimand* and *reprove* tend to co-occur with, will yield important clues as to the semantic differences between them:

***rebuking* tends to be done *sharply*, whereas *reprimanding* tends to be done severely; only reproving can be done *gently* but cannot be done *sharply, severely*.**

Compleat Lexical Tutor

Go to

http://www.lextutor.ca/

Learning grammar

Questioning the NORM!

- **Progressives with stative verbs**
- **Prepositions**
- **Verb complementation (Infinitive or Gerund?)**

- **Reflexive pronouns**
- **Countable vs uncountable nouns**
- **Subject–predicate agreement**
- **Clauses**
- **Articles**
- **…**

Error analysis

- **Misuse of Words**

The purpose of error analysis is to seek the causes of the errors and the law of learning a language, but finding out the representative problems or errors requires the amount of the learner's output to be as large as possible. Hence, the analysis of the learners' interlanguage would be much more efficient if the outputs are transferred into an electronic corpus

accumulatively from learners and then explored via a computer program.

Leaner Corpora

Leaner corpora are sources of error analysis for different L2 leaners, especially useful for educators.

Learner Corpus Association (http://www.learnercorpusassociation.org/).

ICLE (The International Corpus of Learner English), Louvain, Belgium (written)

ITAcorp (The International Teaching Assistants Corpus), Pennsylvania State University, USA (spoken)

ICCI (The International Corpus of Crosslinguistic Interlanguage), Tokyo University of Foreign Studies, Japan (written)

CLEC (The Chinese Learner English Corpus), China (written)

Importance of Learner Corpora

The interface between L1 and L2 materials.

Frequency information from native-speaker corpora alone is not sufficient to inform curriculum and materials design. Rather, 'it is important to strike a balance between frequency, difficulty and pedagogical relevance. That is exactly where learner corpus research comes into play to help weigh the importance of each of these' (Meunier 2002: 123).

Shortcomings and caveats of using corpora

- **Learners are overwhelmed by the extremely large number of examples generated by their searches.**
- **Giving undue prominence to what is simply frequent at the expense of rarer but more effective or salient expressions (Cook 1998:61)**
- **Corpus data are traces of texts rather than discourse and must be recontextualized in language teaching (Widdowson 1990,2000)**
- **Corpus samples may have to be adapted when used with low levels or young learners, and when corpus examples contravene or offend sociocultural norms and customs.**
- **The need for learners and teachers to acquire new skills and assume new roles in the studying process.**
- **The need to make corpora more user-friendly.**

Conclusion

In a learner-centered classroom, students are seen as being able to assume a more active and participatory role than in traditional approaches.

Corpora can be very useful tools for language learners, since they allow these learners to quickly and easily see how native speakers use the language in a wide variety of naturally occurring texts. Corpus-based research and teaching has the potential to empower non-native teachers and researchers, since native speaker introspection is no longer considered the one infallible source of insights into language structure and use.

- **Corpus use is not meant to replace existing teaching**

Bibliography

- Aijmer, K. (Ed.). (2009). *Corpora and language teaching.* Amsterdam, The Netherlands: John Benjamins.
- Boulton, A., Carter-Thomas, S., & Rowley-Jolivet, E. (Eds.). (2012). *Corpus-informed research and learning in ESP: Issues and applications.* Amsterdam, The Netherlands: John Benjamins.
- Campoy-Cubillo, M. C., Fortuño, B. B., & Gea-Valor, M. L. (Eds.). (2010). *Corpus-based approaches to English language teaching.* London, UK: Continuum.
- Flowerdew, J. (2009). Corpora in language teaching. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 327-350).. Oxford: Wiley-Blackwell.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, England: Cambridge University Press.
- Jones, M., & Durrant, P. (2010). What can a corpus tell us about vocabulary teaching materials? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics.* London, UK: Routledge.
- McCarthy, M. J. (2004). Using corpora in language teaching. *CALPER Digests*. Retrieved from http://calper.la.psu.edu/publications.php.
- McEnery, T. (2011). What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 364-380), New York, NY: Routledge.
- Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung and S. Petch-Tyson (eds), Computer learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 119-142. Philadelphia: John Benjamins.
- Reppen, R. (2010). Using corpora in the language classroom. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 35-50). Cambridge, UK: Cambridge University Press.
- Sinclair, J. M. (2004). (Ed.). *How to use corpora in language teaching*. Amsterdam, Netherlands: John Benjamins.
- Stevens. V. (1995). Concordancing with language learners: Why? When? What? *CAELL Journal, 6*(2), 2-10.
- Sun, Y., & Wang, L. (2003). Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer Assisted Language Learning, 16*, 83-94.