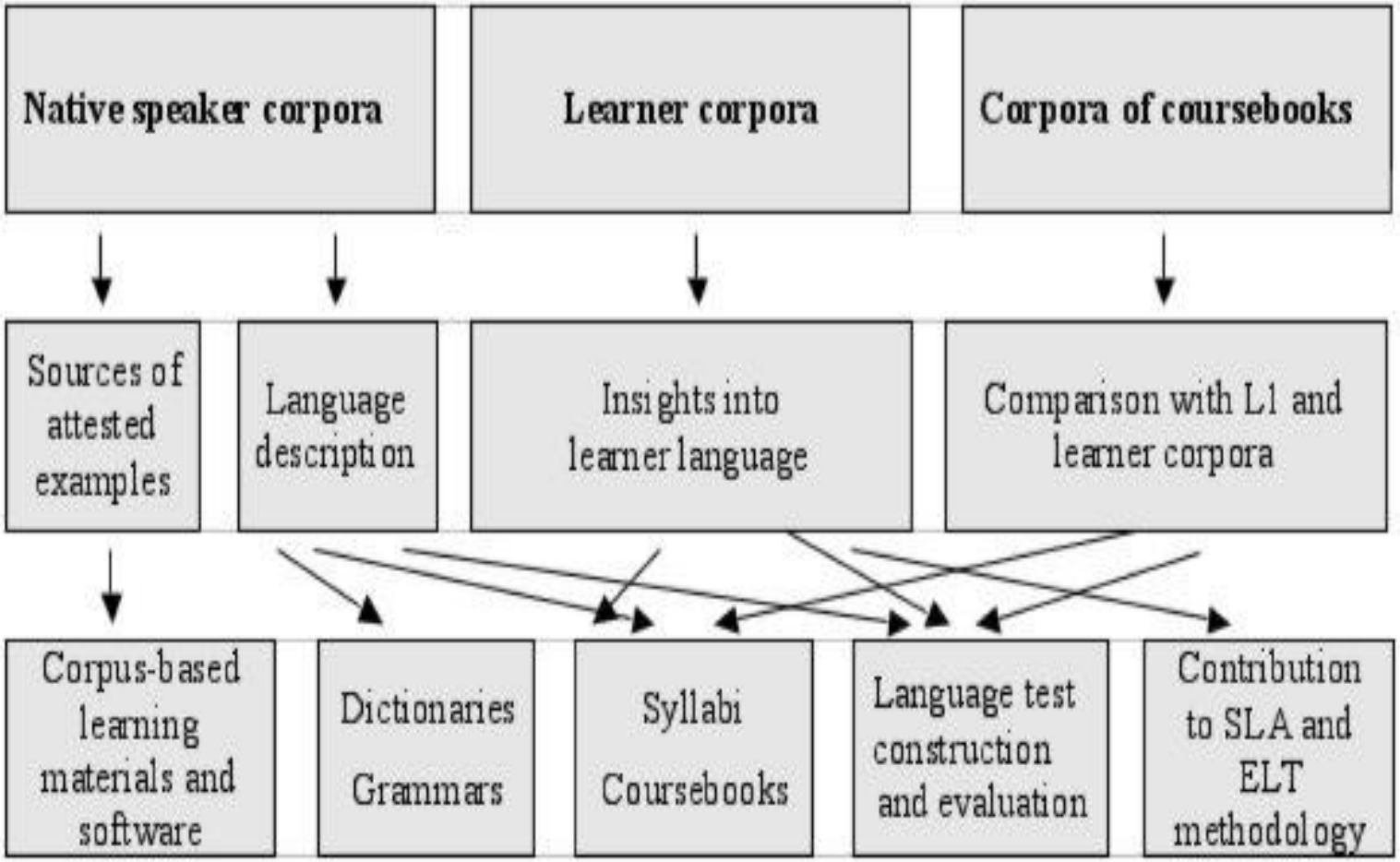
The corpus, the teacher's best friend

Elena Tarasheva, PhD
New Bulgarian University

20th BETA Annual Conference
29th April – 1st May 2011, Sofia, Bulgaria

Type of corpora and their use:



Corpora and Language Teaching: Just a fling or wedding bells? Costas Gabrielatos, personal website

Native Speaker Corpora

- The British National Corpus
- Freely available at:

  http://www.natcorp.ox.ac.uk/

- My task – get students to elicit attributes used with *course* (corpus based learning materials)

The search ATTRIBUTE + COURSE

- B3C **896** Most will have taken a Foundation **course** in Art and Design,
- CCM **2509** and sent to a secretarial **course** when she had her heart set on going to university
- FUA **373** Framework English is specifically described as a self-study **course**.
- GoW **814** The ways in which these topics would be introduced into a PGCE **course** would be as varied
- G4X **1049** And it may well be that we'll have to make sure everybody who's running a training **course**
- HTE **284** This means that you would be wise not to set your heart on one career but to keep options open, even if you are taking a vocational **course**.

NOTATION

- **HTE** *Edinburgh undergraduate prospectus.* Smith, David. Portsmouth: Grosvenor Press, 1993, pp. 1-157. 2716 s-units, 55423 words

  gives an idea about the source of the material
- Which word in the text
- What part of speech (POS)

Multi modal corpora

- **Michigan Corpus of Academic Spoken English**

http://micase.elicorpora.info/sound-files-online

Native speaker corpora – language test evaluation

- Marking students' written or oral production, have you asked yourself – can you say this?
- How can you check what is acceptable or what is not?

- BNC answers to a case study – what to say at Easter as a greeting?

A BNC search returns:

- ARG **918** Christ has risen again. - 1 solution
- Christ is risen – 7 solutions
- Christ has been resurrected – 0 hits

Why not Google?

- "Christ has been resurrected" – returns 84,700 results.

How authoritative is that?

Christ has been resurrected from the dead - 1Cor 15:3-8: a direct quotation from the Bible

But do people say it?

BNC is:

- **Monolingual:** It deals with modern British English, not other languages used in Britain. However non-British English and foreign language words do occur in the corpus.
- **Synchronic:** It covers British English of the late twentieth century, rather than the historical development which produced it.
- **General:** It includes many different styles and varieties, and is not limited to any particular subject field, genre or register. In particular, it contains examples of both spoken and written language.
- **Sample:** For written sources, samples of 45,000 words are taken from various parts of single-author texts. Shorter texts up to a maximum of 45,000 words, or multi-author texts such as magazines and newspapers, are included in full. Sampling allows for a wider coverage of texts within the 100 million limit, and avoids over-representing idiosyncratic texts.

Google/BNC

- Oriented to one language variety – we do not want one British word, one Australian, one American etc in the same sentence
- It may be the case that one organisation has created too many web-sites and 'drowned' cyberspace with its idiosyncratic language use
- A balanced corpus tries to be representative of a language variety, Google embraces language samples regardless of their representativeness.

Learner corpora

Different corpora exist:

- English children learning Spanish and French:

http://www.splloc.soton.ac.uk/dosearch.php

Where you can hear sound files recording the students performing different tasks at different levels and see the development

- The Cambridge Learner Corpus – paid
- Japanese students learning English

My own learner corpus

- I set my students the task to analyse a film – Michael Palin's documentary about Bulgaria – following a plan
- The first draft was collected in a corpus, as were the second drafts in a second corpus for analysis and comparison.
- Analytical procedures: word list, key words, concordances of the keywords; comparison between the two.
- Corpus size: first draft - 5000 words in 14 works

  second draft – 3 000 words in 9 works

Word lists

- Software – WordSmith Tools (Scott 1999)
- The word list is a list of all the words in the texts, which can be arranged alphabetically, or in the order of frequency.
- The word list was considered indicative of linking devices to check the coherence of the texts

Linking devices:

Word	1st draft	2nd draft
For example	3	6
however	1	6
therefore	2	7
Secondly	1	5
Also	7	4
Despite	2	1

Key words

- The frequency of the words in the students' texts is compared to the frequency of the respective words in the BNC
- A statistical procedure conducted by the Wordsmith based on chi square
- Normally, proper names, nouns specific for the about-ness of the text tend to be higher

Key words in the first draft:

Key word	Freq.	%	RC. Freq.	Keyness
GYPSIES	70	1.43	279	1,039.74
PALIN	45	0.92	63	746.35
BULGARIA	39	0.80	527	490.08
MESSAGE	33	0.67	6,770	237.96
AZIS	10	0.20	0	198.38
MICHAEL	29	0.59	9,107	184.75
PLOVDIV	9	0.18	4	162.50
DIDN	7	0.14	19	108.58
CREATED	12	0.24	8,695	56.85
INTERESTING	12	0.24	9,461	54.90
OPINION	11	0.22	7,405	53.67
ABOUT	38	0.78	192,022	48.75
IS	102	2.08	974,293	46.35
ARE	57	1.16	458,368	36.95
THIS	54	1.10	454,419	32.06
THE	389	7.94	6,055,105	26.90

Concordances: Gypsy

- A lot of people associate the word **gipsy** with a person who is dirty, disgusting, uneducated and unemployed
- I just would say that Palin is not well informed at all about the Bulgarians **gypsies**, how cruel and criminal they are. He is not doing a careful study
- But they do not know that the **gipsy** doesn't have the wish to work and feels good in this situation
- has chosen to show the weak aspect of Bulgaria shooting the poor area and its **gypsi** occupants
- I perceive this message as offensive and misrepresenting since the **gypsies** are not our native population and do not possess our spirit and individuality
- To me, as I already have a rather unchangeable opinion about **gipsies** and it's far from being positive, this message is another point of view

Concordances: Foreigners

- I know their life I am sad for my country and these people. I am sure that when **foreigners** watch this film they will not be respected by the state.
- The aim of this message is to prevent **foreigners** from visiting this poor country Bulgaria.
- or a person at the street instead of finding a job and working for salary. The **foreigners** would not believe this and would feel sorry about the gipsy people.
- that these people live in almost ruined houses without water and electricity. The **foreigners** would probably think that the country does nothing for the poor gipsy
- points of view are different because we know the real face of a gipsy and the **foreigners** believe what they see. We know that the gipsy people are lazy and pretentious

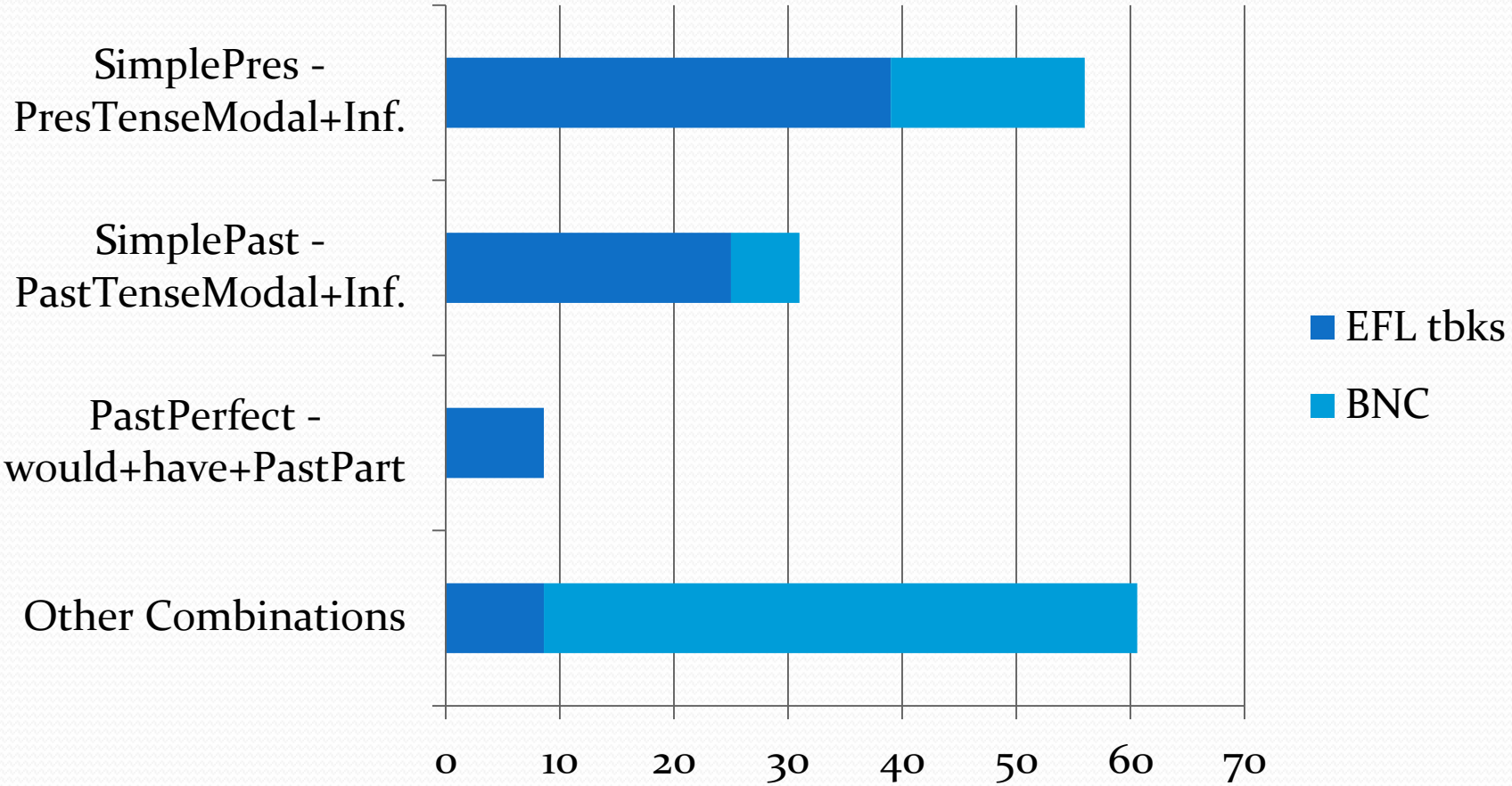
Corpora of course books

Ute Römer : Where the Computer Meets Language, Literature, and Pedagogy: Corpus Analysis in English Studies, personal website

Research question:

How frequent are diferent types of conditional sentences in English-teaching textbooks for German students and in naturally occurring texts

Comparison between the frequency of Conditionals



Conclusions:

- Native speaker corpora help develop materials for our students
- They also help us check which language uses are acceptable
- Learner corpora give indications about mistakes common to the whole population
- They also show whether progress has occurred in the course of learning
- Corpora of teaching materials contrasted to natural speech show how adequate a material is

References

- **Gabrielatos, C.** *Corpora and Language Teaching: Just a fling or wedding bells?* **personal website http://www.gabrielatos.com/Publications.htm**
- *The BNC Sampler*, XML version. 2005. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/
- **Michigan Corpus of Academic Spoken English** http://micase.umdl.umich.edu/m/micase/
- SPLLOC projects (http://www.splloc.soton.ac.uk)
- WordSmith Tools Scott, M., 2008, WordSmith Tools version 5, Liverpool: Lexical Analysis Software
- Ute Römer : Where the Computer Meets Language, Literature, and Pedagogy: Corpus Analysis in English Studies, personal website

http://www.uteroemer.com/publikationen.htm