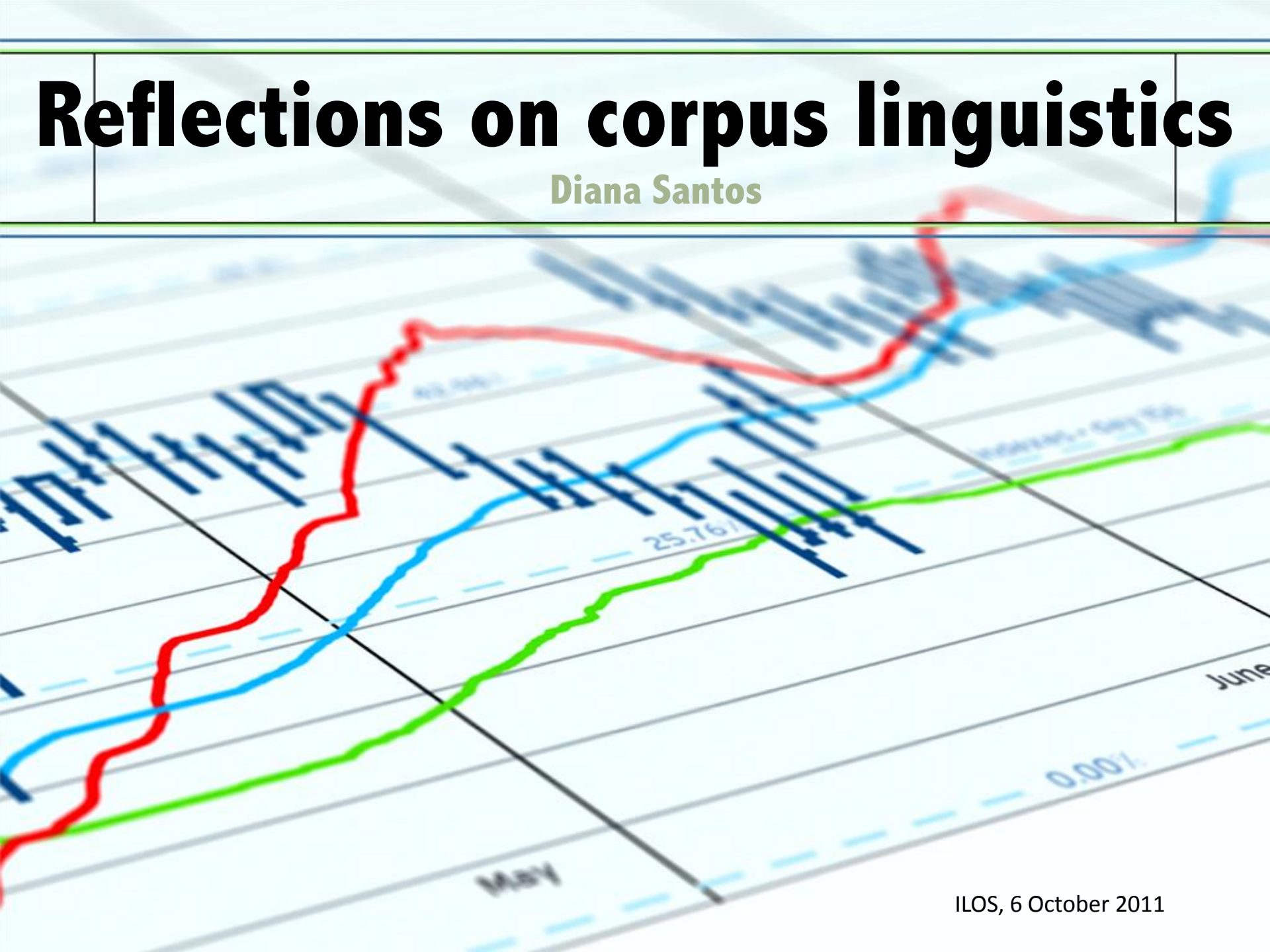
Reflections on corpus linguistics

Diana Santos



ILOS, 6 October 2011

What I will be talking about

Reflections on a 25-years practice

1. How people perceive corpora
   A lot of different views
2. Between compiling and using
   Off-the-shelf versus design
3. Examples of measurements and conclusions
   A wider range of approaches
4. A course on statistics in linguistics
   Would that be useful?
5. What is the defining “past-participle” of corpus linguistics?
   Inspired, based, derived, supported, illustrated, induced?

The perception of corpora

Widely different "perceptions"

You can have developers for own use.

You can have consumers

You can have corpus developers for others

- There is always a tension between corpus use and corpus creation

You can have corpus co-developers

- Collaborative corpus creation

Three objects make up a corpus: texts, CES, shell (including annotation)

Quite a nice trilogy, most people don't distinguish the three things

People usually only pay attention to the texts, and expect the rest to appear automatically!

Kinds of positive attitudes

one can have as far as corpora are concerned

Not everyone has the same view of a corpus and of the way it would be used

From the in principle positive (the more the better, if not, tant pis!) to those who devote their life to building them for others, there are a lot of ways to look at a corpus

Between compiling and using

Off-the-shelf vs. design

You can investigate a known issue in a known corpus: contributing to progress within CL

Or a radically new issue in a known corpus

Or a known issue in new corpora

- There is always a tension between corpus use and corpus creation

You have to devise a corpus for a new issue, or to supplement a given corpus with more data

You have to create a corpus for a new kind of "text" that was never investigated before

There is so much you can do with what is already available!

There are so many interesting research questions out there, if only there was a corpus

Kinds of uses

Users use corpora for a variety of purposes

- To get acquaintance with a theme
- To study a distribution according to the corpus parameters
- To correlate different observations
- To evaluate other people's claims
- To create data for other purposes (e.g. teaching, lexicography, terminology, indexing…)
- To create resources or systems based on the corpus, having the corpus underlying them

An accent box, click to edit the text inside.

An accent box, click to edit the text inside.

Measurements and conclusions

Some examples

This is not a broad overview!

MEASUREMENTS	CONCLUSIONS
Kinds of futures in English	Different teaching material
Frequency differences in two comparable corpora (LOB-Br)	Lexical divergence between varieties of English
Translation of clauses E-P	Wide variation in number
Difference in use of a construction in learner corpora	Need for different teaching strategies for diff. native speakers
Frequency differences in grammar	Different authors involved

The essence of corpus linguistics…

Or, what is our credo?

- To use authentic material
- To trust the (other) language users
- To believe in quantity
- To believe in distribution
- To believe in replicability
- To believe in shared resources, community of users, improvement through the masses
- To trust performance to arrive at competence, from *parole* (only positive) derive *langue* (system, both positive and negative)

There is a lot one can do with corpus material!

But how do you know the corpus is appropriate for what you want?

Why counts?

Or, the quantitative vs. qualitative debate

- yes or no, categorical answers
- subjunctive, indicative, imperative, not 56% ind.
- good, not 89.5% ok
- the girl, not "something fuzzy to which I assign 78% of girlhood, not being able to see well enough"
- *bald-headed, water, butter*

- All our categorization is qualitative… Why should counting over it enlighten us?

It allows one to predict further quantitative relationships

But not reason about language…

Reasons to count

Different answers

- language is probabilistic in nature
- knowing the probabilities we also learn about the surprise and the innovation and can measure style and quality of a text
- it helps teaching it, explaining the most frequent words/constructions first
- it helps dealing with language (storing it, indexing it, sending it across the channels, translating it) : it gives us power over the language

It is a must for applications!

But we still don't know how it works…

If you want to know how language works

This is what interests me most

- I want laws
- I want principles
- I want methods, processes
- I want concepts

- I am not sure I want numbers

- Simulation is not explanation.
- I do not believe we have a random generator in our mind

The most important concept is vagueness

How can you prove it?