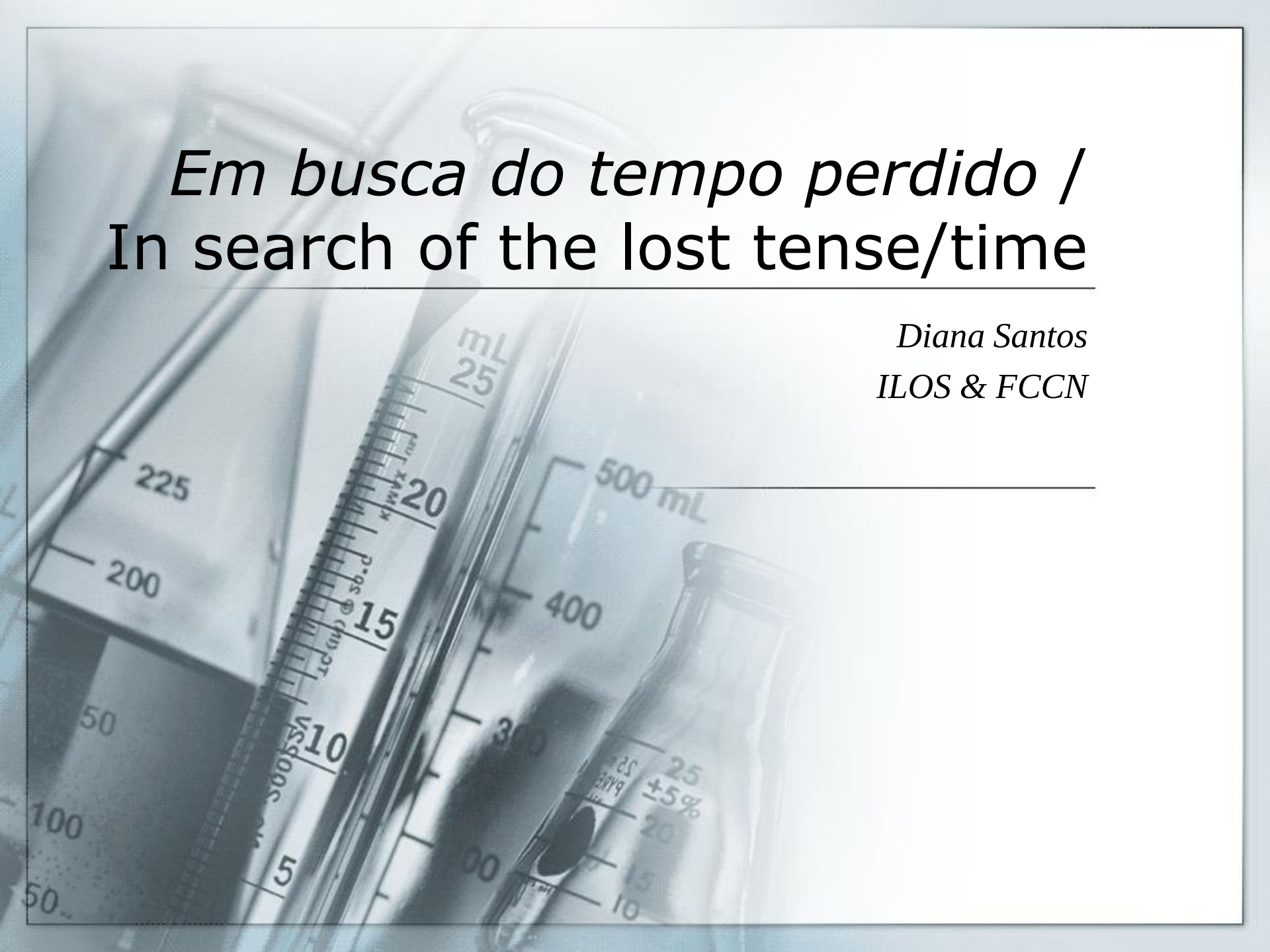
Em busca do tempo perdido / In search of the lost tense/time

Diana Santos

ILOS & FCCN



Once upon a time…

There was a young engineer who was fascinated by language and time, and dreamed of uncovering the depths of human mind in what concerned that most elusive of all concepts (time) in the new field of artificial intelligence (whose key was undoubtedly "natural language understanding").

So semantics of tense and aspect seemed to be the practical way to achieve it (since one has to choose "gavetas"/"skuffer" to study), and Lauri Carlson was the most clever guy I had ever met in person and who was in a position to supervise me.

So in 1990 I started a PhD in Informatics Engineering at Instituto Superor Técnico in Lisbon, with Lauri Carlson as supervisor and Amilcar Sernadas (mathematician) as co-supervisor.

What happened?

Six years later (June 1996), I delivered a 600 pages dissertation and the next year (17 January 1997) I got my PhD degree

Later I rewrote/ improved the text and it was published as a book in 2004, thanks to Stig Johansson who included me in his group at IBA and made it possible for me to get it accepted at Rodopi

In 2008 I was invited to write a chapter of Binnick's *Handbook of Tense and Aspect* on tense and translation (named "Translation")

I am currently trying to go (finally!) further by trying to empirically evaluate or assess the contents of my proposal (substantiated in my PhD)

In a nutshell

These are the main contributions of my thesis:

- A model for fine-grained description of translations, the translation network
- A (radically) new description of tense and aspect in Portuguese, with a set of Portuguese aspectual classes
- Several empirical studies based on parallel corpora
- Some methodological points, both on linguistic research, contrastive studies and translation analysis
- The concept of vagueness as a fundamental building block in a theory of natural language

At the same time ;-)

Other activities that may help me achieve this and/or contributed to my expertise (?), knowledge and standpoint(s)

- Development of tools for Portuguese corpora
- Teaching Portuguese grammar to Norwegians in 1995/1996
- Development of a public corpus infrastructure for Portuguese worldwide, the AC/DC project (98-now)
- Developing several parallel corpora (COMPARA: 99-08, CorTrad, 09-now)
- Organizing the HAREM evaluation contest (comparing semantic annotation of Portuguese) and taking part in the detailed revision of syntactic annotation of Portuguese (Floresta Sintá(c)tica)

Recently

- (2010) Developed an interest in statistical methods (as a reaction to the machine learning wave that is "infecting" NLP)
- (Jan 2011) Started as associate professor in Portuguese language at ILOS and (partially) merged the Linguateca and the ILOS research time into the same goals
- Decided to create a new, corpus-based, grammar of Portuguese based on the "AC/DC cluster" material
- Developed AC/DC-based tools for grammar teaching
  - Ensinador
  - PoNTE: Portuguese <-> Norwegian translation of short texts by ILOS students, for which I've just got a ILOS grant

Oh, really?

- Did you say: **Decided to create a new, corpus-based, grammar of Portuguese based on the "AC/DC cluster" material …**
- This is not so easy as you may expect… can you be a little more concrete, please?
- Do you know how long it took for English?
- It: Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & E. Finegan. *The Longman grammar of spoken and written English*. 1999, London: Longman.
- Maybe you should break this down into subparts?
- And what exactly is your material, and your method(s)?
- And by the way, what do you mean by "grammar"?

Materials

- The corpora included in the AC/DC
- The parser that annotated them
- The semantic domains included
- The information that is available
- The quality of the material

Procedure: for tense and aspect

- Choose the (semantic) aspects that seem more relevant/fundamental in your language
- Check or make sure you can detect them automatically
- Gather quantitative data on them
- Gather lexical data on them
- Gather interesting facts on them
- Write a quantitative-inspired description,
- Present the data using visualization capabilities in R
- Illustrate with good examples

Some apetizers…

Loosely based on the structure of the grammar course ☺

- Simple present vs. progressive
- Imperfeito, Perfeito, PPC and "neutral" verbs
- Passive (*ser*, *estar*, *ficar*) vs. non-passive
- Reflexive use (or use of reflexive pronouns)
- Future vs. periphrastic future
- Use of aspectualizers (*andar*, *começar a*, …)
- Use of Presente in the future sense
- Fundamental verbs (*passar*, *andar*, *ir*)
- Use of subjunctive in completives
- Some temporal conjunctions
- Temporal quantification and *vez*

Some *aperitifs…*

Again, loosely based on the structure of the grammar course ☺

- The domain of colour
- Physical appearance
- Kinds of adjectives (before or after)
- Emotions
- Place(s) and placement
- Possessives (position and system)
- Diminutives
- Quantifiers and the partitive construction

Conclusion: the beginning

An idea of what should be in the grammar, and preliminary proof that data can be gathered on these subjects

Some possible authentic examples of relevant distinctions and uses

Some idea of the quantitative distributions of these phenomena

But

How different or interesting will be such a grammar?

Is it worth while, or “yet another grammar”?

The most important* issue, is

How well (or not) will a corpus-based investigation support or even help the model of tense and aspect put forward in my thesis?

Or are they simply incommensurable?

In other words:

Is it possible – by any means whatsoever – replicate, or reproduce, in computational terms, the generalization/induction done by me as a human researcher to come up with the model I want to validate?

* For me, and for this presentation

Doubts

Interpretation is not available in corpora

- Except in translation corpora (and only when the languages differ in the overt marks and we know reasonably clear what one means)
- Or when annotation has been performed (which is extremely time-consuming and hard to reach a consensus)

Formal markings have always multiple uses

- It would be extremely uneconomical (and un-natural language) that the same marker only had one function
- Parsing is precisely the task of deciding which (set of) meaning(s) are relevant in the particular case

So, how can one arrive at semantics from the surface?

Acknowledgements

Linguateca

Titan cluster, run by Research Computing Services (USIT)
NoTur, project PILS (**P**ortugisisk **I**LOS-**L**inguateca **S**amarbeid)