MorDebe & NeoTrack

Using Neologism Research for Lexical Resources

Maarten Janssen

Overview of the Talk

- NeoTrack
  - Brief Presentation
- MorDebe
  - Morphological DataBase
  - Design, Philosophy, Solutions
- Open Source Lexical Information Network
  - *Portal da L ▪ íngua Portuguesa*

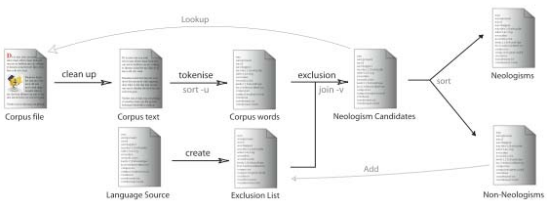
Part I : NeoTrack

Semi-Automatic Neologism Detection

NeoTrack - Design

- Semi-automatic neologism detection
  - Orthographic neologism
  - *Diário de Notícias, Público*
- Completely Web Based
  - Accessible by any computer anywhere
- Extended Lexicographic Diachronic Criterion
  - Neologism Candidate list by lexicographic exclusion
  - Candidate verification by Corpus lookup

NeoTrack Flow Chart



1. Create list of words in a text
2. Subtract the list of known words (Exclusion List)
3. Verify whether the results are neologisms

Exclusion List Subtraction

Corpus Words		Exclusion List		Neologism Candidates
a and as be **belissimo** bold but cleaned codes first **HTML** in interesting is it Italians italics later might need nice order other proceed process **raelly** say text that the this to up with would	-	a aalii aardvark aardvarks aardwolf aardwolves aba abac abacs abaca abacas aback aback abactinal abaft abalone abalones	=	belissimo HTML raelly

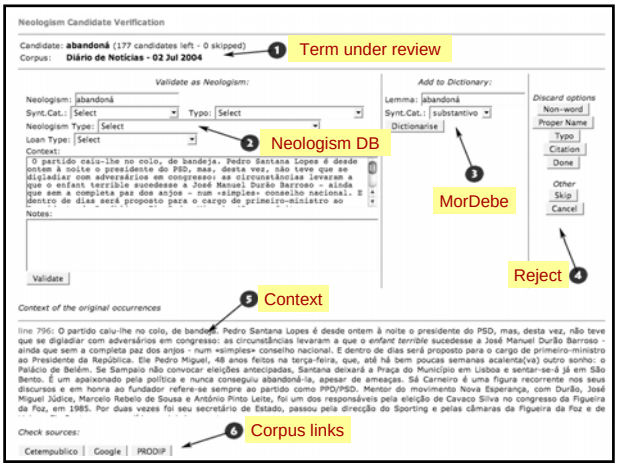
Technically : **cat corpuswords.txt | sort -u | join -v 1 - exclusionlist.txt**

Neologism Candidate Sorting

- Verify each candidate manually
- Reject non-words
  - Codes
  - Proper names
  - Typographic errors
  - Foreign language citations / stretches
- Verify the correct candidates against corpora
  - Because of the incompleteness of dictionaries
  - Words that occur >10 in reference corpus non-neologisms

Reference Corpus

- Large good corpora
  - *CETEMPublico, AC/DC, CLUL, REDIP*
  - Texts > 3 years old
- Manually verification
  - Corpus occurrences same criteria:
  - No proper names, typos, foreign quotes, codes
- Recurring words non-neologistic
  - 10 occurrences



Loss of Work

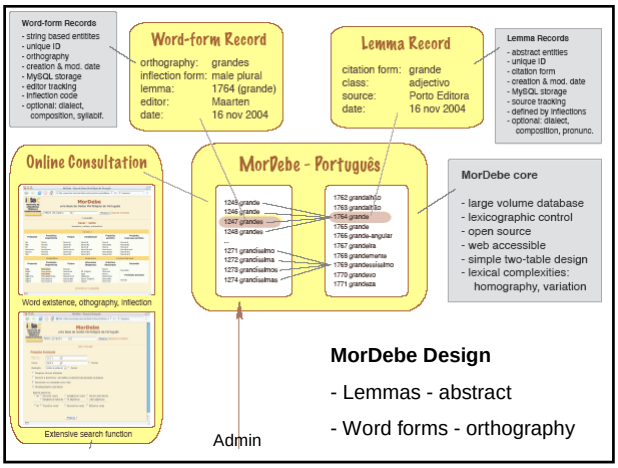
- All candidates are checked manually
  - Trained linguist
- Many words are non-neologistic
  - Will not be stored in the neologism database
- Pity to throw them away
  - They will reappear on the candidate list
  - They should be in the dictionary
  - Dictionary should be structured

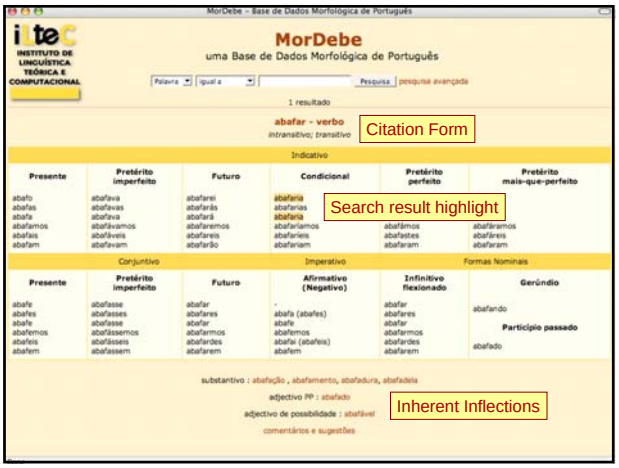
Part II : MorDebe

Morphological Database

Basic Design

- Web Based Database
  - MySQL - multi-table structure
  - Fully storage based - no rules
- Two tables
  - Lemmas
  - Word-forms
- All additions done via admin tools
  - Consistency
  - Semi-automatic Inflection



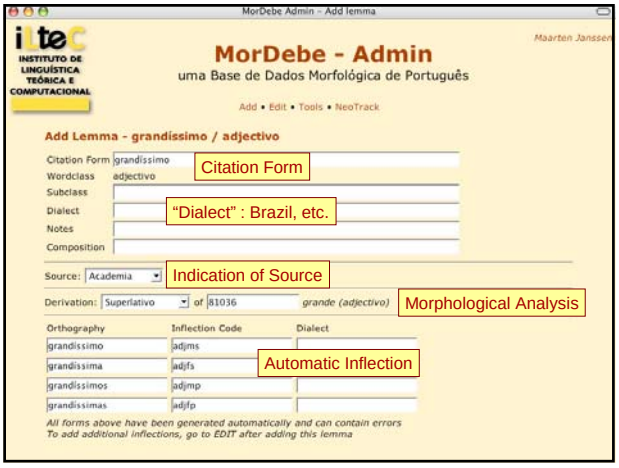


Need for Admin Tools

- MorDebe is a Distributed Database
  - Lemma and word forms in distinct databases
- Implicit relational database
  - No hard links - only references
  - *Walked* past tense of lemma nr 13.453
- Consistency important
  - No word-forms referring to non-existent lemmas
  - Adding and deleting by means of tools

Constantly Updated

- Integrated with NeoTrack
  - Daily additions
- Different project to include different resources
  - Manually adding words from dictionaries - Academia
  - Open - semi-automatically process CETEMPúblico
  - All projects with linguistic goals in mind
- New Additions
  - Foreseen: pronunciation
  - Subcategorisation grammar



Popular Resource

- Minimal Announcements
  - *Ciberduvidas da Lingua Portuguesa*
  - *Linguist List*
- Many visitors
  - 700 each day - and rising
  - Very positive reactions
- Linguistically relevant
  - Advanced search tools make it very useful
  - Phonetic patterns, minimal pairs, etc.

External Consistency Checks

- Admin tools for consistency
  - Errors are unavoidable - but fatal
- External checks
  - Word form without lemma, lemma without word forms
  - Incompatible inflection codes
  - Duplicate entries
  - Missing derivations
  - Citform = Pertinent form
  - Spaces in words (beginning and end)
- Extendible with more
  - Ad-hoc: Brazilian forms, orthographic variations

Regular Manual Check

- Weekly check of new additions
  - Ideally - check for errors
- Daily backups
  - No roll-back - extensive change log files
- Check by use
  - Multiple linguistic projects
- User feedback
  - Mostly irrelevant questions
  - Sometimes indications of errors
  - New feature: user words

Inherent Inflection

- Between inflection and derivation
  - Different entries with strong ties
- Modeled by "Lexical Functions"
  - Inflectional Functions
  - s0a, s0v, adv0, a2, able2, max, min, aum, dim, female
- Orthographic variation
  - alt, pe
  - doirar / dourar - 2 different paradigms
  - "same" word

Special Cases

- Defective Paradigms
  - Non-existing forms marked by "-" (nevo - I snow)
- Adundant Paradigms
  - Two entries for same lemma with same inflection code
  - Waaide / woei
- Homonymy
  - Not semantically driven
  - Formally driven
  - *redar - redou - re+dar / redo - rede+ar*

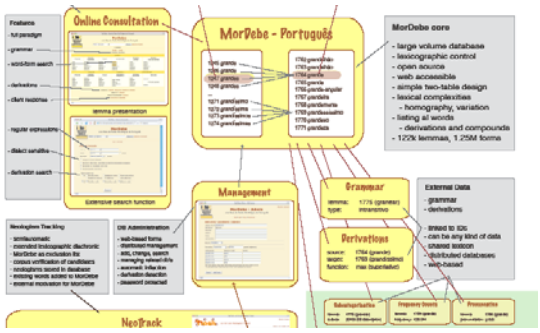
Part III : OSLIN

Open Source Lexical Information Network

Open Source - Modular Network

- All Lexical resources need a lemma list
  - MorDebe could provide that
  - Open source - useable for everybody
- MorDebe modular design
  - Word forms separate database
  - Grammar yet another
  - Extendible quasi indefinitely
- Web Based
  - Different parts can be maintained by different people
  - Cooperation rather than competition

Open Source Lexical Information Network



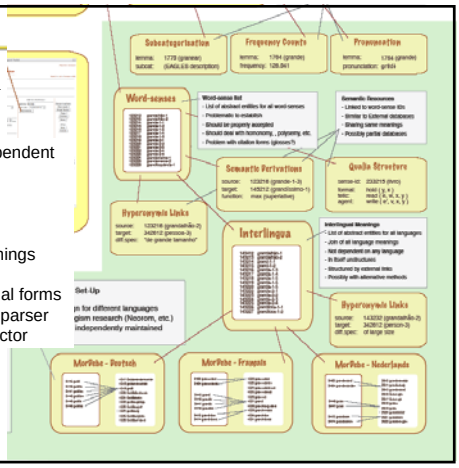
Multilingual

Language Independent

- Not Unicode
- No Chinese

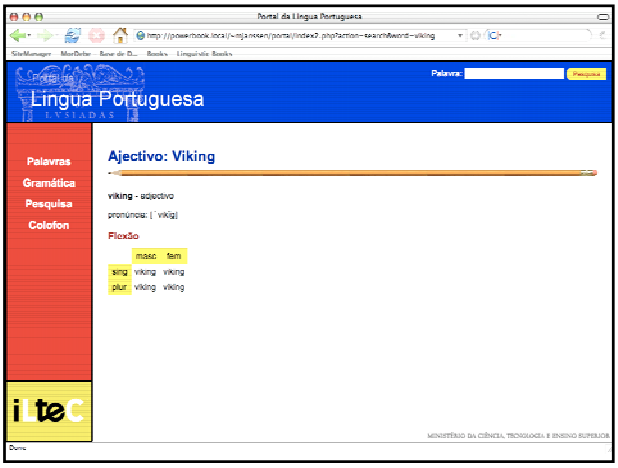
Some specific things

- list of inflectional forms
- morphological parser
- automatic inflector



Computationally Weak

- All programming done by me
  - Limited time
  - Little computational background - “hacking”
- No optimalisation
  - Complex search queries too slow
  - Regular pattern search disabled
- Multilingual use means more staff
  - Improve language independent tools
  - Data and tools are open source



Reusable Resources

- Hype Concept
  - Many old resources unusable by now
  - EAGLES, ISLE-MILE, etc.
  - Reusability by standardization
  - Build new sources upon old ones
- OSLIN different
  - Reusable by extension
  - The lexicon can be used live rather than once
  - Politically difficult - but conceptually better