Robust Morphologic Analyzer for highly inflected Languages

Andrés Tomás Hohendahl [1,3]

José Francisco Zelasco [1,2]

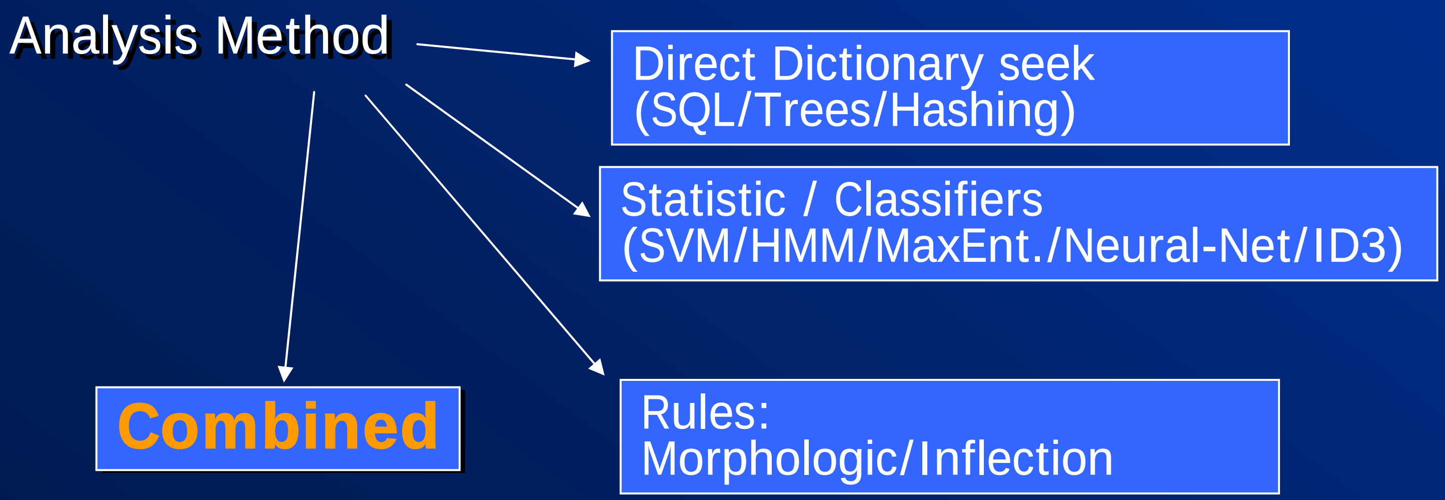
1 Lab. de Estereología y Mecánica Inteligente
Facultad de Ingeniería, U.B.A.

2 INTIA Facultad de Ciencias Exactas, UNCPBA

3 Instituto de Ingeniería BioMédica, U.B.A.

Languages & Dictionaries

- Inflected Languages
  - Slighltly Inflected:
    - English ~ 80k roots (x2.3) < 200k total
  - Highly Inflected + Parasynthetic
    - Spanish ~ 80k roots (x10k) > 3000M total
    - Huge Word-Space 1.0 E22 words for 15 letters
    - Similar for Polish, French, Italian, Portuguese, etc.



Analysis Method

Direct Dictionary seek (SQL/Trees/Hashing)

Statistic / Classifiers (SVM/HMM/MaxEnt./Neural-Net/ID3)

Rules: Morphologic/Inflection

Combined

Languages & Dictionaries

- Lexical Word analysis (goals)
  - Minimum stored amount of data
  - Obtain Semantic and Grammatical Information
  - Tolerate Misspelling & Suggest Corrections
  - Do all above: efficiently

- Used Method
  - Store word-Roots along with applicable Rules & base Tags
  - Each Morphologic Rule contains Grammatical & Semantic info.
  - Fast in-memory data structures: (Patricia-Trie y Tst)
  - Recursive & Greedy Algorithm: Seek / Lemmatize
  - Intelligent Spelling Suggestion: min. seek / max. probability

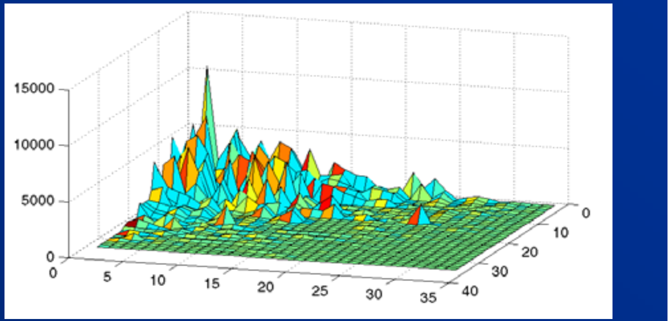
Languages & Dictionaries

- Language Recognition

Statistic

- Fast (few operations)
- Compact Datasets ~5kb/language
- Good Recall (F-Score >95%)
- Reduces Unnecessary Seeks
- State of the Art in 2004

• Hohendahl, A.T. Zelasco, J.F. WICC 2006 (art.694)



ES- Distribución de diletras
por: inicial, segunda (48424words)

Other Methods

Proprietary (MS, etc.)

Brute Force (high cost)

• Padró, Lluís/Munsa. TALP 2004 UPC

Efficient Index Structures

- SQL / TSQL / CQL (high level)
  - → Inefficient for partial words
  - → High cost (resources, licensing, maintenance, TOC)
- Binary Trees, M-Trees, Radix-Trees
  - → Less efficient for partial-matching

✓ Tries & TST (Ternary Search Trees)

- → Linear Time **O**(word length)
- → Useful for error detection/correction
- → Easy to finding Sub-Ranges for Similarity
- → Flexible: Linkable & Combinable

Reversible Morphologic Rules

- Spanish vocabulary using enhanced ASPELL compression

~ 3.900 inflection rules (300 prefix/infix + 3600 suffix)

~ 200 Semantic/Grammatical Attributes.

~ 79.000 Root words (Lemmas)

~ 300 kb compacted (*.zip)

ASPELL.org (GNU)

Yields ➔ >5 Million exact recognizable words

+ Phonetic Guess Sampa (Sound-Like)

+ Enhanced Spell Correction (statistic-guess)

+ Morphologic Guessing (statistic + rules)

+ Parasynthesis (multiple combinations)

➔ Huge >5000 Million word space !

(not including correctable mis-spells)

Morphologic Analysis Algorithm

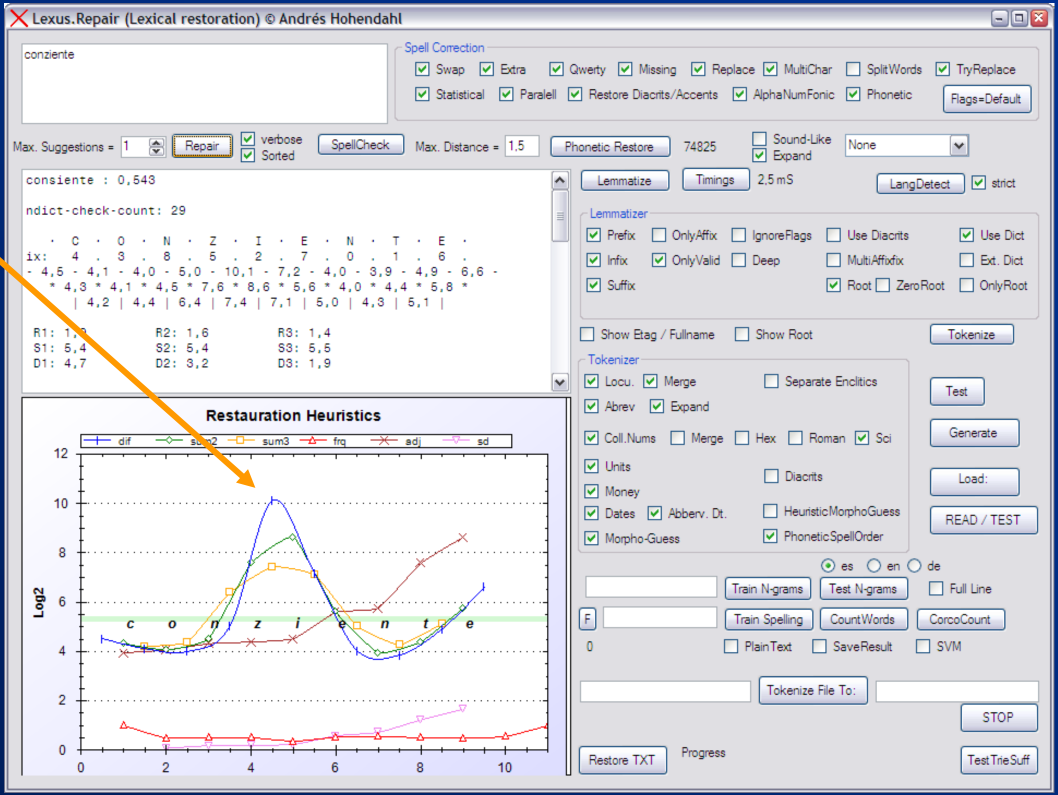
- Pseudo-Logic Diagram (very simplified) for Finding a Word

```
if (word in Roots) → found
Acumulate word in [seePrefix]
foreach Affix Rule in Suffix-Rules
    if Rule Applicable to word → strip-Suffix
        if (stripped in Roots  & bears Rule)
            → found
        else accumulate in [seePrefix]

foreach word in [seePrefix]
    foreach Affix Rule in Prefix-Rules
        if Rule Applicable to word → strip-Prefix
            if (stripped in Roots & bears Rule)
                → found
```

Spell Error Detection-Correction

- Using Bigram & Trigram Freq. from Language Detector.
- Heuristics to find best fitting replacement.
- Reduced seek count.
- Detects promptly unusual zones.
- Uses language specific rules.
- Usually finds the best “human” word in the first (few) trials
- Shares TST/Tries with Analyzer.
- Simply based on: **Poor-Man-Speller**



Lexus.Repair (Lexical restoration) © Andrés Hohendahl

conziente

Spell Correction
☑ Swap ☑ Extra ☑ Qwerty ☑ Missing ☑ Replace ☑ MultiChar □ SplitWords ☑ TryReplace
☑ Statistical ☑ Paralell ☑ Restore Diacrits/Accents ☑ AlphaNumFonic ☑ Phonetic

Max. Suggestions = 1 ☑ verbose ☑ Sorted Max. Distance = 1,5 74825 □ Sound-Like ☑ Expand None

```
consiente : 0,543

ndict-check-count: 29

  · C · O · N · Z · I · E · N · T · E ·
ix:  4 . 3 . 8 . 5 . 2 . 7 . 0 . 1 . 6 .
- 4,5 - 4,1 - 4,0 - 5,0 - 10,1 - 7,2 - 4,0 - 3,9 - 4,9 - 6,6 -
   * 4,3 * 4,1 * 4,5 * 7,6 * 8,6 * 5,6 * 4,0 * 4,4 * 5,8 *
     | 4,2 | 4,4 | 6,4 | 7,4 | 7,1 | 5,0 | 4,3 | 5,1 |

 R1: 1,[illegible]       R2: 1,6       R3: 1,4
 S1: 5,4       S2: 5,4       S3: 5,5
 D1: 4,7       D2: 3,2       D3: 1,9
```

2,5 mS ☑ strict

Lemmatizer
☑ Prefix □ OnlyAffix □ IgnoreFlags □ Use Diacrits ☑ Use Dict
☑ Infix ☑ OnlyValid □ Deep □ MultiAffixfix □ Ext. Dict
☑ Suffix ☑ Root □ ZeroRoot □ OnlyRoot

□ Show Etag / Fullname □ Show Root

Tokenizer
☑ Locu. ☑ Merge □ Separate Enclitics
☑ Abrev ☑ Expand
☑ Coll.Nums □ Merge □ Hex □ Roman ☑ Sci
☑ Units □ Diacrits
☑ Money
☑ Dates ☑ Abberv. Dt. □ HeuristicMorphoGuess
☑ Morpho-Guess ☑ PhoneticSpellOrder

● es ○ en ○ de
□ Full Line
0 □ PlainText □ SaveResult □ SVM

Progress

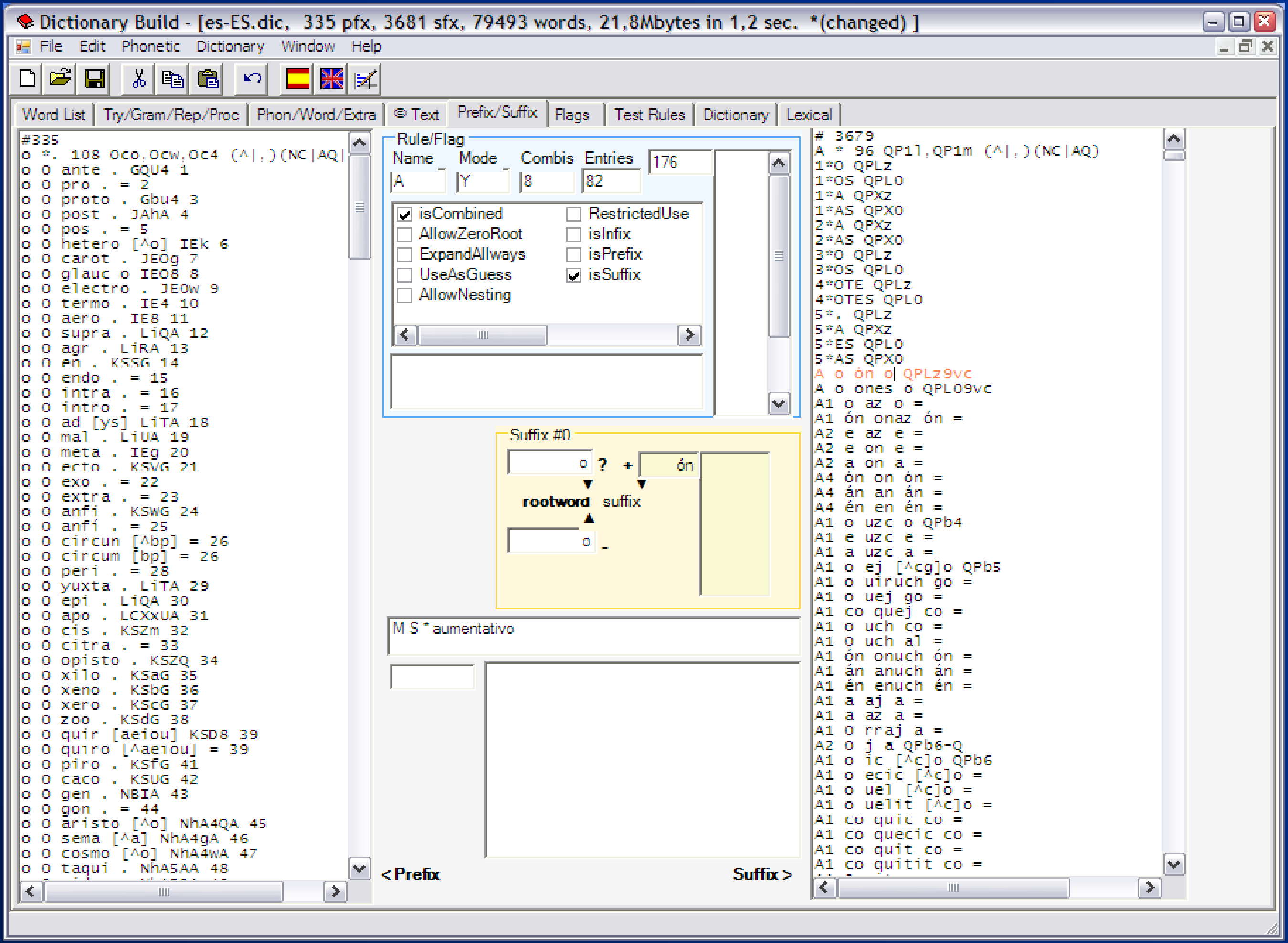
Restauration Heuristics

Dictionary + Rule Editor/Utility

Runs on Windows (.NET 2.0 C# Platform)

Features:

- Builds Rules
- Test Rules
- Analyzes words
- Expands words
- Benchmarks
- Imports
  - IFFIX
  - ASPELL
  - Word-List
- Does all kind of word/tag operations



Dictionary Build - [es-ES.dic, 335 pfx, 3681 sfx, 79493 words, 21,8Mbytes in 1,2 sec. *(changed)]

```
#335
o *. 108 Oco,Ocw,Oc4 (^|.)(NC|AQ|
o O ante . GQU4 1
o O pro . = 2
o O proto . Gbu4 3
o O post . JAhA 4
o O pos . = 5
o O hetero [^o] IEk 6
o O carot . JEOg 7
o O glauc o IEO8 8
o O electro . JEOw 9
o O termo . IE4 10
o O aero . IE8 11
o O supra . LiQA 12
o O agr . LiRA 13
o O en . KSSG 14
o O endo . = 15
o O intra . = 16
o O intro . = 17
o O ad [ys] LiTA 18
o O mal . LiUA 19
o O meta . IEg 20
o O ecto . KSVG 21
o O exo . = 22
o O extra . = 23
o O anfi . KSWG 24
o O anfi . = 25
o O circun [^bp] = 26
o O circum [bp] = 26
o O peri . = 28
o O yuxta . LiTA 29
o O epi . LiQA 30
o O apo . LCXxUA 31
o O cis . KSZm 32
o O citra . = 33
o O opisto . KSZQ 34
o O xilo . KSaG 35
o O xeno . KSbG 36
o O xero . KScG 37
o O zoo . KSdG 38
o O quir [aeiou] KSD8 39
o O quiro [^aeiou] = 39
o O piro . KSfG 41
o O caco . KSUG 42
o O gen . NBIA 43
o O gon . = 44
o O aristo [^o] NhA4QA 45
o O sema [^a] NhA4gA 46
o O cosmo [^o] NhA4wA 47
o O taqui . NhA5AA 48
```

Rule/Flag

Name	Mode	Combis	Entries	
A	Y	8	82	176

- ☑ isCombined
- ☐ AllowZeroRoot
- ☐ ExpandAllways
- ☐ UseAsGuess
- ☐ AllowNesting
- ☐ RestrictedUse
- ☐ isInfix
- ☐ isPrefix
- ☑ isSuffix

Suffix #0

o ? + ón

rootword suffix

o -

M S * aumentativo

```
# 3679
A * 96 QP1l,QP1m (^|.)(NC|AQ)
1*O QPLz
1*OS QPLO
1*A QPXz
1*AS QPXO
2*A QPXz
2*AS QPXO
3*O QPLz
3*OS QPLO
4*OTE QPLz
4*OTES QPLO
5*. QPLz
5*A QPXz
5*ES QPLO
5*AS QPXO
A o ón o QPLz9vc
A o ones o QPLO9vc
A1 o az o =
A1 ón onaz ón =
A2 e az e =
A2 e on e =
A2 a on a =
A4 ón on ón =
A4 án an án =
A4 én en én =
A1 o uzc o QPb4
A1 e uzc e =
A1 a uzc a =
A1 o ej [^cg]o QPb5
A1 o uiruch go =
A1 o uej go =
A1 co quej co =
A1 o uch co =
A1 O uch al =
A1 ón onuch ón =
A1 án anuch án =
A1 én enuch én =
A1 a aj a =
A1 a az a =
A1 O rraj a =
A2 O j a QPb6-Q
A1 o ic [^c]o QPb6
A1 o ecic [^c]o =
A1 o uel [^c]o =
A1 o uelit [^c]o =
A1 co quic co =
A1 co quecic co =
A1 co quit co =
A1 co quitit co =
```

Phonetic Similarity Module

New Algorithm

- Measures what humans “think-sounds” written text.
- Based on analogical phono-articulartory model (uses non-linear kernel on: pitch-vibration, nasal-lingual position/occlusion + openness + fricative energy)
- Measures Distance among words with a Real Number 0..max where (1.0 is the mean phonemic distance)
- Correlation with human-perception over 85%.
- Establishes a good parameter for spell correction delivering the correct word even with worse misspells,

Example:

VAHIEMA → BALLENA

(only 1 guess d~0.69 0.001sec)

A. Hohendahl,
S. Zanutto,
A.Wainselboim
SLAN 2007

Phonetic Similarity Algorithm

Highlights

- Very little literature found on the subject.
- Outperforms classic Lexical distance for cognitive perception experiments and measures (Levenshtein, etc.)
- Very Fast (over 30k/second)
- Small Memory usage

Windows Utility for Testing

Benchmarking

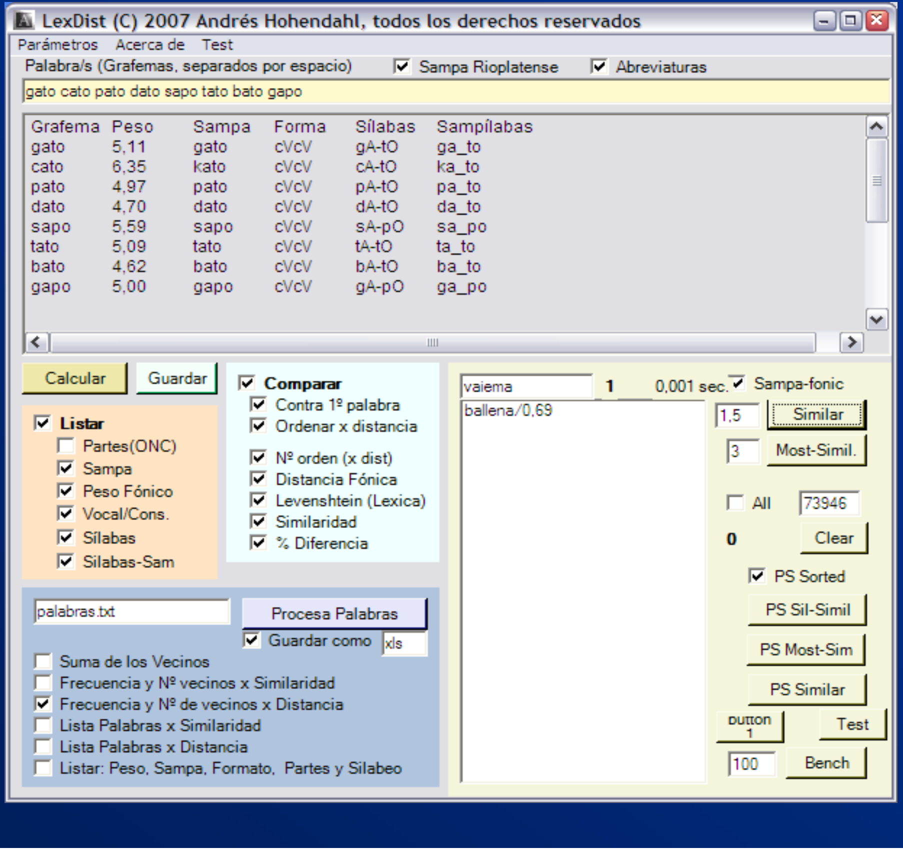
- Fast-Find
- Best-Fit

Processing

- Word Lists

Making

- Word Matrix



Palabra/s (Grafemas, separados por espacio) ☑ Sampa Rioplatense ☑ Abreviaturas

gato cato pato dato sapo tato bato gapo

Grafema	Peso	Sampa	Forma	Sílabas	Sampílabas
gato	5,11	gato	cVcV	gA-tO	ga_to
cato	6,35	kato	cVcV	cA-tO	ka_to
pato	4,97	pato	cVcV	pA-tO	pa_to
dato	4,70	dato	cVcV	dA-tO	da_to
sapo	5,59	sapo	cVcV	sA-pO	sa_po
tato	5,09	tato	cVcV	tA-tO	ta_to
bato	4,62	bato	cVcV	bA-tO	ba_to
gapo	5,00	gapo	cVcV	gA-pO	ga_po

☑ **Listar**
- ☐ Partes(ONC)
- ☑ Sampa
- ☑ Peso Fónico
- ☑ Vocal/Cons.
- ☑ Sílabas
- ☑ Silabas-Sam

☑ **Comparar**
- ☑ Contra 1º palabra
- ☑ Ordenar x distancia
- ☑ Nº orden (x dist)
- ☑ Distancia Fónica
- ☑ Levenshtein (Lexica)
- ☑ Similaridad
- ☑ % Diferencia

palabras.txt

☑ Guardar como xls

- ☐ Suma de los Vecinos
- ☐ Frecuencia y Nº vecinos x Similaridad
- ☑ Frecuencia y Nº de vecinos x Distancia
- ☐ Lista Palabras x Similaridad
- ☐ Lista Palabras x Distancia
- ☐ Listar: Peso, Sampa, Formato, Partes y Silabeo

vaiema **1** 0,001 sec. ☑ Sampa-fonic

ballena/0,69

1,5

3

☐ All 73946

0

☑ PS Sorted

100

Features & Applications

Features

- Find best-human like guess on mistyped or bad orthographically written (but sounding like) text.
- Delivering EAGLES 2.0 compatible, semantic-extended tags
- Uses Open-source Dictionaries, and spell checkers so it's adaptable to many languages based on free existing data.
- Delivers N-alternative Tags, ordered by phonetic distance.
- Detects foreign words (tagging language) + capable to handle many mixed languages (one must be principal)

Real World Human Computer Interface (HCI)

- Fast and Lightweight, engineered to fit into small appliances.
- Recognition + Guessing of parasynthetic O.O.V. (Out Of Vocabulary) in Scientific Text, Medical Records, etc.
- Robust Open-Lexicon Dialog System (free text)
- Automatic Speech Recognition (ASR) with huge-open Lexicon
- Teaching Aid / Support (Intelligent conversational agents)
- Artificial Understanding, AI, Context-based tagging, etc.

Future Research Lines

H.C.I.

- Cognitive Modeling for fast Storage-Retrieval
- Spanish Dialog Subsystem
  - Robust GLR Compiler (Tomita-Like w/Scrödinger Tokens)
  - Cognitive Run-Time
    - Implied verbal Logic (Math, Set & Boolean Logic)
    - Simple Scientific Math (numeric + algebraic)
    - Scientific Units Cognitive Operations
    - Artificial Shallow Understanding
  - Information Extraction on OOV. & mistyped words (morphologically correctly constructed, even with errors)
  - Conversational-Space Resolution (Me-You-They)
  - On-The Fly Anaphora Resolution & used as context
  - Ontology Driven Contextual Conditional Parsing

Questions?

Thank you

(dissertant)

José Francisco Zelasco

jfz@fi.uba.ar

Andrés T. Hohendahl

andres.hohendahl@fi.uba.ar