Portuguese Linguistic Tools: What, Why and How

Valeria de Paiva, Nuance Communications,

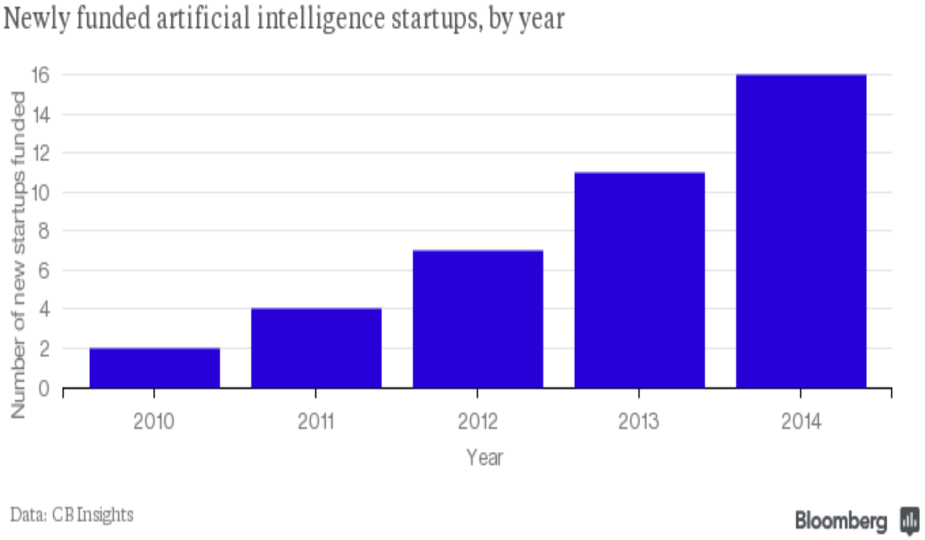
NL and AI Lab, Sunnyvale, CA, USA, Sept 2015

Talk at IBM Research, Brazil

The Future is Meaning

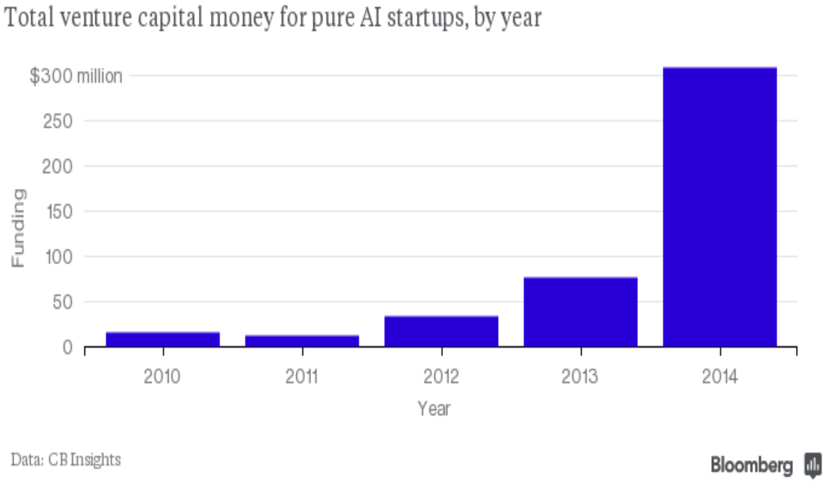
http://www.wired.com/2013/03/conversational-user-interface/

Setting the Scene: AI as an ecosystem



Newly funded artificial intelligence startups, by year

Data: CB Insights



Total venture capital money for pure AI startups, by year

Data: CB Insights

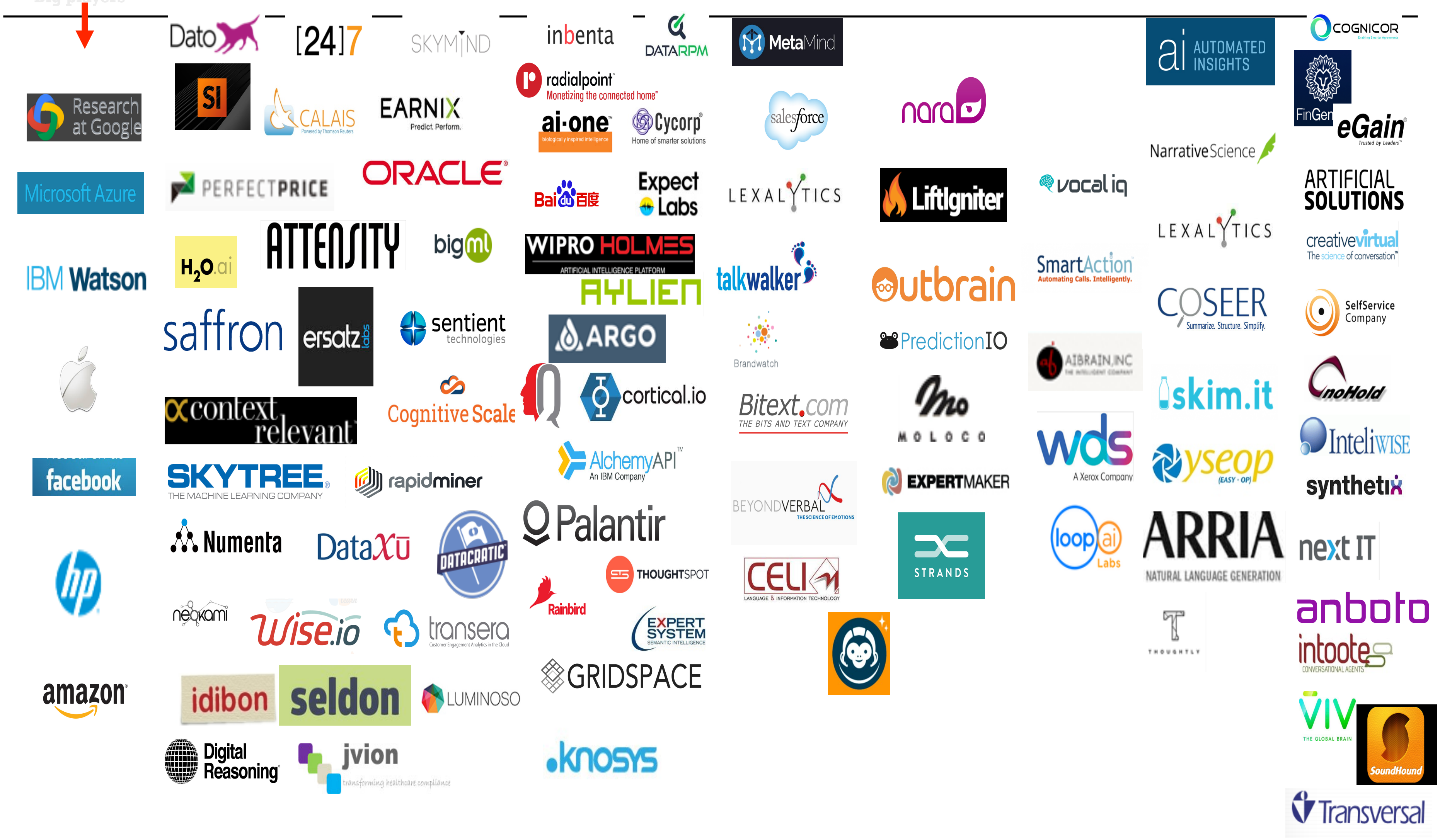
Source: http://www.bloomberg.com/news/articles/2015-02-03/i-ll-be-back-the-return-of-artificial-intelligence

Hundreds of startups & many capabilities

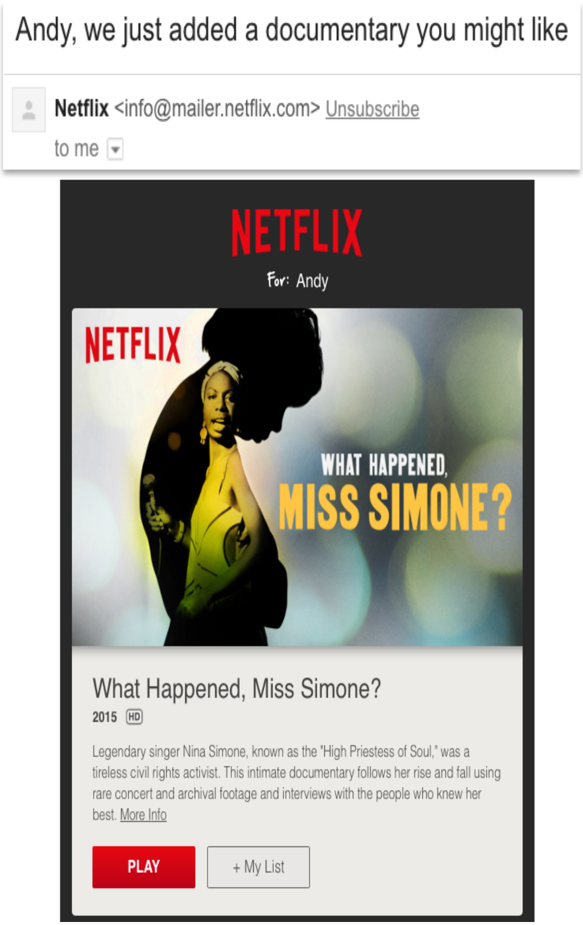
Functional Areas

Prediction | Knowledge Retrieval | Sentiment | Personalization | Learning | Summarization | VA

Big players



Intelligence is expected



Andy, we just added a documentary you might like

Netflix <info@mailer.netflix.com> Unsubscribe

to me

NETFLIX

For: Andy

What Happened, Miss Simone?

2015 HD

Legendary singer Nina Simone, known as the "High Priestess of Soul," was a tireless civil rights activist. This intimate documentary follows her rise and fall using rare concert and archival footage and interviews with the people who knew her best. More Info

PLAY + My List

Personalized & proactive.

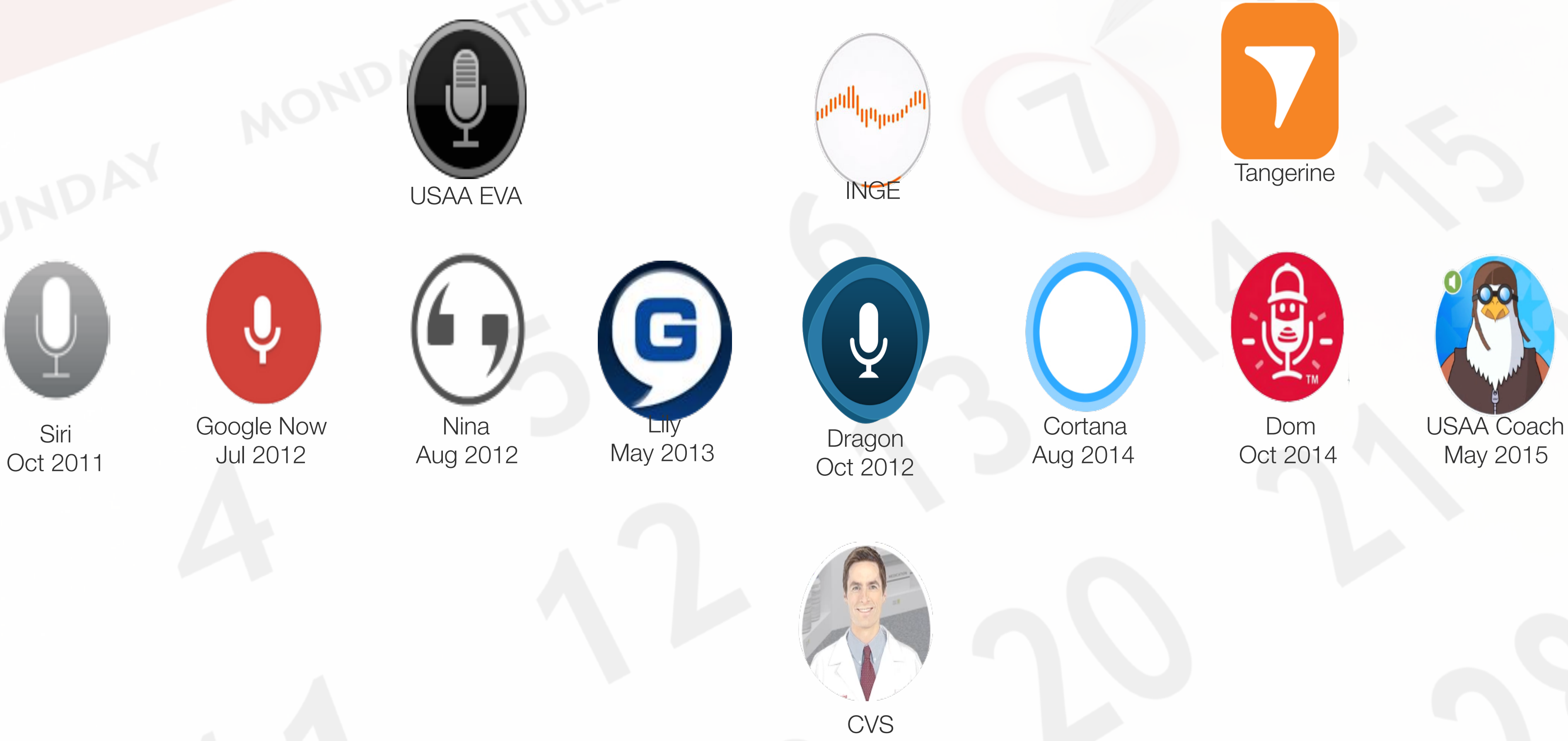
Hello, Miranda.
What can I do for you?
I've gathered a look at the day for you.
Here's the latest in financial news.

Consistent assistance on every channel.

Hey Siri, is the weather nice in Lake Tahoe today?

Knows how to talk.

Voice based Virtual Assistants paving the way for general acceptance of AI



USAA EVA

INGE

Tangerine

Siri
Oct 2011

Google Now
Jul 2012

Nina
Aug 2012

Lily
May 2013

Dragon
Oct 2012

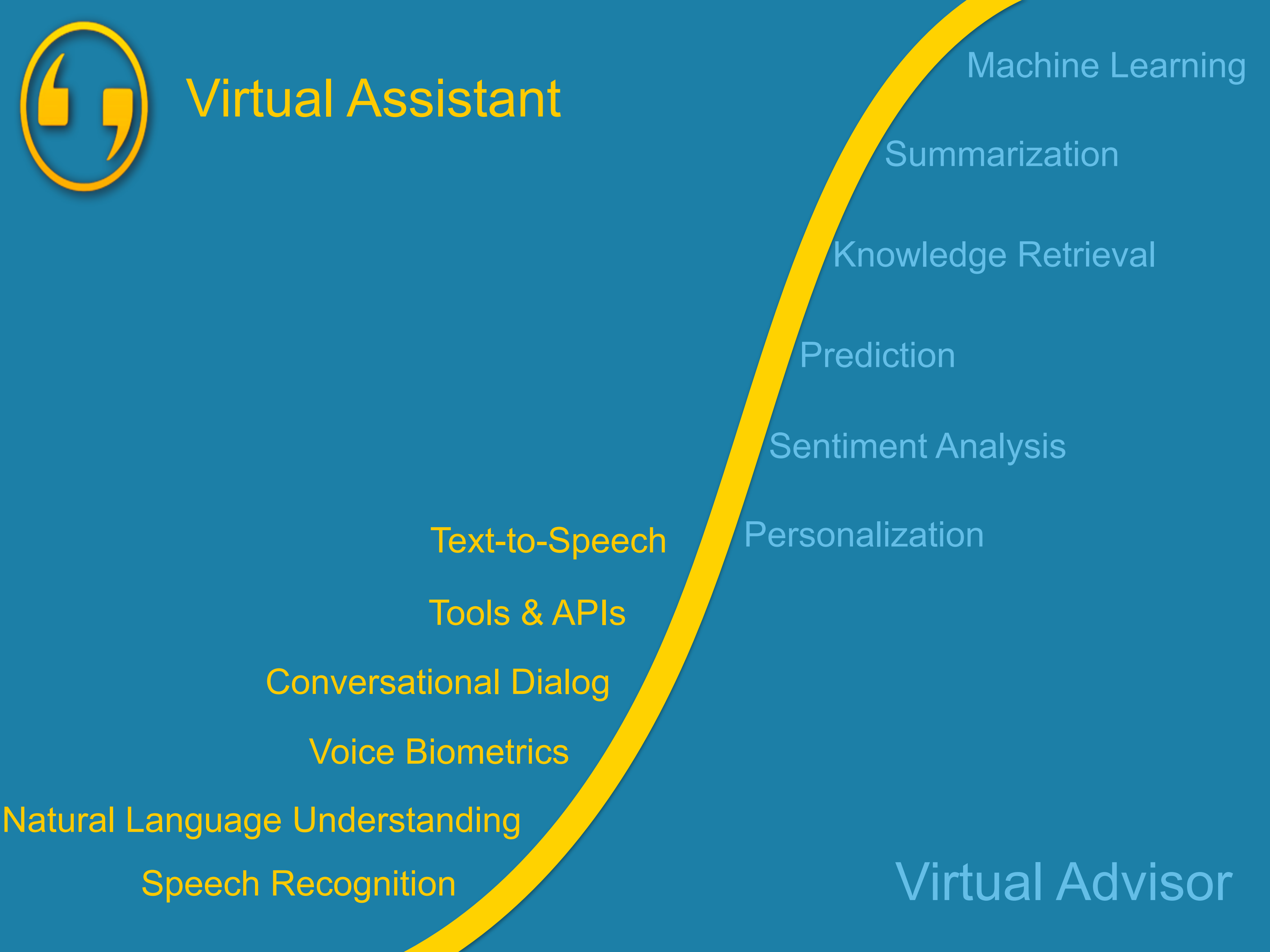
Cortana
Aug 2014

Dom
Oct 2014

USAA Coach
May 2015

CVS

Virtual Assistant



Machine Learning

Summarization

Knowledge Retrieval

Prediction

Sentiment Analysis

Personalization

Text-to-Speech

Tools & APIs

Conversational Dialog

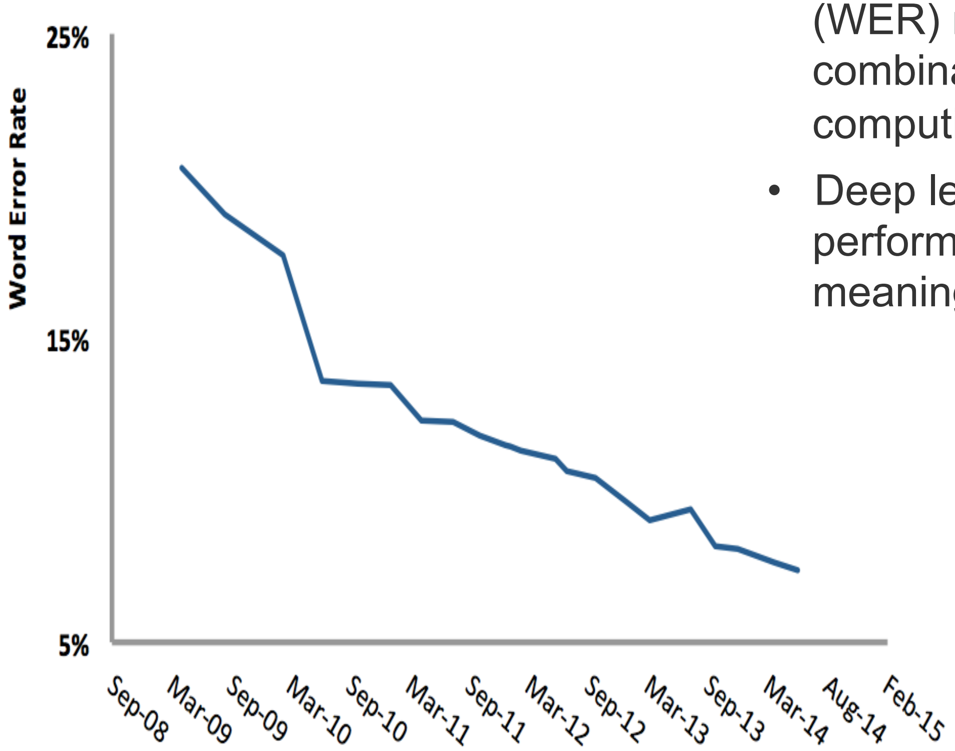
Voice Biometrics

Natural Language Understanding

Speech Recognition

Virtual Advisor

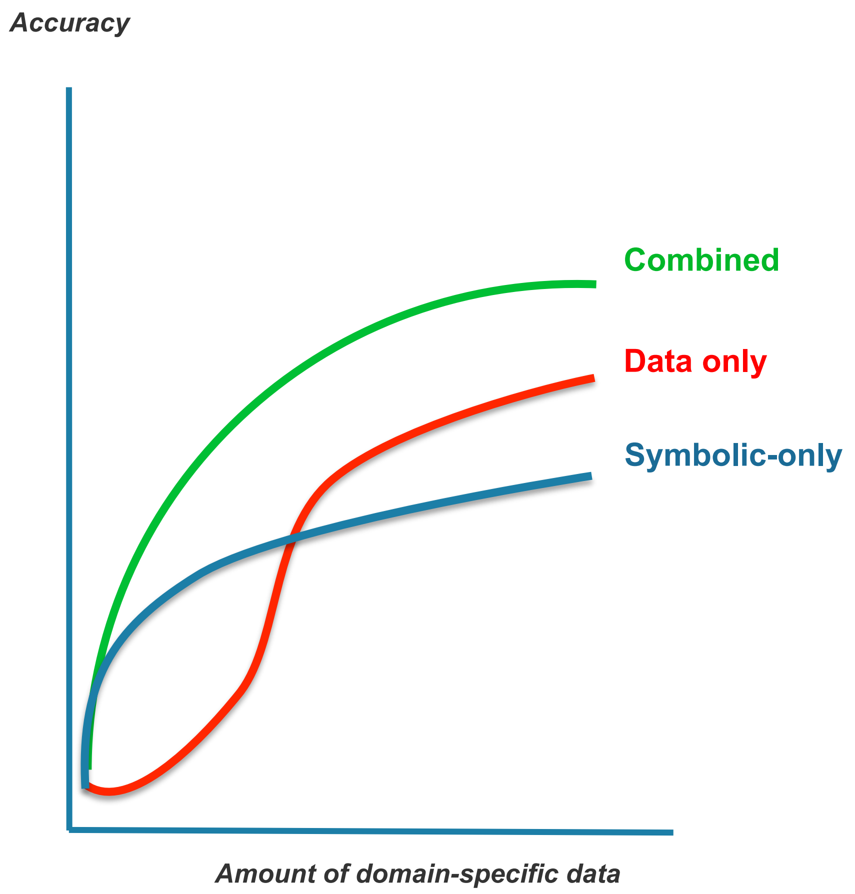
Core ASR performance



- Continuous server ASR word error rate (WER) reduction ~18% / year: combination of algorithms, data, and computing
- Deep learning (DNNs) is driving recent performance improvements in ASR and meaning extraction

Deep Language Understanding

Symbolic methods complement machine learning in a common architecture

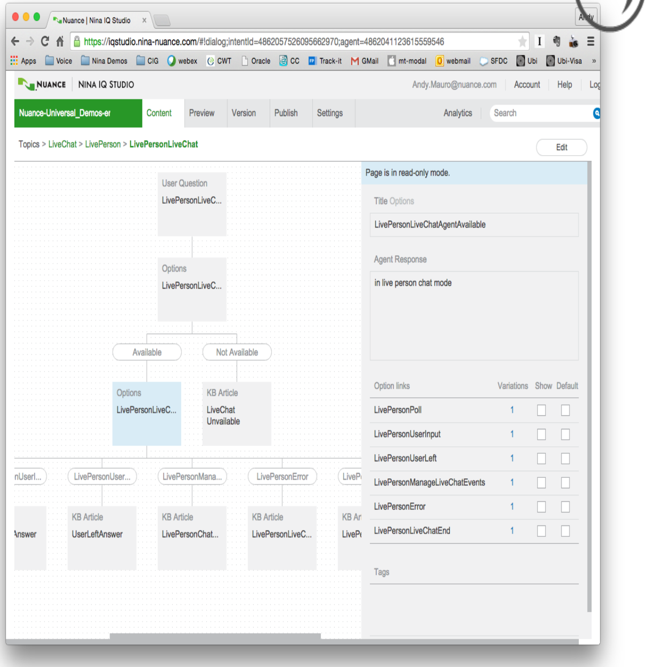


How Virtual Assistants are built today…

Structured Data

VA Building Tools & Runtime (NLU + Dialog + Knowledge)

SDK/API



Topics > LiveChat > LivePerson > LivePersonLiveChat

Page is in read-only mode.

Title Options

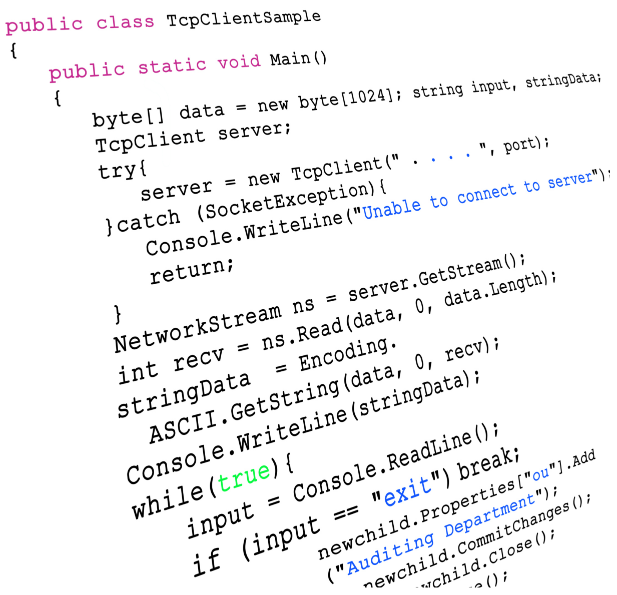
LivePersonLiveChatAgentAvailable

Agent Response

in live person chat mode

Option links	Variations	Show	Default
LivePersonPoll	1	□	□
LivePersonUserInput	1	□	□
LivePersonUserLeft	1	□	□
LivePersonManageLiveChatEvents	1	□	□
LivePersonError	1	□	□
LivePersonLiveChatEnd	1	□	□

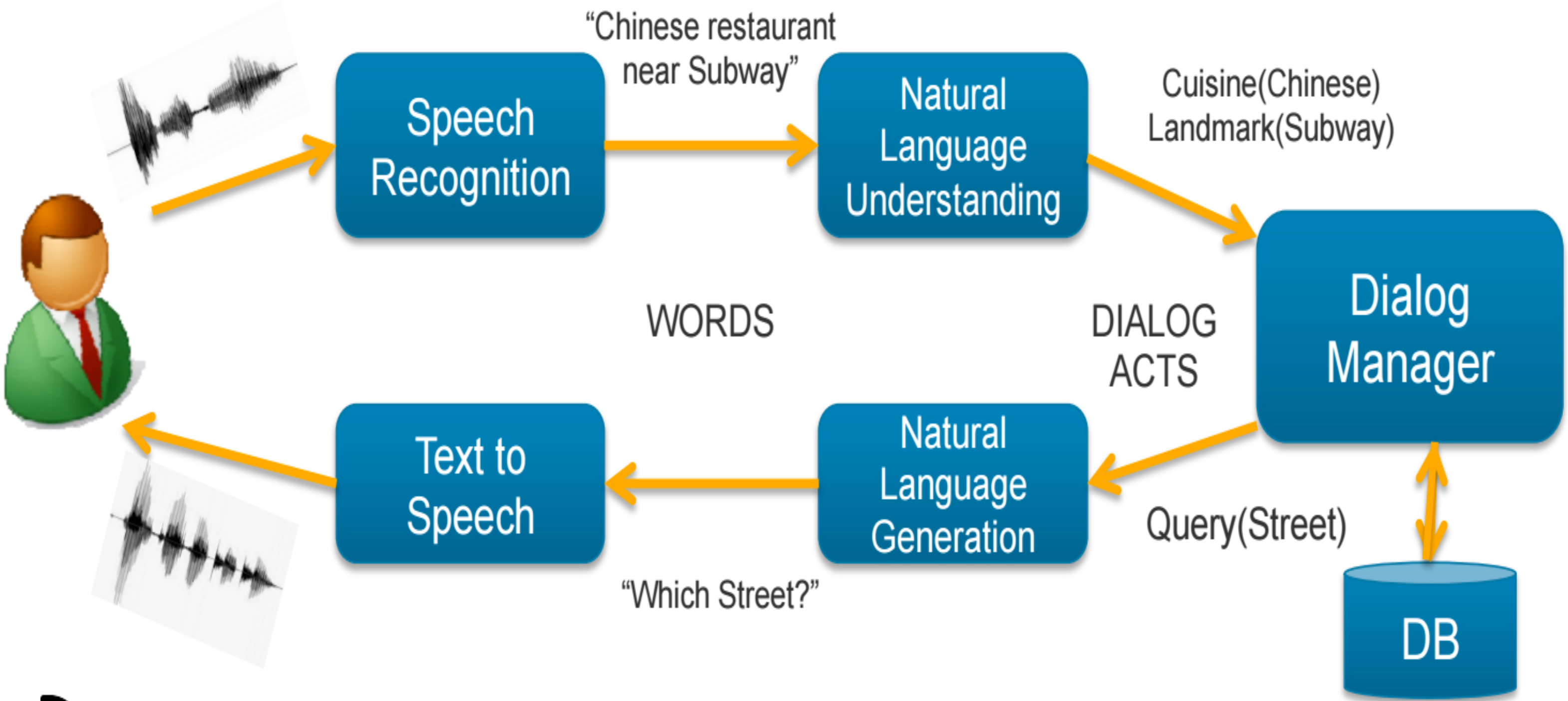
Tags



```
public class TcpClientSample
{
    public static void Main()
    {
        byte[] data = new byte[1024]; string input, stringData;
        TcpClient server;
        try{
            server = new TcpClient(" . . . . ", port);
        }catch (SocketException){
            Console.WriteLine("Unable to connect to server");
            return;
        }
        NetworkStream ns = server.GetStream();
        int recv = ns.Read(data, 0, data.Length);
        stringData  = Encoding.
            ASCII.GetString(data, 0, recv);
        Console.WriteLine(stringData);
        while(true){
            input = Console.ReadLine();
            if (input == "exit") break;
                newchild.Properties["ou"].Add
                ("Auditing Department");
                newchild.CommitChanges();
                wchild.Close();
                e();
```

Spoken Dialog Systems

Standard architecture of a Spoken Dialog System



Speech Recognition

"Chinese restaurant near Subway"

Natural Language Understanding

Cuisine(Chinese)
Landmark(Subway)

Dialog Manager

DB

Query(Street)

Natural Language Generation

"Which Street?"

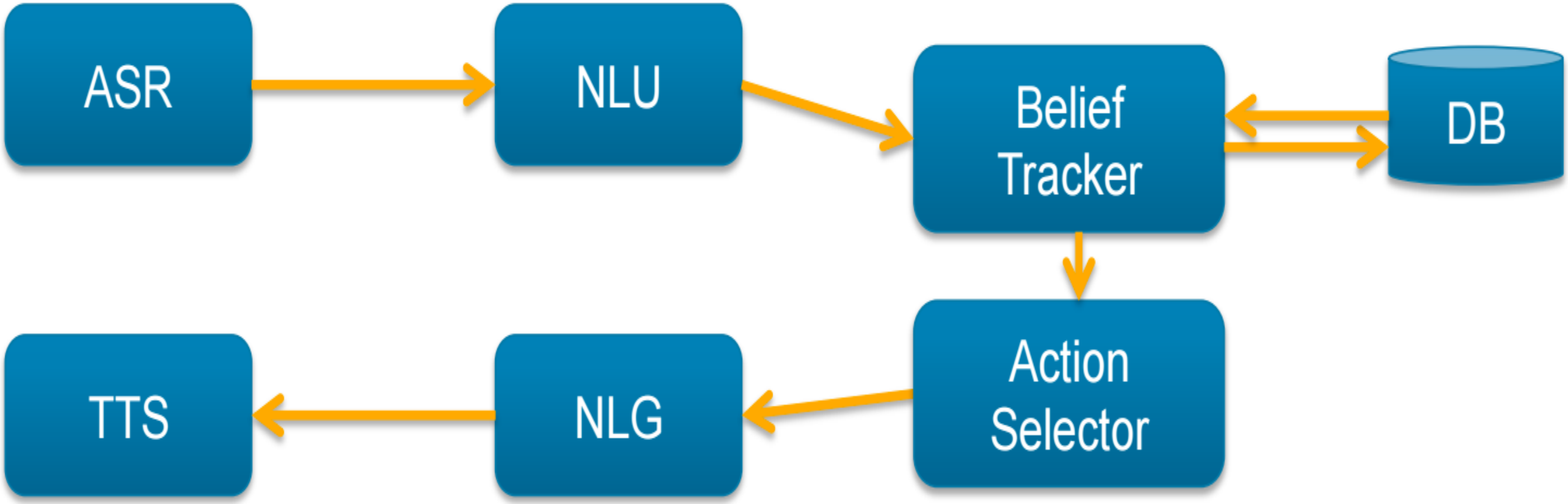
Text to Speech

WORDS

DIALOG ACTS

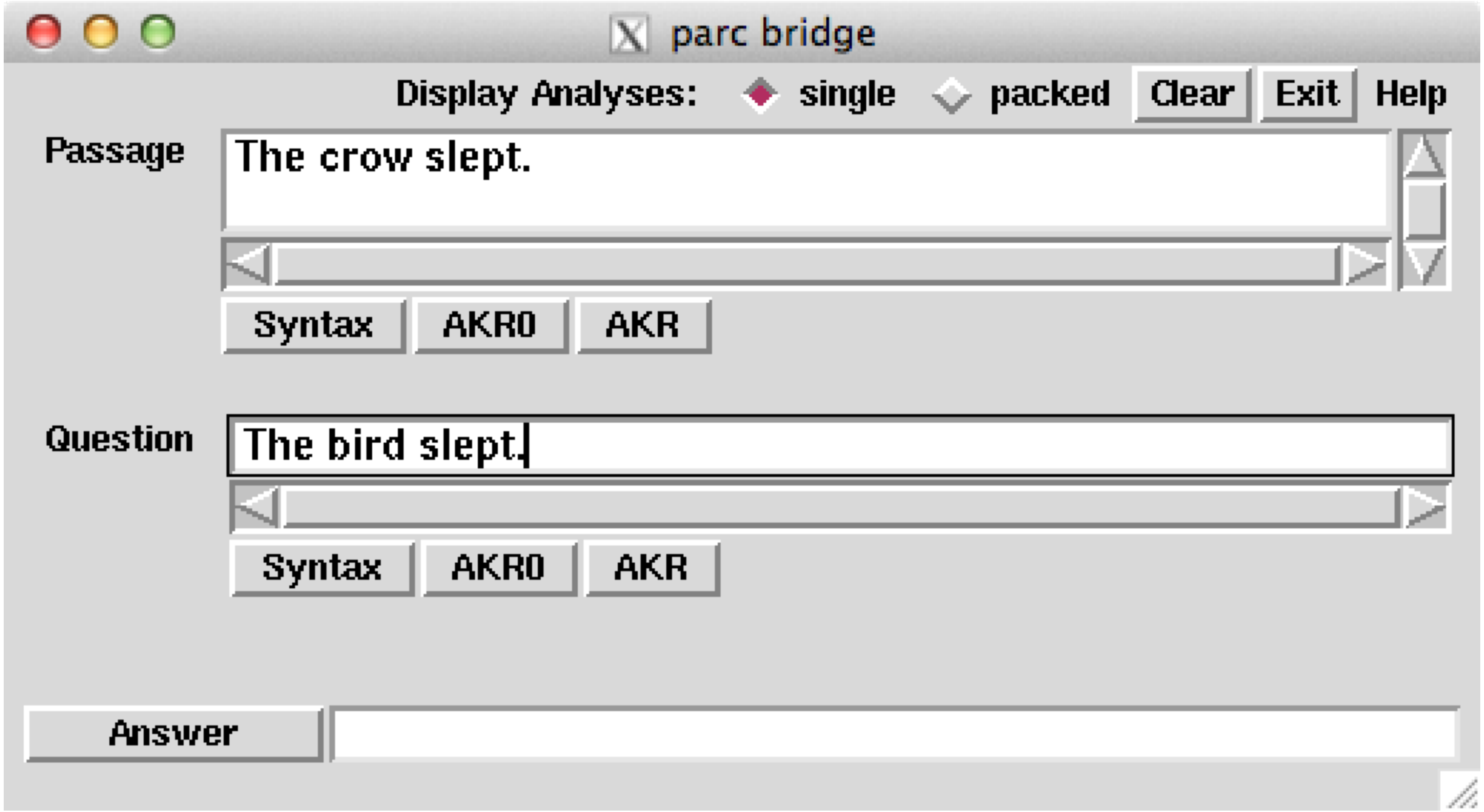
Spoken Dialog Systems

We propose:



Personal assistants in Portuguese

The Past: PARC's Bridge System (1999-2008)

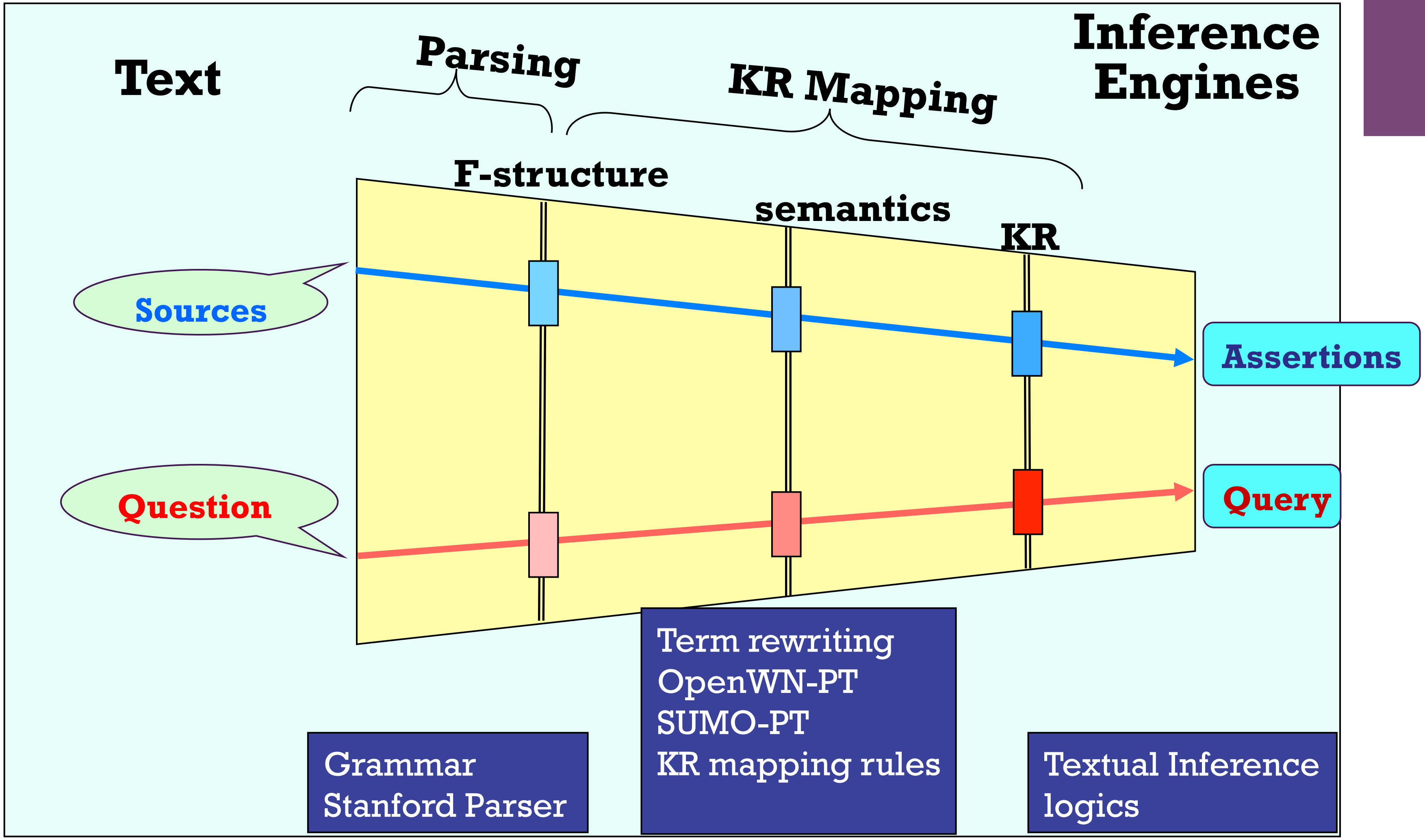


Display Analyses: ● single ○ packed

Passage: The crow slept.

Question: The bird slept.

+ Bridge System (1999-2008)

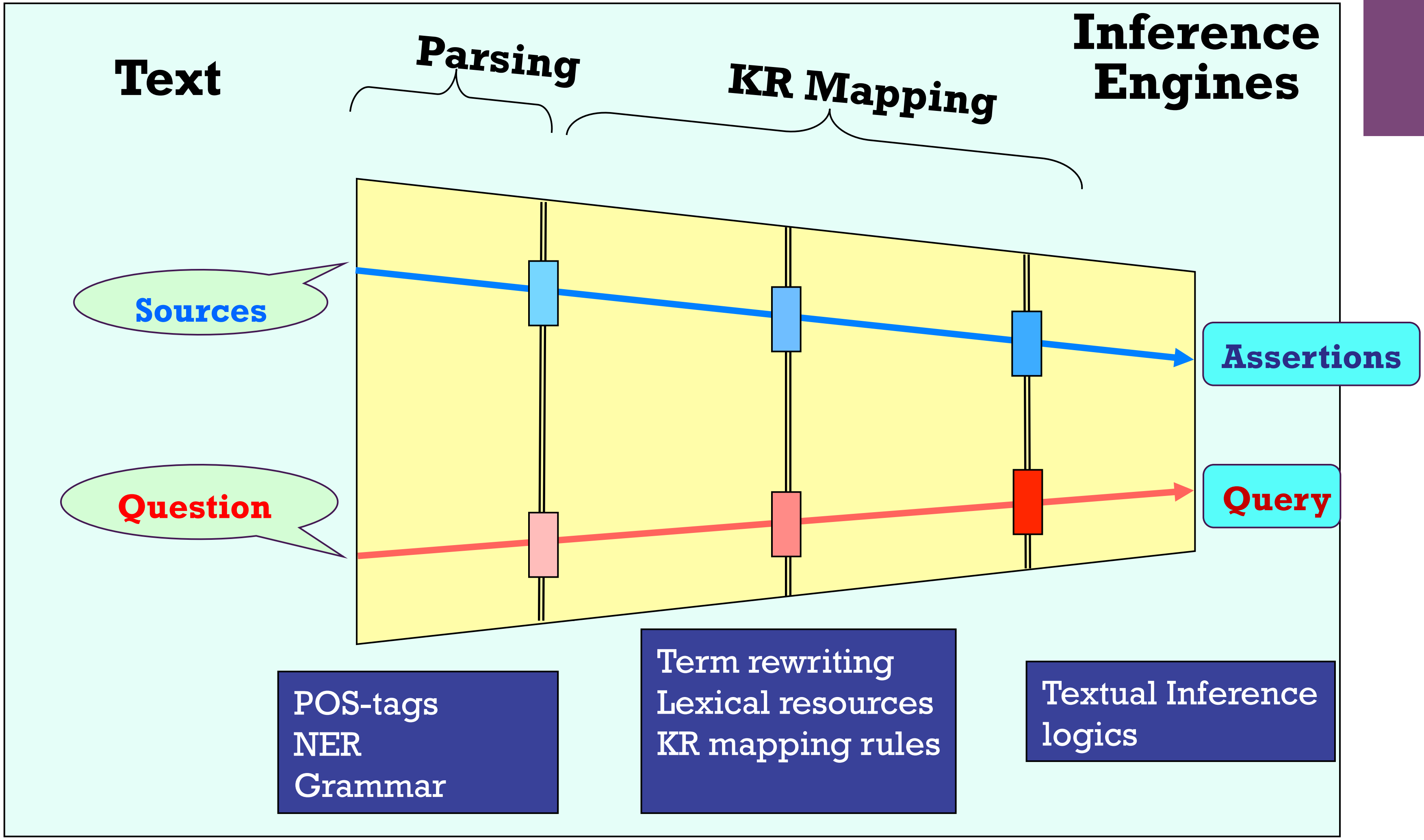


Term rewriting
OpenWN-PT
SUMO-PT
KR mapping rules

Grammar
Stanford Parser

Textual Inference
logics

New Bridge System



POS-tags
NER
Grammar

Term rewriting
Lexical resources
KR mapping rules

Textual Inference
logics

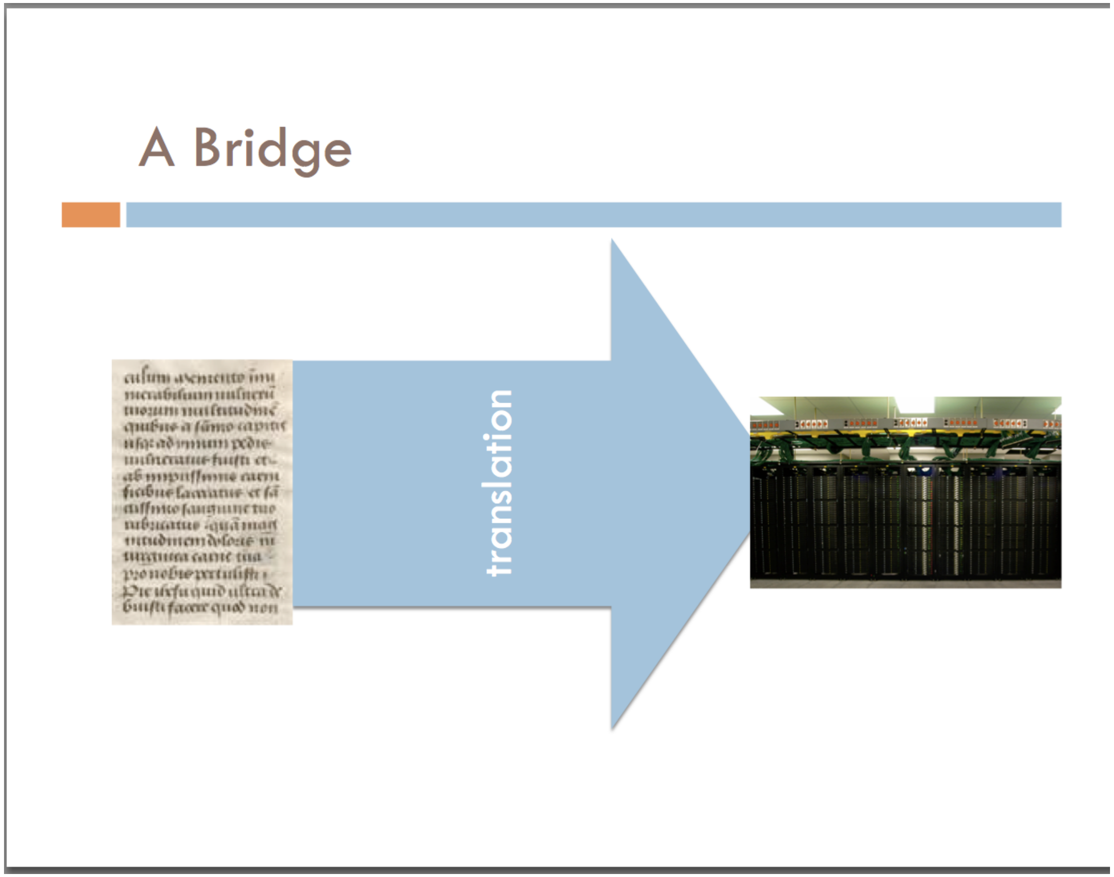
Redoing PARC work in Portuguese

PROCESSAMENTO DE LINGUAGEM NATURAL: UMA PERSPECTIVA PESSOAL WINGING NLP…

Valeria de Paiva, Cuil, Inc.

FGV Setembro de 2010

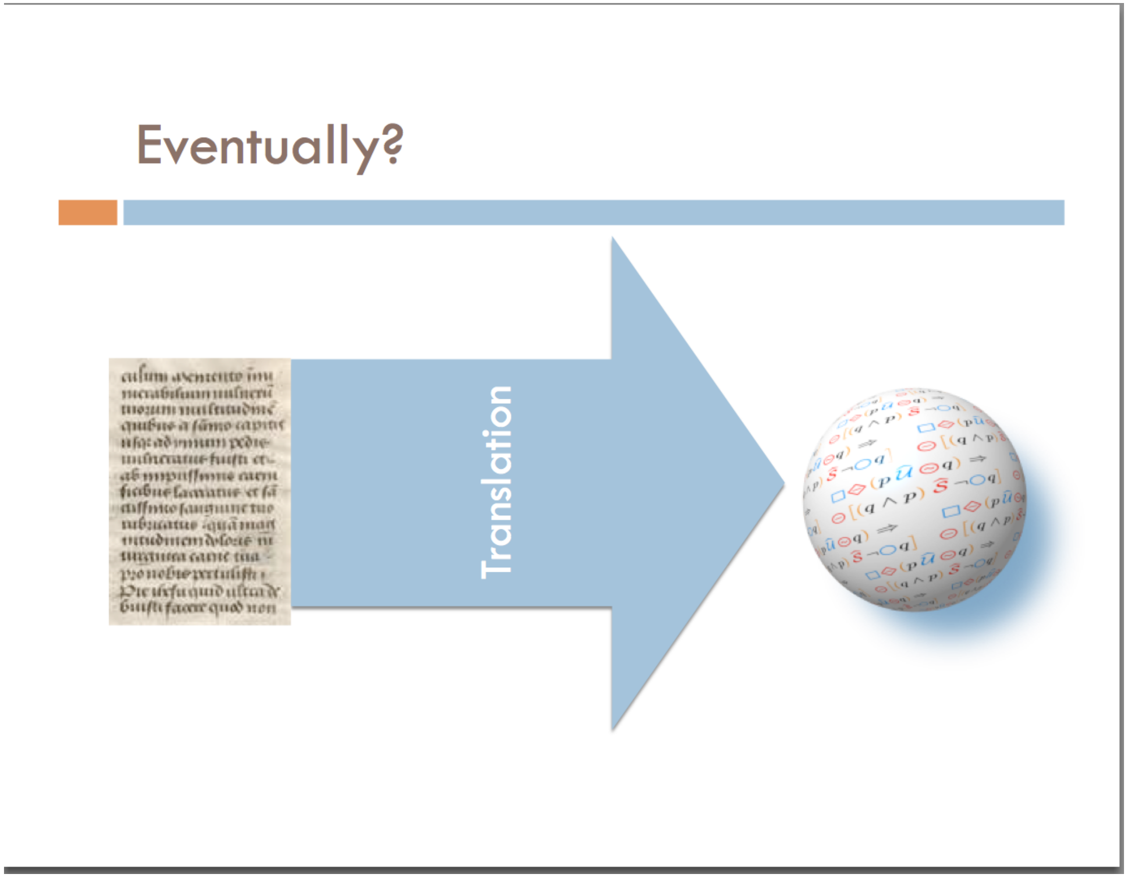
Goals in 2010



A Bridge

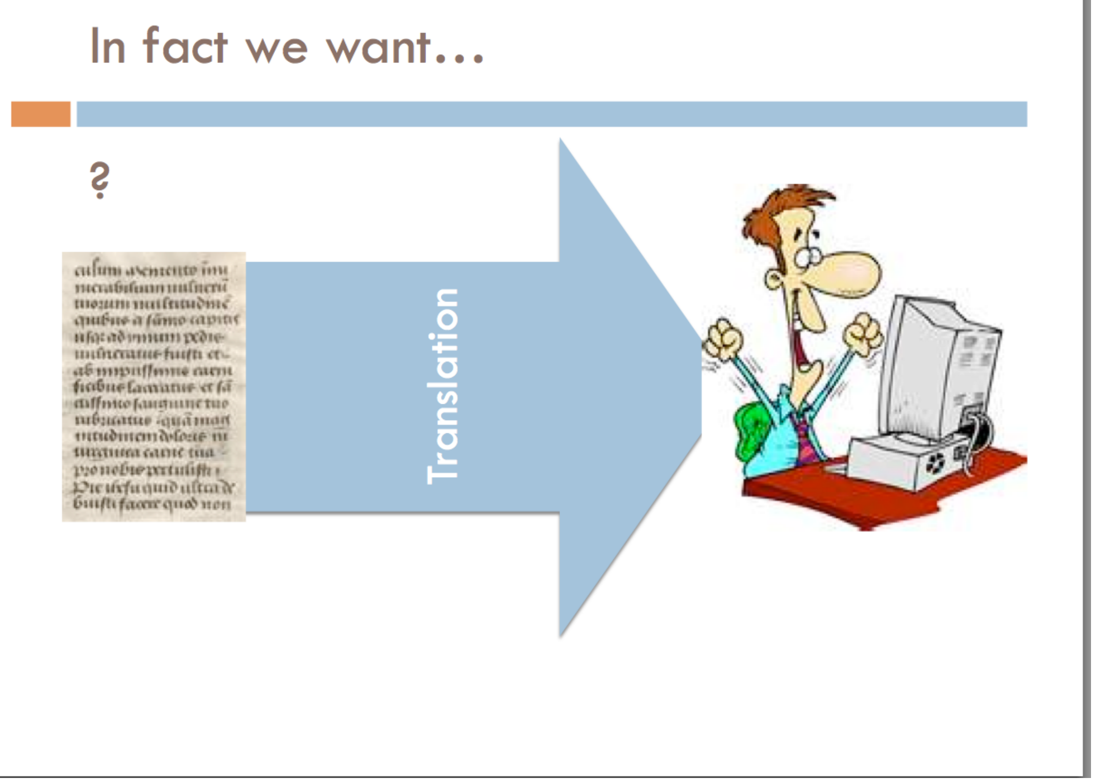
translation

Goals in 2010



Eventually?

Goals in 2010…



In fact we want…

?

□ Content analysis /large-scale intelligent information extraction, access and retrieval

□ Text understanding

□ Text generation

□ Text simplification

□ Automatic summarization

□ Dialogue systems

□ Question answering

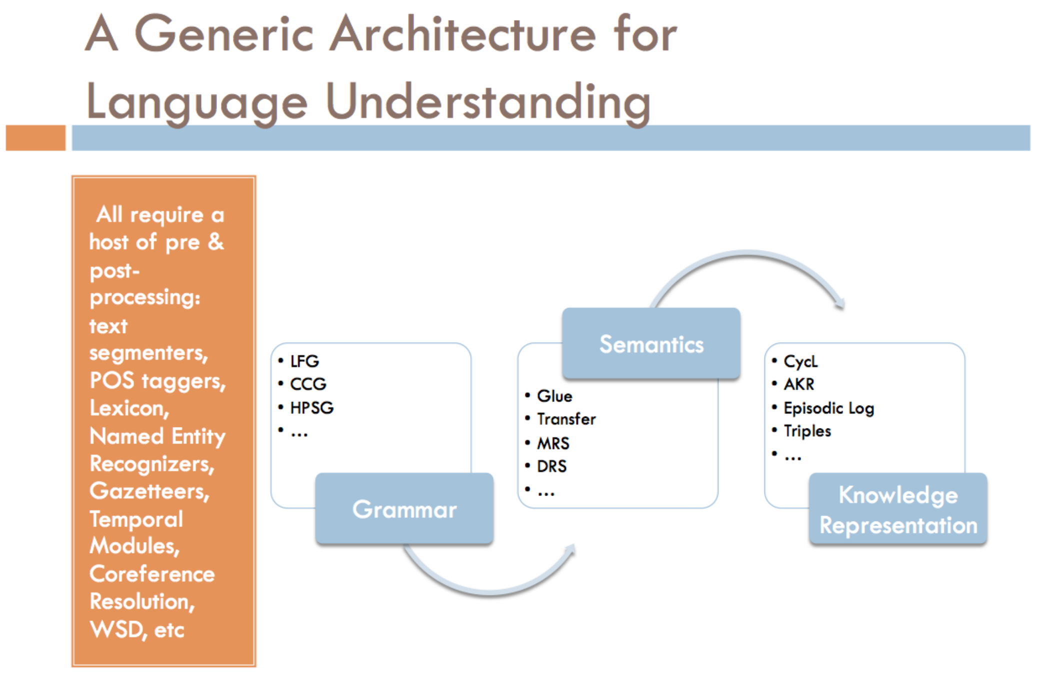
□ Machine Translation

□ Named Entity Recognition,

□ Anaphora/co-reference resolution,

□ Reading, writing, grammar aids, etc…

Pipeline envisaged in 2010 Plug and Play…



A Generic Architecture for Language Understanding

All require a host of pre & post-processing: text segmenters, POS taggers, Lexicon, Named Entity Recognizers, Gazetteers, Temporal Modules, Coreference Resolution, WSD, etc

Reality Check…

- Pre-processing is MOST of the processing!
- Need several lexicons that DO NOT exist openly for Portuguese, notably WordNet.
  - Spent the next four years working on OpenWordNet-PT.
- Not done BUT
- Google Translate, Open MultiLingual Wordnet, BabelNet, FreeLing use our Portuguese wordnet.

Goal: Reasoning & Inference

- **Which kind?**

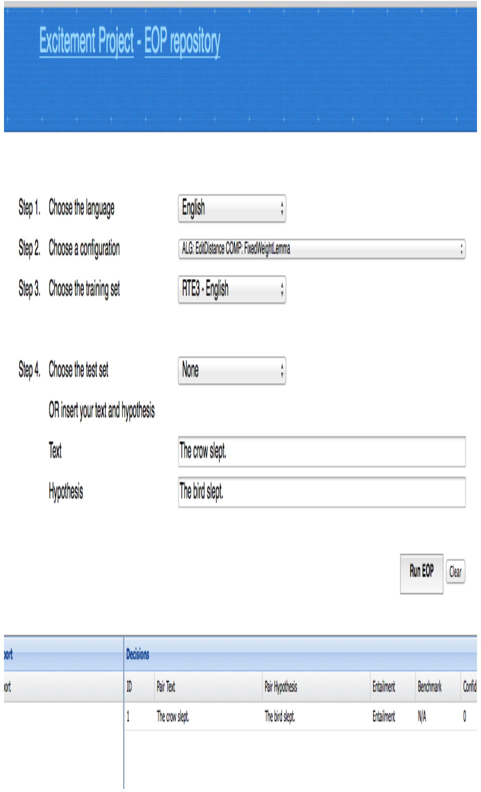
- **Textual entailment** methods recognize, generate, and extract pairs ⟨T,H⟩ of natural language expressions, such that a human who reads (and trusts) T would infer that H is most likely also true (Dagan, Glickman & Magnini, 2006)

- Example:
  (T) The drugs that slow down Alzheimer's disease work best the earlier you administer them.
  (H) Alzheimer's disease can be slowed down using drugs.
  $T \Rightarrow H$

- A series of competitions since 2004, ACL "Textual Entailment Portal", many different systems, some open source…

- **EOP Demo: EXCITEMENT Open Platform**

http://hlt-services4.fbk.eu/eop/index.php



Excitement Project - EOP repository

Step 1. Choose the language: English

Step 2. Choose a configuration: ALG: EditDistance COMP: FixedWeightLemma

Step 3. Choose the training set: RTE3 - English

Step 4. Choose the test set: None

OR insert your text and hypothesis

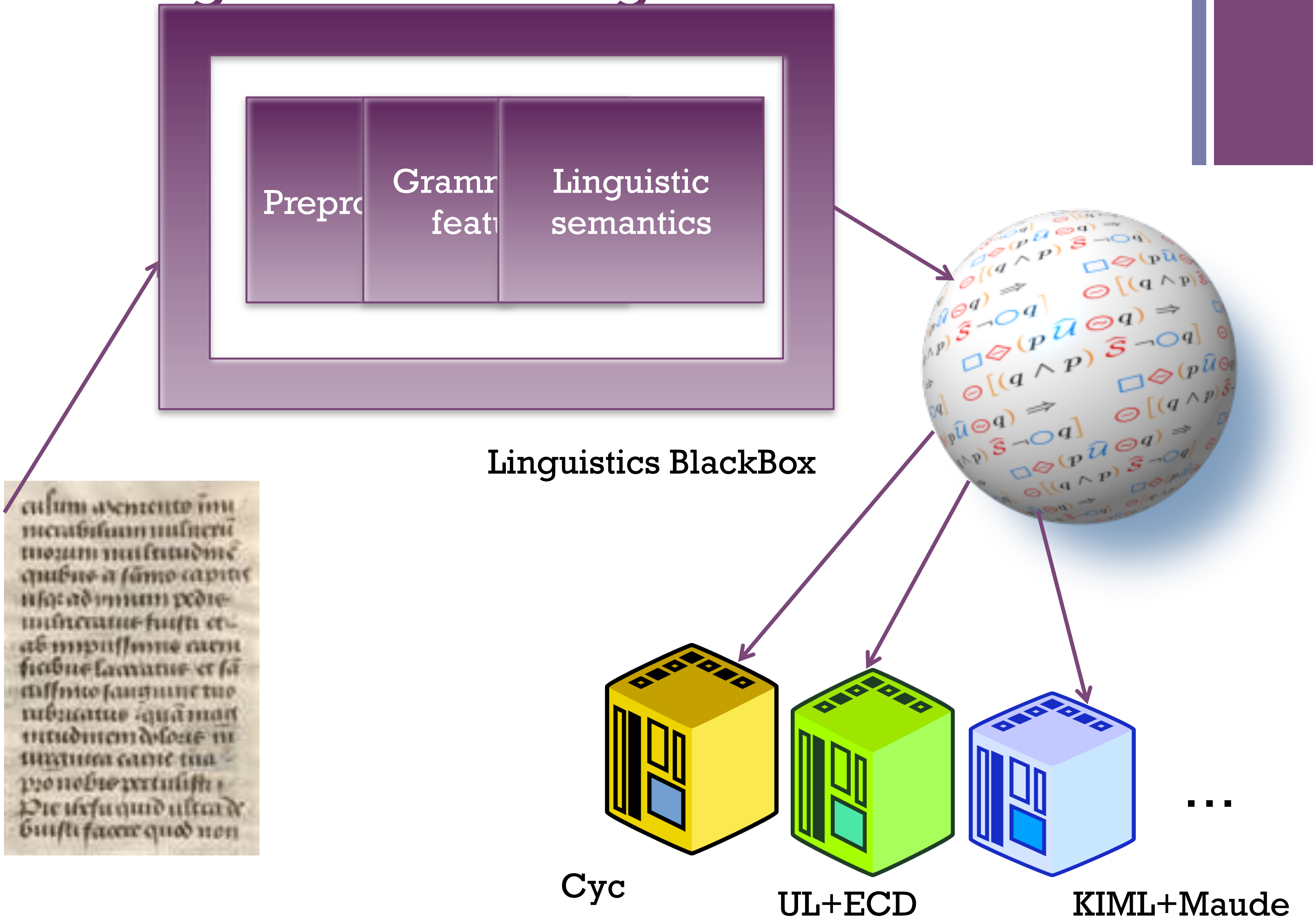
Text: The crow slept.

Hypothesis: The bird slept.

Decisions

ID	Pair Text	Pair Hypothesis	Entailment	Benchmark	Confid
1	The crow slept.	The bird slept.	Entailment	N/A	0

+ Bridge: Text to Logic



Model Theoretic and Proof Theoretic semantics

- Semantics for natural languages about creating **representations** for the sentences in a logical formalism (the **logical forms**).

- Also uncovering **truth conditions** for these translated sentences.

- This is Model Theoretic semantics

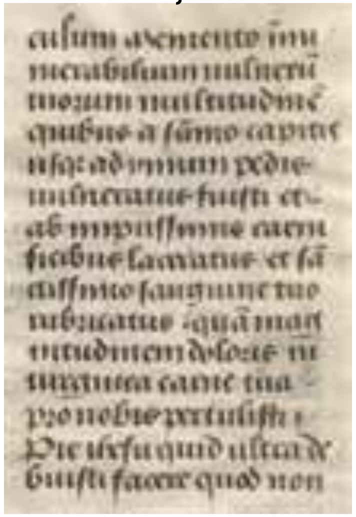
- An alternative view, the ***proof-theoretic paradigm*** of semantics: basic criterion is to establish when sentences follow from others, when they are consistent with each other, when they contradict each other. In short their entailment behavior. (cf. Schroeder-Heister, Francez)

- Relations of entailment and contradiction are key data of semantics. The ability to recognize such semantic relations is clearly **not a sufficient criterion** for language understanding: there is more than just being able to tell that one sentence follows from another. But we would argue that it is a **minimal, necessary criterion**.

- Hence **Lean Logic**

What can we do?
Logic and Lexical Ontologies

Group: Alexandre Rademaker, Livy Real, Claudia Freitas, Fabricio Chalub, Gerard de Melo, Suemi Higuchi, Hermann Haeusler, Bruno Lopes, Luiz Carlos Pereira, Vivek Nigam and Valeria de Paiva



Improving Lexical Resources and Inferential Systems to work with Logic coming from free form text.

The way to inference? KIML

- A representation language based on events (neo-Davidsonian), concepts, roles and contexts, McCarthy-style

- Using events, concepts and roles is traditional in NL semantics (Lasersohn)

- Usually equivalent to FOL (first-order logic), ours a small extension, contexts are like modalities. our Language based on linguists' intuitions

- Exact formulation still being decided: e.g. not considering temporal assertions, yet…

- What do we need for this representation language?

Example: a crow slept/um corvo dormiu

- Conceptual Structure: role(cardinality restriction,crow-1,sg) role(sb,sleep-4,crow-1) subconcept(crow-1, [crow#n#1,crow#n#2,brag#n#1]) subconcept(sleep-4, [sleep#v#1,sleep#v#2]

- Contextual Strux instantiable(crow-1,t) instantiable(sleep-4,t) top context(t) Temporal Structure: trole(when,sleep-4,interval(before,Now))

Need Concepts and Subconcepts

- These will be coming from **OPENWORDNET-PT**
- In the example “crow” and “sleep” → “corvo” and “dormir”
- Lexical concepts, organized into a lexical hierarchy
- Need taxonomy: a crow is a bird, sleep is a bodily function
- Need adaptation: different birds in Brazil, “corvo”= crow, but what is “gralha, grauna” ?
- Much harder on verbs, sleep=dormir, but cochilar, adormecer, pestanejar, tirar uma soneca…
- Mix “lab examples” with corpus ones.

Need predicate-argument structure

- in the example: role(sb,sleep-4,crow-1)
- Who's doing what to whom?
- How much do we canonicalize? Ted broke the window=the window was broken by Ted.
- ``the crow slept'' not the same as ``the crow was put to sleep'', but "Ed took a shower" is the same as "Ed showered."
- To get it, need parsing and grammar? (At least)Traditionally.
- To get parsing and grammar need POS-tagging, NER, MWEs, roles, semantic role labelling, etc…
- How to go about it?
- FreeLing our original option

FREELING

- FreeLing (Padro and Stanilovsky, 2012) is an open-source library of multilingual Natural Language Processing (NLP) tools that provide linguistic analysis for written texts.
- Freeling has been developed for more than ten years. It is a complete NLP pipeline built on a chain of modules that provide a general and robust linguistic analysis.
- Available tools in FreeLing: sentence recognition, tokenization, named entity recognition, tagging, chunking, dependency parsing, word sense disambiguation, and coreference resolution.

KIML versus FOL

- In FOL could write ∃Crow∃Sleep.Sleep(crow)
  Instead we will use basic concepts from a parameter ontology O
- O (could be Cyc, SUMO, UL, KM, etc...) But to begin with it will be WordNet or Open WordNet-PT.
- Instead of FOL have Skolem constant crow-1 a subconcept of an ambiguous list of concepts: subconcept(crow-1, [crow#n#1,crow#n#2,brag#n#1])
- Same for sleep-2 and have roles relating concepts
  role(sb,sleep-4,crow-1)
  meaning that the sb=subject of the sleeping event is a crow concept

What is Different?

- Corresponding to formulas in FOL, KIML has a collection of assertions that, read conjunctively, correspond to the semantics of a (fragment of a) sentence in English.

- Concepts in KIML – similar to Description Logic

  concepts primitive concepts from an idealized version of the chosen

- Ontology on-the-fly concepts, always sub-concepts of some primitive concept. concepts are as fine or as coarse as needed by the application

- Roles connect concepts: deciding which roles with which concepts a big problem… for linguists

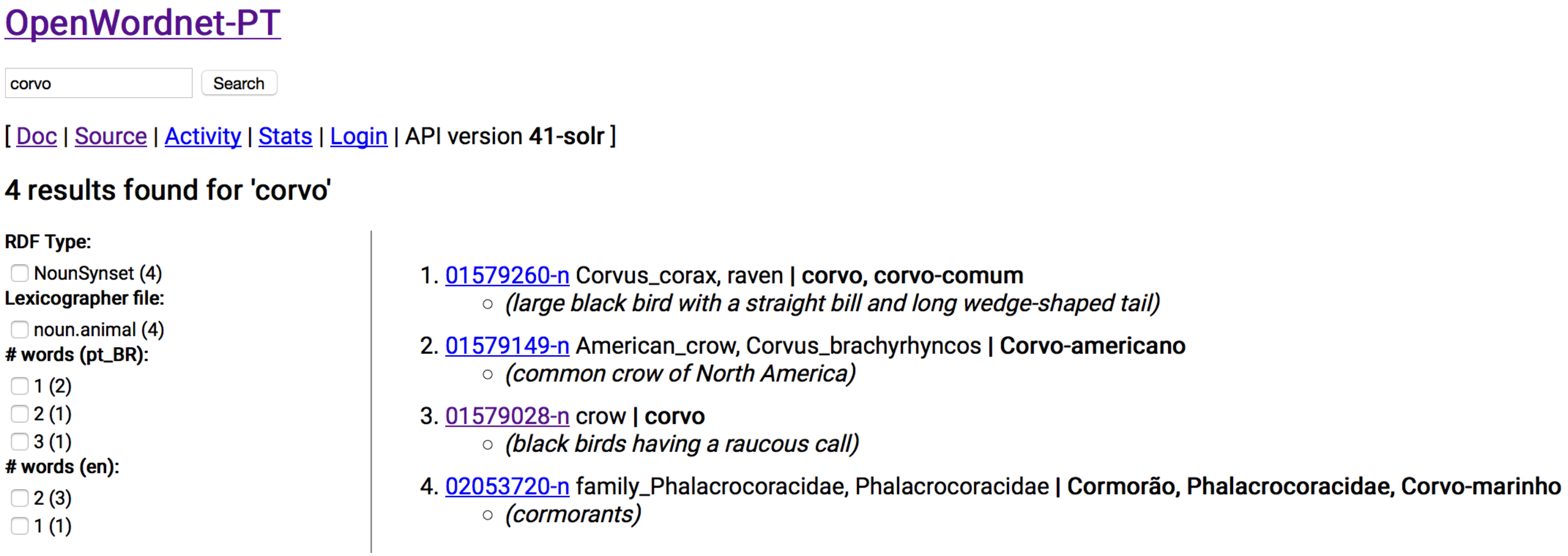
- Roles assigned in a consistent, coherent and maximally informative way by **the NLP module**

How to do this in Portuguese?

- Need concepts in Portuguese WordNet, as accurate as possible
- Need to decide if we're accurate enough
- Need syntactic structure (deep parsing? shallow will do?) in Portuguese
- Need NER, POS-tagging, multiword expressions recognition in Portuguese, ETC..
- Need to leverage extant open source work

Recent work: A collaborative editor for OpenWordnet-PT

- Available from https://github.com/own-pt/cl-wnbrowser
- Web interface in http://wnpt.brlcloud.com/wn/



OpenWordnet-PT

[Doc | Source | Activity | Stats | Login | API version **41-solr**]

4 results found for 'corvo'

RDF Type:

- □ NounSynset (4)

Lexicographer file:

- □ noun.animal (4)

# words (pt_BR):

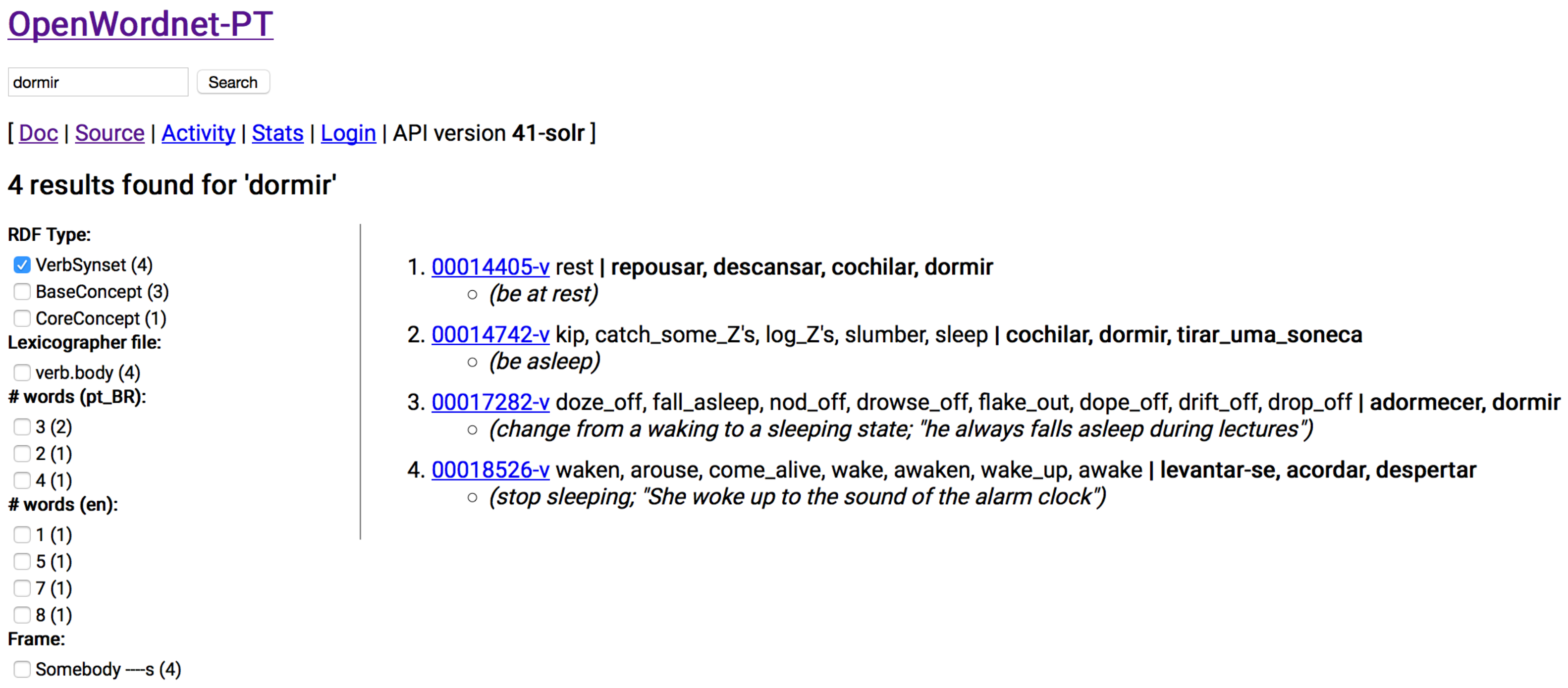
- □ 1 (2)
- □ 2 (1)
- □ 3 (1)

# words (en):

- □ 2 (3)
- □ 1 (1)

1. 01579260-n Corvus_corax, raven | **corvo, corvo-comum**
   - *(large black bird with a straight bill and long wedge-shaped tail)*
2. 01579149-n American_crow, Corvus_brachyrhyncos | **Corvo-americano**
   - *(common crow of North America)*
3. 01579028-n crow | **corvo**
   - *(black birds having a raucous call)*
4. 02053720-n family_Phalacrocoracidae, Phalacrocoracidae | **Cormorão, Phalacrocoracidae, Corvo-marinho**
   - *(cormorants)*

Example



OpenWordnet-PT

dormir

[Doc | Source | Activity | Stats | Login | API version **41-solr**]

4 results found for 'dormir'

RDF Type:

☑ VerbSynset (4)
□ BaseConcept (3)
□ CoreConcept (1)

Lexicographer file:

□ verb.body (4)

# words (pt_BR):

□ 3 (2)
□ 2 (1)
□ 4 (1)

# words (en):

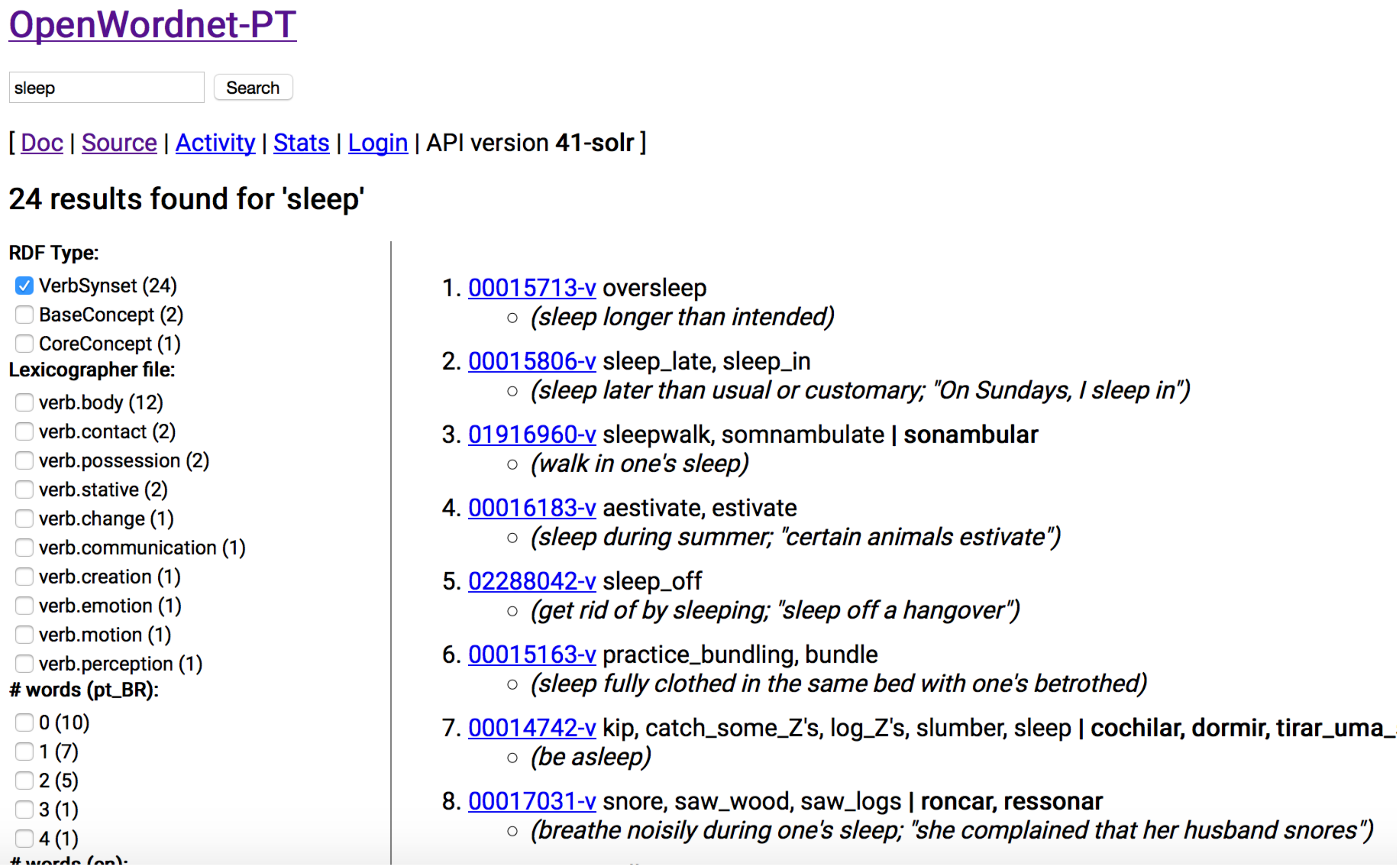
□ 1 (1)
□ 5 (1)
□ 7 (1)
□ 8 (1)

Frame:

□ Somebody ----s (4)

1. 00014405-v rest | **repousar, descansar, cochilar, dormir**
   - *(be at rest)*
2. 00014742-v kip, catch_some_Z's, log_Z's, slumber, sleep | **cochilar, dormir, tirar_uma_soneca**
   - *(be asleep)*
3. 00017282-v doze_off, fall_asleep, nod_off, drowse_off, flake_out, dope_off, drift_off, drop_off | **adormecer, dormir**
   - *(change from a waking to a sleeping state; "he always falls asleep during lectures")*
4. 00018526-v waken, arouse, come_alive, wake, awaken, wake_up, awake | **levantar-se, acordar, despertar**
   - *(stop sleeping; "She woke up to the sound of the alarm clock")*

Example again



OpenWordnet-PT

sleep Search

[Doc | Source | Activity | Stats | Login | API version **41-solr**]

24 results found for 'sleep'

RDF Type:

☑ VerbSynset (24)
□ BaseConcept (2)
□ CoreConcept (1)

Lexicographer file:

□ verb.body (12)
□ verb.contact (2)
□ verb.possession (2)
□ verb.stative (2)
□ verb.change (1)
□ verb.communication (1)
□ verb.creation (1)
□ verb.emotion (1)
□ verb.motion (1)
□ verb.perception (1)

# words (pt_BR):

□ 0 (10)
□ 1 (7)
□ 2 (5)
□ 3 (1)
□ 4 (1)

# words (en):

1. 00015713-v oversleep
   - *(sleep longer than intended)*
2. 00015806-v sleep_late, sleep_in
   - *(sleep later than usual or customary; "On Sundays, I sleep in")*
3. 01916960-v sleepwalk, somnambulate | **sonambular**
   - *(walk in one's sleep)*
4. 00016183-v aestivate, estivate
   - *(sleep during summer; "certain animals estivate")*
5. 02288042-v sleep_off
   - *(get rid of by sleeping; "sleep off a hangover")*
6. 00015163-v practice_bundling, bundle
   - *(sleep fully clothed in the same bed with one's betrothed)*
7. 00014742-v kip, catch_some_Z's, log_Z's, slumber, sleep | **cochilar, dormir, tirar_uma_**
   - *(be asleep)*
8. 00017031-v snore, saw_wood, saw_logs | **roncar, ressonar**
   - *(breathe noisily during one's sleep; "she complained that her husband snores")*

Linked Data OpenWordNet-PT

- Freely available since Dec 2011
- RDF based since its beginning
- SPARQL endpoint and RDF download
- Bootstrapped from the Portuguese subset of UWN (Gerard de Melo)
- Manually curated constantly improved either manually or by making use of corpora.
- Extended with nominalizations (NomLex-PT)

OpenWordNet-PT challenges

- Need new synsets that only exist in Portuguese. How to go about it?
- Need to decide guidelines for concepts that are lexicalized as a single word in English, e.g "oversleep", "jog"
- Need to decide what to do with Princeton's WN concepts with no direct correspondent in Portuguese? E.g "quarter", "United States Department of Treasury"
- Need to decide on ABOX-like concepts, e.g Barack Obama or Rio de Janeiro.

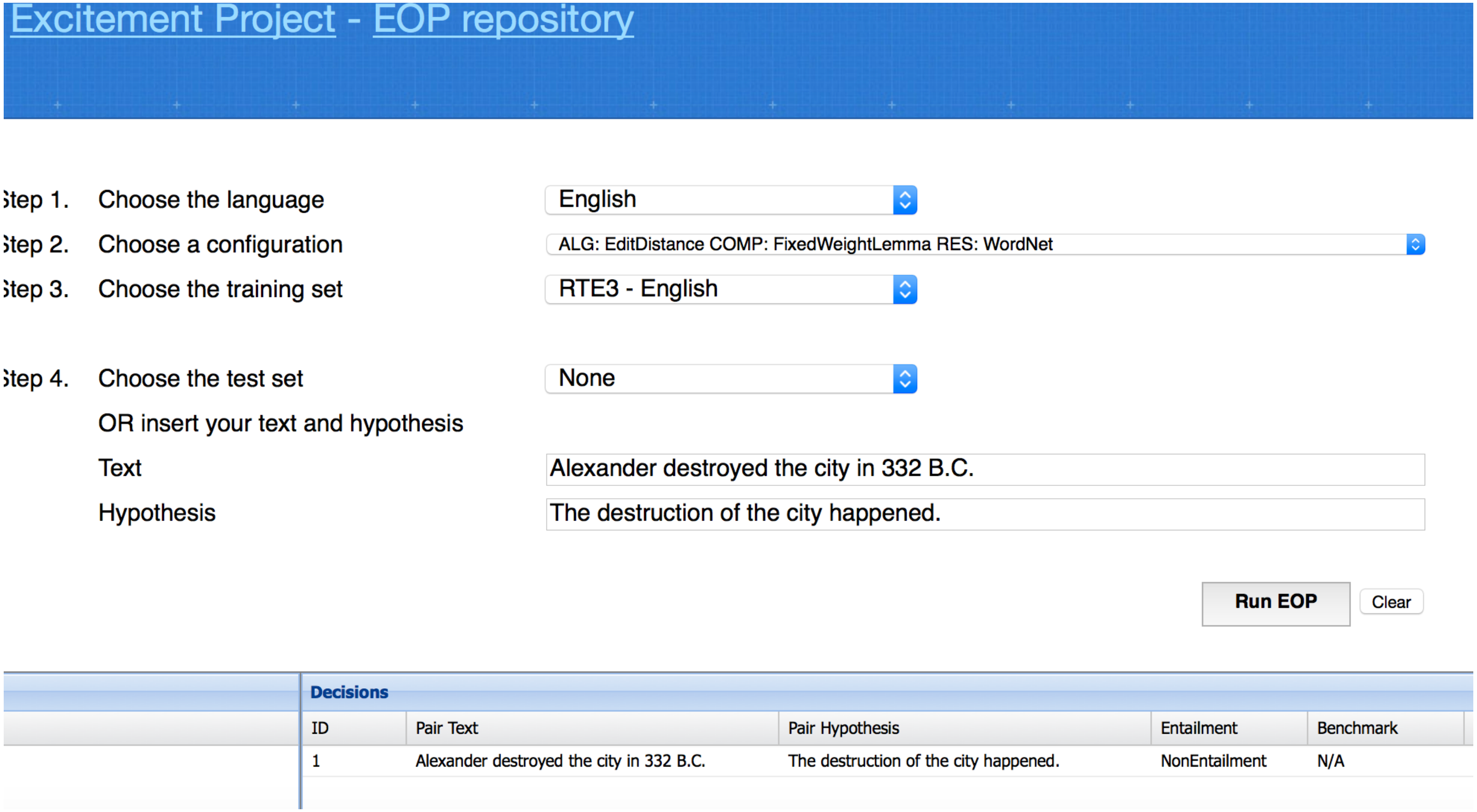
OpenWordNet-PT status

- 43,925 synsets, of which 32,696 correspond to nouns, 4,675 to verbs, 5,575 to adjectives and 979 to adverbs.
- Much smaller than PWN 117K synsets
- But more than twice the size of the Russian wordnet, bigger than the Spanish and just a little smaller than the French wordnet.
- Many minor mistakes such a capitalization and gender, number
- Some Galician, Spanish as Portuguese
- Portuguese variants?

OpenWordNet-PT status

- Faceted search for activites and synsets.
- Implemented with Common Lisp, Solr/Cloudant, NodeJS running in IBM BlueMix platform.
- couple of months online, over 4000 manual suggestions, and over 7000 votes have been cast and over 2600 comments. (5 people team)
- Links to other resources, PULO (over 125000 suggestions of glosses to evaluate,…)
- Voting mechanism Reddit-style
- ARE WE THERE YET?

Anyway need Inference to build reps and to reason with them



Excitement Project - EOP repository

Step 1. Choose the language: English

Step 2. Choose a configuration: ALG: EditDistance COMP: FixedWeightLemma RES: WordNet

Step 3. Choose the training set: RTE3 - English

Step 4. Choose the test set: None

OR insert your text and hypothesis

Text: Alexander destroyed the city in 332 B.C.

Hypothesis: The destruction of the city happened.

Decisions

ID	Pair Text	Pair Hypothesis	Entailment	Benchmark
1	Alexander destroyed the city in 332 B.C.	The destruction of the city happened.	NonEntailment	N/A

NomLex-PT incorporation would do it

- But other ways too.
- Configuration BIUTEE with CatVar does it too!
- No CatVar for Portuguese, no EOP for Portuguese
- OUR JOB.

How to get there?

- POS-tagging
- NER
- MWE
- Need all three for yesterday.
- Have discussed some of the issues to make up our minds on on Google tags paper.

Google POS-tags for Portuguese

- Need to do it.
- No one is doing it, yet.
- How to go about it?
- Need to see which issues arise.

Issues from J. Nivre's laundry list

- http://universaldependencies.github.io/docs/issues.html
- #4, #6, #7, #8, #9, #10 are particularly relevant
- Includes light verb constructions, MWEs, abbreviations, names and particles.
- Must discuss and decide.

Conclusions

- Proof-of-concept framework(s)
- Introduced a program of construction of lexical resources for construction of representations and reasoning
- Resources not good, but better than (all free?) others
- Use them to create representations
- Demonstrated by example that framework can be used to prove in an semi-automated fashion whether a sentence follows from another (proof by reference..)
- Many problems: removing black box, ambiguity, temporal information, etc..
- Many possible, sensible ways to move forward, how to organize the work?
- ’shallow theorem proving’ for common sense applications? Adam Pease

Thanks!

Coda

- Improve OpenWordNet-PT until what?
- Need a better Verb Lexicon
- Need a better adjectives lexicon
- Need subcategorization frames
- Need connections: Morpho-semantic links !
- Want Google tags and UDs for Portuguese, may be AMR too
- Want to do a bag of concepts experiment (DHBB?)
- Corpus work to make sure coverage in place

+

References

Revisiting a Brazilian Wordnet. Valeria de Paiva, Alexandre Rademaker, (2012)
Proceedings of Global Wordnet Conference, Global Wordnet Association, Matsue.
OpenWordNet-PT: An Open Brazilian WordNet For Reasoning. de Paiva, Valeria, Alexandre Rademaker, and Gerard de Melo. In *Proceedings of the 24th International Conference On Computational Linguistics*. http://hdl.handle.net/10438/10274.
OpenWordNet-PT: A Project Report. Alexandre Rademaker, Valeria de Paiva, Gerard de Melo, Livy Real and Maira Gatti.
Proceedings of the 7th Global Wordnet Conference, Tartu, Estonia. Global Wordnet Association, 2014.
Embedding NomLex-BR Nominalizations Into OpenWordnet-PT. Coelho, Livy Maria Real, Alexandre Rademaker, Valeria De Paiva, and Gerard de Melo. 2014. In *Proceedings of the 7th Global WordNet Conference*. Tartu, Estonia

References

Towards a Universal Wordnet by Learning from Combined Evidence Gerard de Melo, Gerhard Weikum (2009)
18th ACM Conference on Information and Knowledge Management (CIKM 2009), Hong Kong, China.

Bridges from Language to Logic: Concepts, Contexts and Ontologies Valeria de Paiva (2010)
Logical and Semantic Frameworks with Applications, LSFA'10, Natal, Brazil, 2010.

`A Basic Logic for Textual inference'', AAAI Workshop on Inference for Textual Question Answering, 2005.

``Textual Inference Logic: Take Two'', CONTEXT 2007.

``Precision-focused Textual Inference'',   Workshop on Textual Entailment and Paraphrasing, 2007.

PARC's Bridge and Question Answering System Proceedings of Grammar Engineering Across Frameworks, 2007.

+ Ed knows that the crow slept

- alias(Ed-0,[Ed])
  role(prop,know-1,ctx(sleep-8))
  role(sb,know-1,Ed-0)
  role(sb,sleep-8,crow-6)
  subconcept(Ed-0,[male#n#2])
  subconcept(crow-6,
  [crow#n#1,crow#n#2,brag#n#1])
  subconcept(know-1,[know#v#1,...,sleep-together#v#1]) subconcept(sleep-8,
  [sleep#v#1,sleep#v#2]) context(ctx(sleep-8)),
  context(t) context-lifting-relation(veridical,t,ctx(sleep-8)) context-relation(t,ctx(sleep-8),crel(prop,know-1))
  instantiable(Ed-0,t)
  instantiable(crow-6,ctx(sleep-8))
  instantiable(sleep-8,ctx(sleep-8))