Parallel Corpora-based Dictionaries and Their Applications

Summer School of Linguistics 2013
Braga, 10th September 2013

Nuno Ramos Carvalho, Rui Brito, Luís Miguel Braga,
Alberto Simões and José João Almeida

http://natura.di.uminho.pt

Table Of Contents

Part 0

Background Review

Background – Corpus

- collection of structured texts
- usually large and domain restricted
- stored in easy to process formats

Background – Parallel Corpora

- collection of texts in different languages, where each of them is a translation of each other
- used in NLP and Computational Linguistics
- aligned at the sentence level
- different formats
  - Text Encoding Initiative (TEI)
  - Translation Memory eXchange (TMX)
  - XML Corpus Encoding Standard (XCES)
- tools for aligining
  - Sentence Alignment: Hunalign, Vanilla Aligner, WinAlign,...
  - Word Alignment: GIZA++, NATools, ...

Background – Parallel Corpora Example

Estes resultados constituem a base do Programa Europeu de defesa do Mar de Barents e, por esse motivo, peço-lhe que analise um projecto de carta que lhe expõe os factos mais importantes, e que, de acordo com as decis oes do Parlamento, torne clara esta posiçãao na Russia.

No entanto, somos tambem da opiniãao de que deveria haver um debate sobre esta estrategia da comissão que seguisse um procedimento ordenado, e não so com base numa declaraçãao oral pronunciada aqui no Parlamento Europeu, mas tambem com base num documento que seja decidido na comissão e que apresente uma descricão deste programa para um perodo de cinco anos.

These fndings form the basis of the European Programmes to protect the Barents Sea, and that is why I would ask you to examine a draft letter setting out the most important facts and to make Parliament's position, as expressed in the resolutions which it has adopted, clear as far as Russia is concerned.

We believe, however, that the commission's strategic plan needs to be debated within a proper procedural framework, not only on the basis of an oral statement here in the European Parliament, but also on the basis of a document which is adopted in the commission and which describes this programme over the five-year period.

Background – Parallel Corpora Example

```
<?xml version="1.0"?>
<!DOCTYPE tmx SYSTEM "tmx11.dtd">
<tmx version="version 1.0">
<header creationtool="cwb-utils" creationtoolversion="1.0"
  segtype="sentence" adminlang="EN-US" srclang="pt">
</header>

<tu><!--1:1-->
   <tuv lang='pt'><seg>Constituição da República Portuguesa</seg></tuv>
   <tuv lang='es'><seg>CONSTITUCIÓN DE LA REPÚBLICA PORTUGUESA</seg></tuv>
</tu>

<tu><!--1:1-->
   <tuv lang='pt'><seg>IV REVISÃO CONSTITUCIONAL</seg></tuv>
   <tuv lang='es'><seg>Cuarta Revisión 1997</seg></tuv>
</tu>

<tu><!--1:1-->
   <tuv lang='pt'><seg>PREÂMBULO</seg></tuv>
   <tuv lang='es'><seg>PREÁMBULO</seg></tuv>
</tu>

<tu><!--1:1-->
   <tuv lang='pt'><seg>A 25 de Abril de 1974 , o Movimento das Forças Armadas ,
   <tuv lang='es'><seg>El 25 de Abril de 1974 , el Movimiento de las Fuerzas Ar
```

Part I

Introduction to Probabilistic Translation Dictionaries

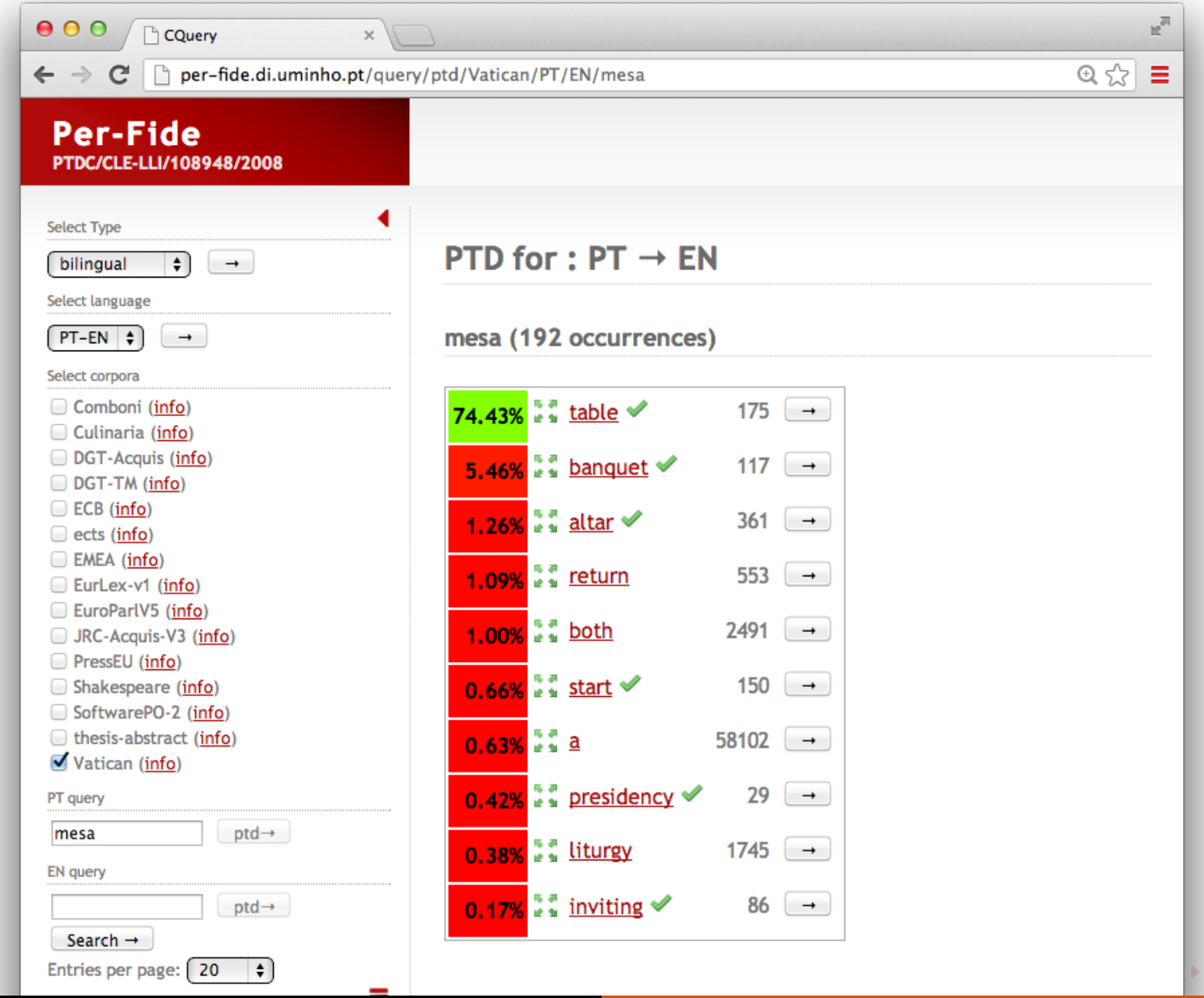
Introduction to PTD

- translation dictionaries
- probabilistic translations
- automatically extracted
- domain constrained
  - from parallel corpora, sentence aligned
- not a conventional dictionary
  - ... but can be transformed
- actually, a pair of dictionaries
  - example: $TMX_{pt \rightarrow en} \longrightarrow PTD_{pt \rightarrow en} \times PTD_{en \rightarrow pt}$
  - PTD from parallel corpora source to target language
  - PTD^{-1} (*inverse PTD*) from parallel corpora target to source language

PTD - Example 1

```
europe => {     ocorr => 42853,
                trans => {      europa    => 0.9471,
                                europeus  => 0.0339,
                                europeu   => 0.0081,
                                europeia  => 0.0011,
                          },
           },
stupid => {     ocorr => 180,
                trans => {      estúpido  => 0.1755,
                                estúpida  => 0.1099,
                                estúpidos => 0.0741,
                                avisada   => 0.0565,
                                direita   => 0.0558,
                                impasse   => 0.0448,
                           },
            },
```

PTD - Example 2



PTD Query - Example 1

```
QUERY> europa
  Occurrences: 39917
  Translations:
      88.50%  europe
       5.73%  european
       2.37%  europa
       1.16%  (none)
       0.57%  eu
       0.23%  unece
       0.17%  the
       0.16%  auto
```

PTD Query – Example 2

```
QUERY> we
  Occurrences: 300431
  Translations:
       17.81%  (none)
        8.25%  que
        6.02%  temos
```

PTD Query – Example 3

```
QUERY> read
  Occurrences: 2435
  Translations:
      29.32%  ler
      13.75%  li
       8.36%  read
       5.96%  lido
       3.54%  lemos
       1.60%  leio
       1.46%  estar
       1.45%  leu
```

```
QUERY> represent
  Occurrences: 2538
  Translations:
      17.87%  representam
      11.57%  representar
       8.93%  represento
       7.54%  representamos
       4.93%  constituem
       3.63%  representa
       3.37%  (none)
       2.35%  representante
```

PTD Query – Example 4

```
QUERY> palavra
 Occurrences: 6337
 Translations:
      35.75%  floor
      16.88%  word
      13.57%  (none)
      9.28%   speak
```

PTD Query – Example 5

```
QUERY> aceitável
 Occurrences: 1713
 Translations:
      71.48%  acceptable
      8.56%   unacceptable
```

PTD Extraction Algorithm

$$createPTD :: TU* \longrightarrow PTD$$

1. pre-processing stage
2. statistical algorithm stage
    1. build co-occurrence count table
    2. search highest values in matrix, to define *correct* relations
    3. when no highest values are found, use all others attenuated

- a flor cresce / a casa é grande / a casa azul tem flores
- the flower grows / the house is big / the blue house has flowers

Co-Occurrence Table

	a	flor	cresce	casa	é	grande	azul	tem	flores
the	3	1	1	2	1	1	1	1	1
flower	1	1	1	0	0	0	0	0	0
grows	1	1	1	0	0	0	0	0	0
house	2	0	0	2	1	1	1	1	1
is	1	0	0	1	1	1	0	0	0
big	1	0	0	1	1	1	0	0	0
blue	1	0	0	1	0	0	1	1	1
has	1	0	0	1	0	0	1	1	1
flowers	1	0	0	1	0	0	1	1	1

PTD – In The Forge

Currently working on:

- produce PTD from lemmas
- POS tagging

Part II

Resources Created from PTD

Bi-Words (BWs)

- strongly related word pair lists
- words may be translations
- example:
  - $\{\text{informed}\}_{en} = \{\text{informados}\}_{pt}$
  - $\{\text{approved}\}_{en} = \{\text{aprovado}\}_{pt}$
  - $\{\text{modern}\}_{en} = \{\text{modernos}\}_{pt}$

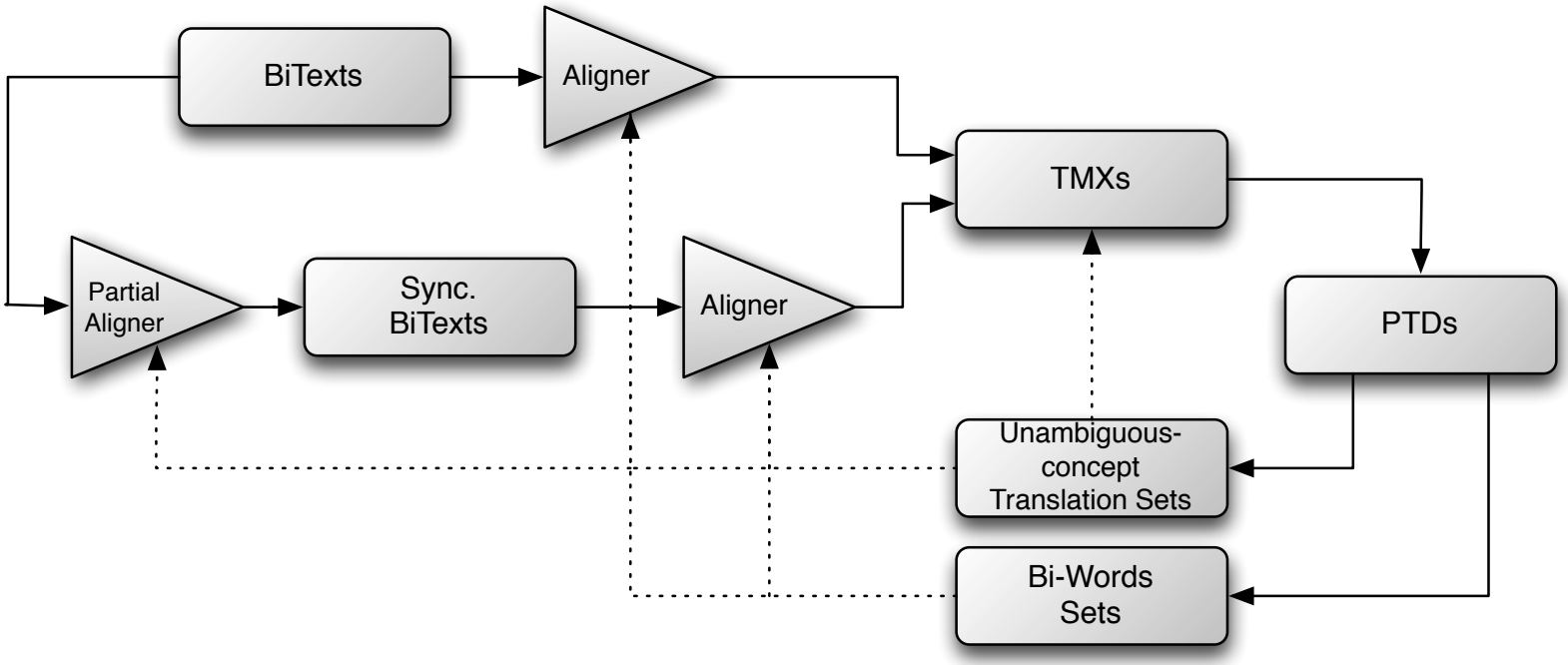
Unambiguous-Concept Translation Sets (UCTS)

- sets of equivalent words for two languages
- words that have always (or almost) the same translation
- stronger relations that BWs (smaller sets)
- examples:
  - $\{\text{London}\}_{en} = \{\text{Londres}\}_{pt}$
  - $\{\text{wolphram, tungsten}\}_{en} = \{\text{volfrâmio, tungsténio}\}_{pt}$
- useful for bootstrapping resources

BWs and UCTS

- extracted automatically from PTD
- more important when few resources are available
- provide connection clues
- used for:
  - partial synchronization/alignment
  - alignment assessment
  - translation quality assurance

BWs and UCTS – Bootstrapping Technique



PTD Algebra

- PTD can be automatically extracted
- PTD formal definition (not discussed here)
- other resources can be created
- set of operations that described how these resources are created
- way to clearly share workflows

Creating a *PTD* composed only of verbs (PTD^V) can be defined using the filter function as:

$$PTD^V = filter(PTD, verb)$$

where, *verb* is a function defined as:

$$verb(entry) = \begin{cases} True & \text{if } word \text{ in } entry \text{ is a verb} \\ False & \text{otherwise} \end{cases}$$

PTD Algebra – Operations Summary

Function	Syntax	Domain		Range
domain	$\mathrm{dom}\,(\_)$	$\mathrm{ptd}_{A\to B}$	$\longrightarrow$	${w_A}^*$
range	$\mathrm{ran}\,(\_)$	$\mathrm{ptd}_{A\to B}$	$\longrightarrow$	${w_B}^*$
union	$\_ \cup \_$	$\mathrm{ptd}_{A\to B} \times \mathrm{ptd}_{A\to B}$	$\longrightarrow$	$\mathrm{ptd}_{A\to B}$
intersection	$\_ \cap \_$	$\mathrm{ptd}_{A\to B} \times \mathrm{ptd}_{A\to B}$	$\longrightarrow$	$\mathrm{ptd}_{A\to B}$
composition	$\_ \circ \_$	$\mathrm{ptd}_{A\to B} \times \mathrm{ptd}_{B\to C}$	$\longrightarrow$	$\mathrm{ptd}_{A\to C}$
domain restrict	$\_ / \_$	$\mathrm{ptd}_{A\to B} \times {w_A}^*$	$\longrightarrow$	$\mathrm{ptd}_{A\to B}$
domain subtract	$\_ \setminus \_$	$\mathrm{ptd}_{A\to B} \times {w_A}^*$	$\longrightarrow$	$\mathrm{ptd}_{A\to B}$
totalization	$\mathrm{tot}\,(\_)$	$\mathrm{ptd}_{A\to B}$	$\longrightarrow$	$\mathrm{ptd}_{A\to B}$
filter	$\mathrm{filter}\,(\_,\_)$	$\mathrm{ptd}_{A\to B} \times (\mathrm{entry}_{A\to B} \longrightarrow \mathit{Bool})$	$\longrightarrow$	$\mathrm{ptd}_{A\to B}$
map	$\mathrm{map}\,(\_,\_)$	$\mathrm{ptd}_{A\to B} \times (\mathrm{entry}_{A\to B} \longrightarrow \mathrm{entry}_{A\to B})$	$\rightarrow$	$\mathrm{ptd}_{A\to B}$

Part III

Other Applications

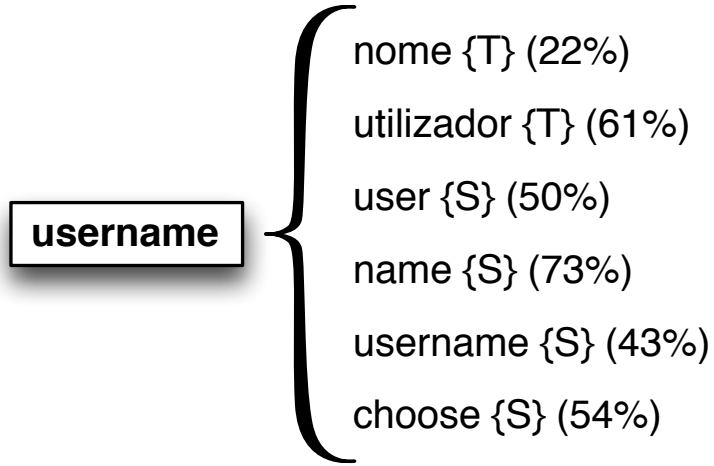
Dictionary Composition

- some languages have a limited amount of available resource
- less available man power to create them
- come up with strategies to create them more efficiently

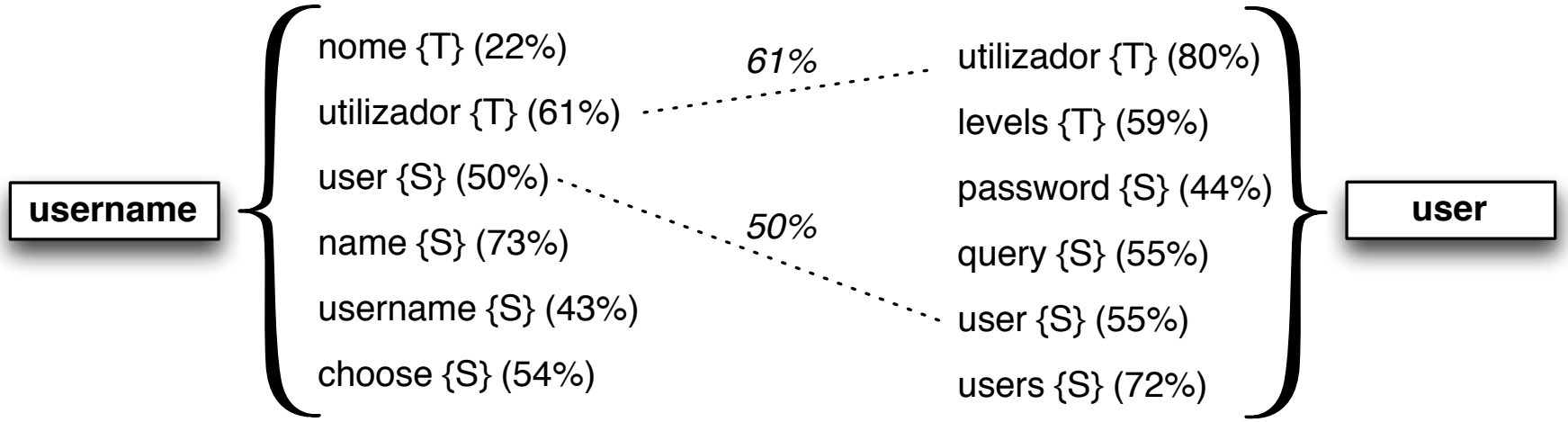
- start with: $PTD_{A \to B}$ and $PTD_{B \to C}$
- apply a composition function so that: $PTD_{A \to C} = PTD_{B \to C} \circ PTD_{A \to B}$
- addressing number of occurrences, which words appear, and how probabilities are mapped
- result is: $PTD_{A \to C}$

Dictionary Semantic Transitivity

- Syn Sets are required by many algorithms and techniques
- to discover words with the same meaning (or close)
- compute Probabilistic Syn Sets (PSS) from PTD
  - multi-lang Syn Sets, with weights, and domain constrained
- to compute $PSS(term)$ include:
  - all the possible translations and probabilities for $PTD(term)$
  - and, for each of these possible translations the inverse PTD translations and corresponding probabilities
- example:



- synonym search
- word similarity
- concepts similarity



Export *PTD* to use in OmegaT, as glossary

```
$ nat-ptd toTSV my_ptd.ptd glossary.tsv
```

Export *PTD* to use in OmegaT, as dictionary

```
$ nat-ptd toStarDict my_ptd.ptd star_dict
```

Software

- Lingua::NATools
  `http://search.cpan.org/dist/Lingua-NATools/`
- Lingua::PTD
  `http://search.cpan.org/dist/Lingua-PTD/`

Online

- `http://per-fide.di.uminho.pt/query`
- `http://ptd.natura.di.uminho.pt/`

References

- Structural Alignment of Plain Text Books
  in LREC2012
- Defining a Probabilistic Translation Dictionaries Algebra
  in EPIA2013 (*forthcoming*)
- Translation Dictionaries Triangulation
  in Iberian SLTech Workshop 2012
- XML Schemas for Parallel Corpora
  in XATA2011
- Extracção de Recursos de Tradução com base em Dicionários Probabilísticos de Tradução
  PhD thesis

Thank You!

Nuno Carvalho <narcarvcalho@di.uminho.pt>