Workshop on the Apertium free/open-source machine translation platform: basics on how to control the engine through linguistics

5th/6th November 2014

Session 1: Introduction to machine translation [15 min.]

What is “machine translation”?

- Machine translation is translating texts from one language to another with the help of computer programs.

How is it used ? (Assimilation)

To get a rough idea of a text when you don't speak the language or you speak it badly. I don't speak Breton,

```
Ofis Publik ar Brezhoneg: Brudañ ar yezh ha skoazellañ anezhi d'en em zispakañ war
holl dachennoù implij ur yezh zo e-touez kefridioù pennañ ar benveg.
```

nor Basque,

```
Txillidaren obraren katalogo lehen liburukia kalean da.
```

But with machine translation, I can get by – in a limited fashion.

How is it used ? (Assimilation)

To get a rough idea of a text when you don't speak the language or you speak it badly. I don't speak Breton,

```
Ofis Publik ar Brezhoneg: Brudañ ar yezh ha skoazellañ anezhi d'en em zispakañ war
holl dachennoù implij ur yezh zo e-touez kefridioù pennañ ar benveg.
```

l'Office Public de la Langue Bretonne : faire connaître la langue et l'aider à déployer sur tout les terrains de l'emploi d'une langue sont parmi les missions principales de l'outil.

nor Basque,

```
Txillidaren obraren katalogo lehen liburukia kalean da.
```

Txillidaren De la obra katalogo el primer tomo en la calle es.

But with machine translation, I can get by – in a limited fashion.

How is it used? (Dissemination)

You have a text in Spanish, and you want to translate it to Portuguese. You first translate the text with the help of machine translation, and then you need to only make a few changes before it is adequate.

Cheboksary es una ciudad del centro de Rusia europea, capital de la República de Chuvashia y puerto del río Volga. Hay fábricas textiles y de artículos de madera y cuero. También hay una central hidroeléctrica. Fundada en el siglo XIV, Cheboksary se transformó en un importante núcleo económico tras finalizarse el enlace ferroviario con Kanash en 1939. En esta ciudad se encuentra la Universidad Estatal Chuvashia Ulyánov (1967).

How is it used? (Dissemination)

You have a text in Spanish, and you want to translate it to Portuguese. You first translate the text with the help of machine translation, and then you need to only make a few changes before it is adequate.

Cheboksary es una ciudad del centro de Rusia europea, capital de la República de Chuvashia y puerto del río Volga. Hay fábricas textiles y de artículos de madera y cuero. También hay una central hidroeléctrica. Fundada en el siglo XIV, Cheboksary se transformó en un importante núcleo económico tras finalizarse el enlace ferroviario con Kanash en 1939. En esta ciudad se encuentra la Universidad Estatal Chuvashia Ulyánov (1967).

Cheboksary é uma cidade do centro da Rússia européia, capital da República de Chuváchia e porto do rio Volga. Há fábricas têxteis e de artigos de madeira e couro. Também há uma central hidroelétrica. Fundada no século XIV, Cheboksary transformou-se en um importante núcleo económico depois de finalizar-se o enlace ferroviário com Kanash em 1939. Nesta cidade encontra-se a Universidade Estatal Chuváchia *Ulyánov (1967).

How is it used? (Dissemination)

You have a text in Spanish, and you want to translate it to Portuguese. You first translate the text with the help of machine translation, and then you need to only make a few changes before it is adequate.

Cheboksary es una ciudad del centro de Rusia europea, capital de la República de Chuvashia y puerto del río Volga. Hay fábricas textiles y de artículos de madera y cuero. También hay una central hidroeléctrica. Fundada en el siglo XIV, Cheboksary se transformó en un importante núcleo económico tras finalizarse el enlace ferroviario con Kanash en 1939. En esta ciudad se encuentra la Universidad Estatal Chuvashia Ulyánov (1967).

Cheboksary é uma cidade do centro da Rússia européia, capital da República de Chuváchia e um porto do rio Volga. Há fábricas têxteis e de artigos de madeira e couro. Também há uma central hidroelétrica. Fundada no século XIV, Cheboksary transformou-se en um importante núcleo econômico depois da conexão ferroviária com Kanash ser finalizada em 1939. Nesta cidade encontra-se a Universidade Estatal Chuváchia Ulyanov (1967).

How is it used? (Dissemination)

You have a text in Spanish, and you want to translate it to Portuguese. You first translate the text with the help of machine translation, and then you need to only make a few changes before it is adequate.

Cheboksary es una ciudad del centro de Rusia europea, capital de la República de Chuvashia y puerto del río Volga. Hay fábricas textiles y de artículos de madera y cuero. También hay una central hidroeléctrica. Fundada en el siglo XIV, Cheboksary se transformó en un importante núcleo económico tras finalizarse el enlace ferroviario con Kanash en 1939. En esta ciudad se encuentra la Universidad Estatal Chuvashia Ulyánov (1967).

Cheboksary é uma cidade do centro da Rússia européia, capital da República de Chuváchia e um porto do rio Volga. Há fábricas têxteis e de artigos de madeira e couro. Também há uma central hidroelétrica. Fundada no século XIV, Cheboksary transformou-se en um importante núcleo econômico depois da conexão ferroviária com Kanash ser finalizada em 1939. Nesta cidade encontra-se a Universidade Estatal Chuváchia Ulyanov (1967).

Is there a difference?

	Necessary	**Unnecessary**
Assimilation	Understandability Fast translation	Syntactic *correctness* Lexical *correctness* Predictable errors Happy translators
Dissemination	Adequate syntactic transfer Predictable errors High accuracy (WER $\leq$ 15%) Happy translators	Understandability Fast translation

With the binoculars the hat-having man sees the squirrel.

Is there a difference?

	Necessary	**Unnecessary**
Assimilation	Understandability Fast translation	Syntactic *correctness* Lexical *correctness* Predictable errors Happy translators
Dissemination	Adequate syntactic transfer Predictable errors High accuracy (WER $\leq$ 15%) Happy translators	Understandability Fast translation

With the binoculars the hat-having man sees the squirrel.

The man wearing a hat sees the squirrel with the binoculars.

Is there a difference?

	Necessary	**Unnecessary**
Assimilation	Understandability Fast translation	Syntactic *correctness* Lexical *correctness* Predictable errors Happy translators
Dissemination	Adequate syntactic transfer Predictable errors High accuracy (WER $\leq$ 15%) Happy translators	Understandability Fast translation

The migration gave a great deal of criticism when it spoke out.

Is there a difference?

	Necessary	**Unnecessary**
Assimilation	Understandability Fast translation	Syntactic *correctness* Lexical *correctness* Predictable errors Happy translators
Dissemination	Adequate syntactic transfer Predictable errors High accuracy (WER $\leq$ 15%) Happy translators	Understandability Fast translation

The migration gave a great deal of criticism when it spoke out.

The organisation received a great deal of criticism when it spoke out.

Typology of machine translation systems

Kinds of machine translation

Rule-based
dictionaries and rules

lorem	лорем
ipsum	ипсум
dolor	долор
sit	сит
amet	амет
labore	лаборе
sed	сед
do	до
epismod	эписмод

```
SUB V NP → NP V
SUB NP VP → VP NP
SUB NP → DET NP
SUB N A → A N
SUB NP RELP → RELP NP
SUB REL VP → VP REL
...
```

Corpus-based
existing translations
of sentences

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation

Лорем ипсум долор сит амет, консететур адиписичинг элит, сед до эиусмод темпор инсидидунт ут лаборе эт долоре магна аликва Ут эним ад миним вениам, квис ноструд эксерситатион.

Lorem	Лорем	1.0
sit amet	сит амет	0.5
et dolore	амет,	0.3
Ut enim	Ут эним ад	0.6
ad	Миним	0.1
Minim	, квис	0.1
, quis	миним	0.3
sed do	сед до	0.5
elit	элит	1.0
veniam	вениам	0.9

Rule-based machine translation

Strengths	Weaknesses

Rule-based machine translation is like taking a set of dictionaries and a descriptive grammar, and trying to translate from one language you don't know into another.

Corpus-based machine translation

Strengths	Weaknesses

Corpus-based machine translation is like taking two documents in two languages you don't know which are translations of each other and trying to match up words. Then you use these words to build sentences which you put into Google to see if they sound likely.

Corpus-based machine translation works best when...

- You have a big corpus of pre-translated and aligned sentences from one language to another — or programmers who don't mind doing the alignment
- The language to be translated into is not morphologically complex — and the language to be translated from is more morphologically complex.
- The domain you want to translate is the same or similar as the one of your corpus.
- You lack linguists who are interested and motivated.

Rule-based machine translation works best when...

- You don't have any pre-aligned corpora, or the pre-aligned corpora you have are bad.
- The languages to be translated are typologically similar.
- You are translating formal language.
- You have interested and motivated linguists.

Practice 1: Taking a look to machine translation systems [30 min.]

Rule-based machine translation

Strengths	**Weaknesses**
+ Predictable output	- Lack of fluency
+ Predictable errors!	- Lack of idiomaticness
+ Incremental improvements	- “Mechanical” output
+ Translation errors traceable	- Development can be time consuming
+ Terminology control easy	
+ No need for large quantity of existing translations	

Corpus-based machine translation

Strengths	**Weaknesses**
+ Fluent output	- Unpredictable
+ Idiomatic output	- Incremental improvements are hard
+ No need for linguistic resources: - dictionaries - grammars - linguists ☹	- Development can be time consuming

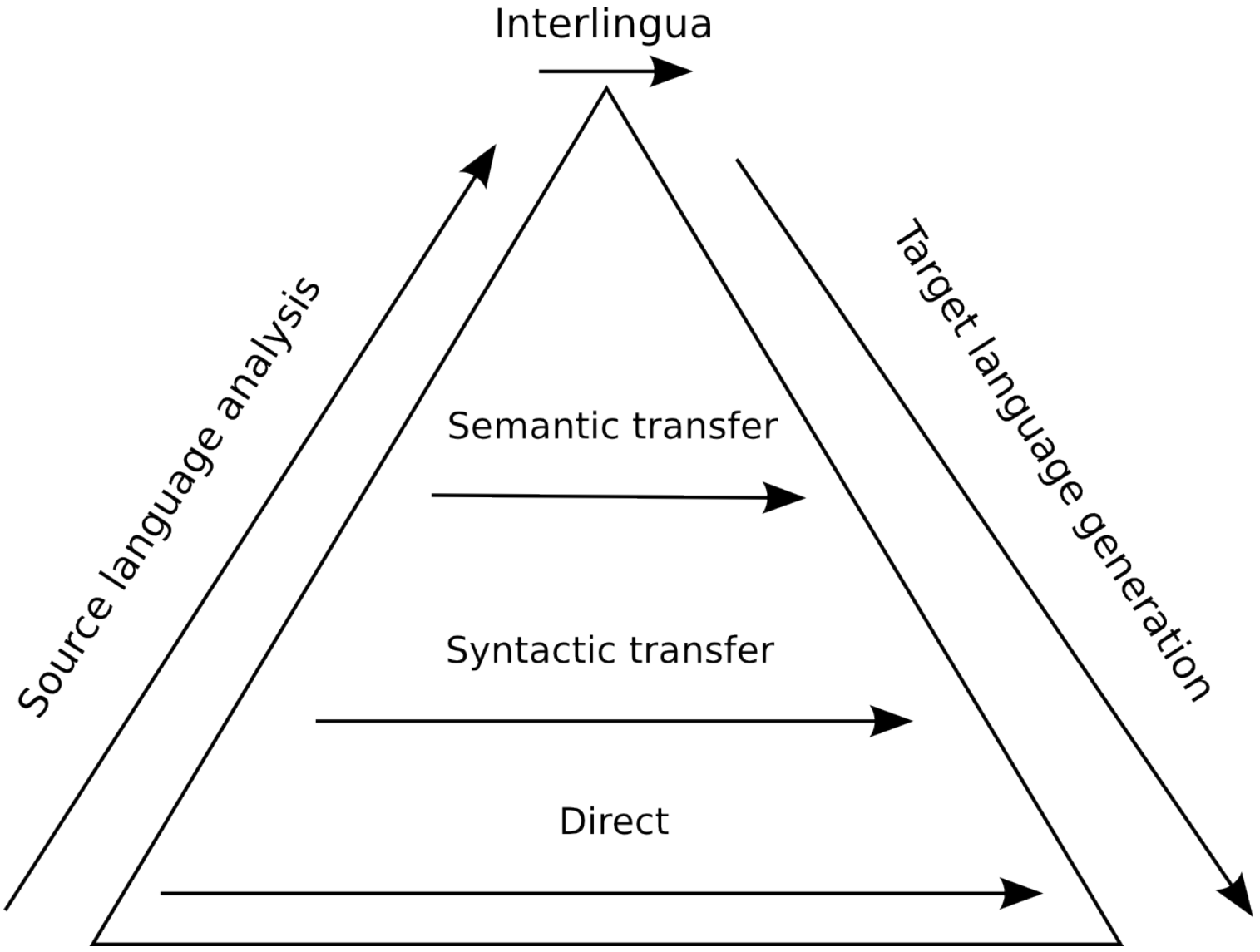
Session 2: Rule-based Machine Translation [20 min.]

Why do we work on rule-based machine translation ?

Machine translation conferences are full of papers about corpus-based MT, so why work on rule-based MT ?

- Sometimes there are no corpora, or only rubbish corpora
- When we codify translation rules, it tells us something about language(s) and translation
- We can produce useful systems! – really!
- Languages are interesting
- It's really fun!

Die Pyramide der maschinellen Übersetzung



Intermediate representation

The idea of an *intermediate representation* is to provide an abstraction of the meaning of the text.

- Direct translation: No intermediate representation
- Syntax transfer: Intermediate representation is either a parse tree, or a graph, along with feature structures
- Semantic transfer: Intermediate representation are predicates with semantic rôles.
- Interlingua: As with semantic transfer, only the same intermediate representation is shared by all languages / language pairs

This traditional division leaves out the Apertium approach:

- Shallow transfer: The intermediate representation can be
  - lexical forms – combinations of lemma and part-of-speech
  - chunks – collections of words into segments broadly reflecting phrases

Problems in rule-based machine translation

Analysis

Form does not entirely determine content.

Many sentences in natural language can have more than one interpretation, and these interpretations may be translated differently in different languages.

- *Traían noticias de Grecia* – theme ‘about’ or source ‘from’?
  - Traziam notícias da Grécia?
  - Traían noticias de Grecía?
- *I saw the girl with the telescope* – who has the telescope?
  - J’ai vu la fille avec le télescope
  - J’ai vu la fille à traves le télescope

The machine only knows as much as you can explain to it.

Synthesis

Content does not entirely determine form.

A single meaning can be expressed in more than one way. A single sentence may have many adequate equivalents.

- What time is it? (en)

- ¿Qué hora es? ¿Qué hora tienes? ¿Me dices la hora? (es)
- Quelle heure est-il? Vous avez l'heure? (fr)
- Que horas são? Que horas tem? (pt)

Transfer

The same content is represented differently in different languages.

Languages differ how they express a particular meaning. Some languages encode facets of meaning which are not encoded by others.

- Definiteness
- Case
- Direction and class of movement

- **Me gusta Croatia**
  - Me = object gusta = verb Croacia = subject
- **I like Croatia**
  - I = subject like = verb Croatia = object
- **Sviđa mi se Hrvatska**
  - Sviđa = verb mi = object se = reflexive Hrvatska = textcolorbluesubject

Description

Representing knowledge about the translation process in machine-readable form

It is difficult to indicate in an explicit and declarative way which is the process we use to translate to a machine. Concepts as ”most common sense” or ”context” are not trivial for machines.

Description

Representing knowledge about the translation process in machine-readable form

Fortunately, sometimes nothing of this is necessary:

- Apertium é um **sistema** de tradução automática. (pt)
- Apertium ei un **sistèma** de traduccion automatica. (oc)
- Apertium és un **sistema** de traducció automàtica. (ca)
- Apertium este o **platformă** de traducere automată. (ro)
- Apertium je **platforma** za računalniško prevajanje. (sl)
- Apertium je **platforma** za računarsko prevođenje. (bs)
- Apertium je **platforma** za računalno prevođenje. (hr)
- Apertium je **platforma** za kompjutersko prevođenje. (sr)

Apertium

Apertium: free/open source RBMT plaftorm

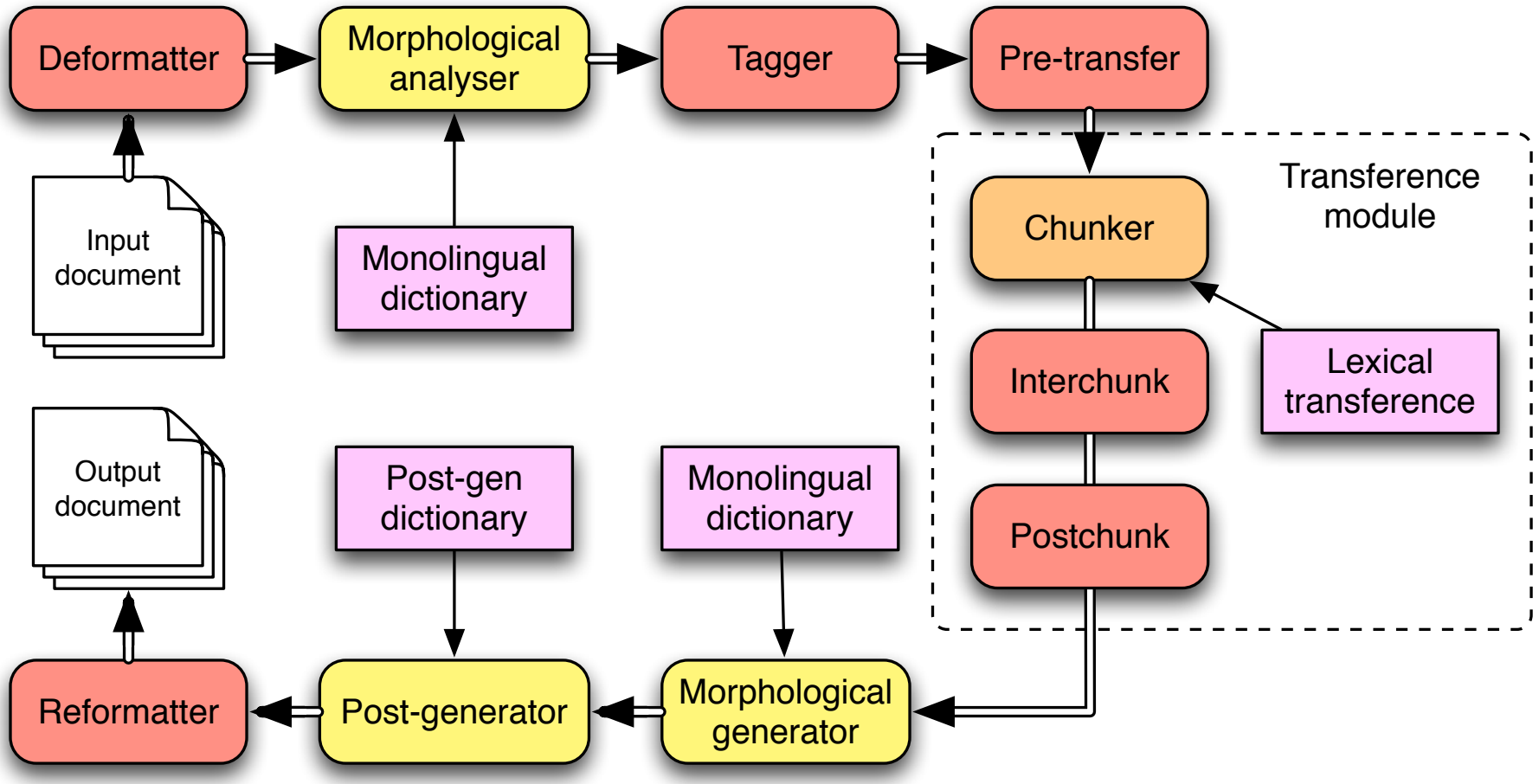
- 2005: engine, tool, language pairs = GNU-GPL v2
- rule-based: focus on related languages and less-resourced languages
- standards: C++, XML, code and linguistic data are decoupled
- modular, robust, fast: Unix pipes, works on any PC, 10.000 words/second
- developed by computer engineers and linguists
- big documentation, support and community

Apertium: free/open source RBMT plaftorm

- funding: public (research projects) and private (companies, GSoC, individuals)
- opportunities:
  - research: 5 masters, 2 PhD, 70 papers, 6 research projects
  - bussiness: services around Apertium – Prompsit (and others)
  - languages: some ”first systems” – Breton, Occitan, Afrikaans

Apertium: architecture

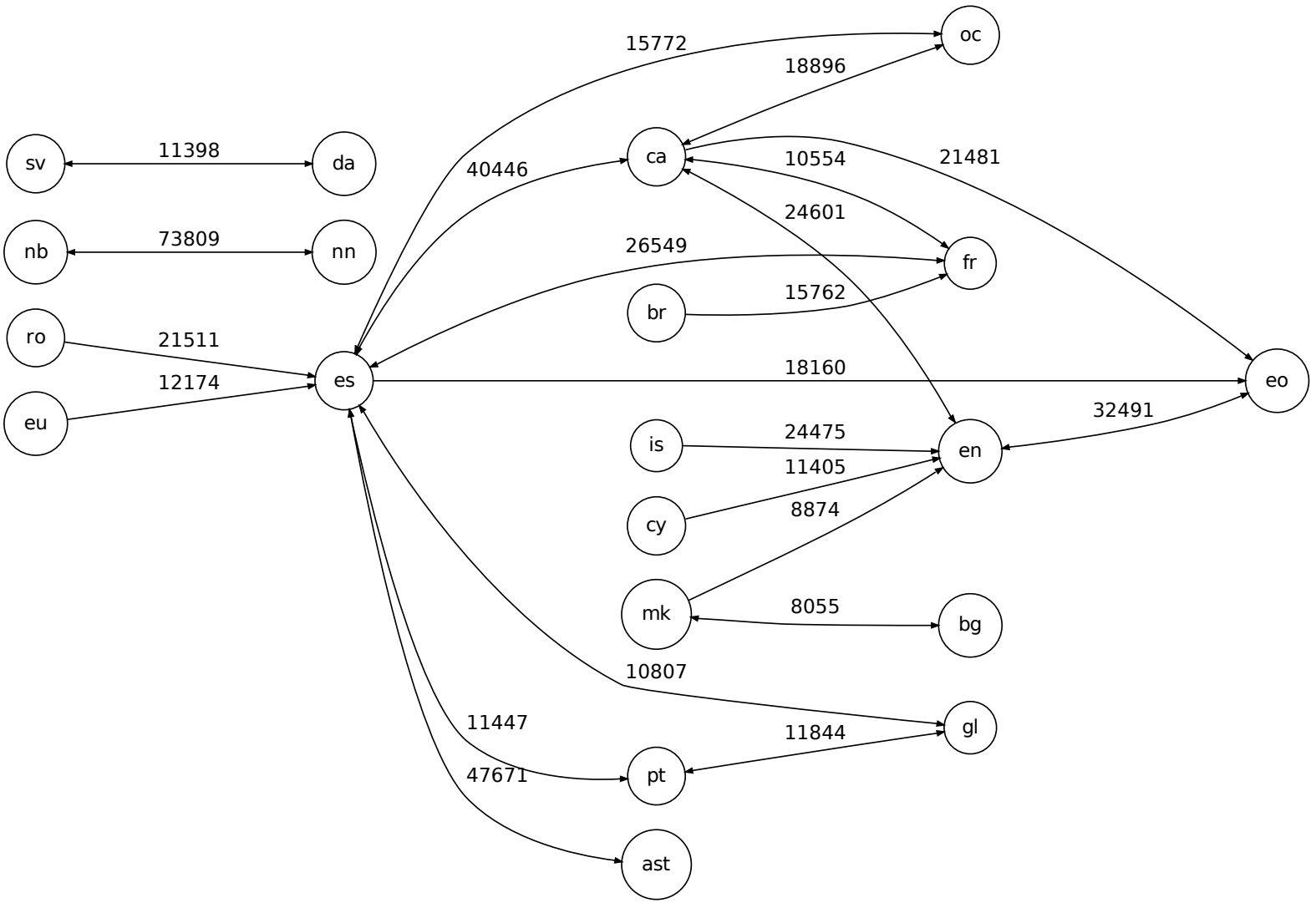
- Classic shallow-transfer system
- Pipeline made by 8 independent modules:



Apertium: performance

- reasonable quality for closely-related languages:
  - word error rate around 5% for general purpose texts
  - naive coverage around 95%
  - dictonaries with a minimum of 10,000 lemas and some 80 frequent transfer rules

Apertium: languages



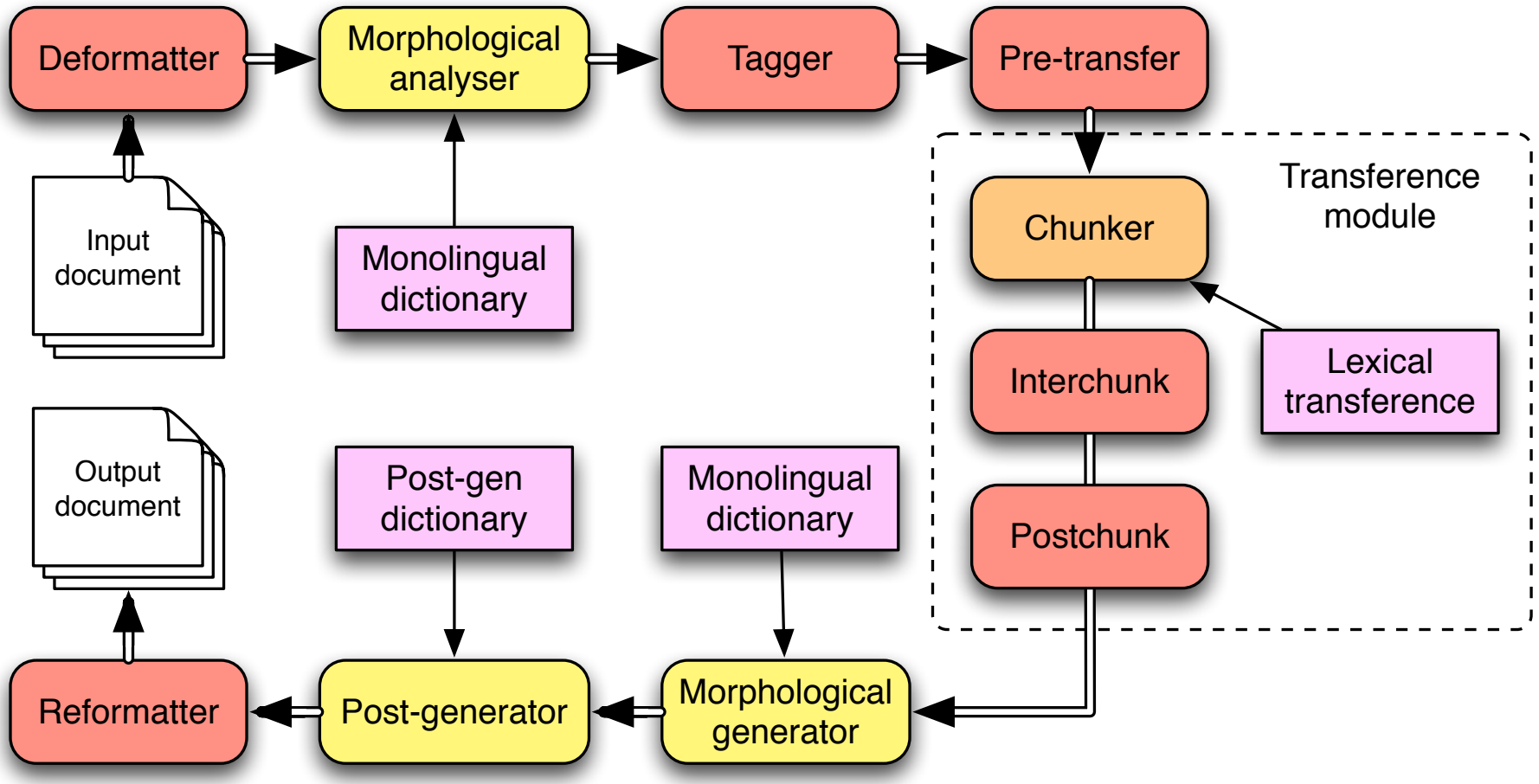
Apertium: more info

- code and languages freely available at Sourceforge `http://sourceforge.net/projects/apertium`
- informationa, developers material, tools, interfaces, chat and much more at: `http://wiki.apertium.org`
- testing interface at:`http://www.apertium.org`

Practice 2:Taking a look to Apertium with apertium-viewer [30 min.]

3. Monolingual entries [20 min.]

Which data are necessary for a language pair?



Language pair structure

- Dictionaries (.dix/.metadix)
  - apertium-es-pt.es.dix, Spanish monodix
  - apertium-es-pt.pt.dix, Portuguese monodix
  - apertium-es-pt.es-pt.dix, bidix Spanish-Portuguese
  - apertium-es-pt.post-es.dix, Spanish post-generator
  - apertium-es-pt.post-pt.dix, Portuguese post-generator
- Tagger (tsx)
  - apertium-es-pt.es.tsx
  - apertium-es-pt.pt.tsx
- Rules (.t1x, .t2x, .t3x)
  - apertium-es-pt.es-pt.t1x
  - apertium-es-pt.pt-es.t1x

Dictionaries

MONOLINGUAL DICTIONARY

- Alphabet definition
- Symbol definition
- Paradigm declaration
- Sections of entries

BILINGUAL DICTIONARY

- Alphabet definition
- Symbol definition
- Sections of entries

English monodix

A (simplified) monodix looks like this:

```
<?xml version="1.0" encoding="UTF-8"?>
<dictionary>
 <alphabet>abcdefghijklmnopqrstuvwxyz</alphabet>
 <sdefs>
  <sdef n="n" c="noun"/>
  <sdef n="sg" c="singular"/>
  <sdef n="pl" c="plural"/>
 </sdefs>
 <pardefs>
  <pardef n="book__n">
   <e><p><l></l><r><s n="n"/><s n="sg"/></r></p></e>
   <e><p><l>s</l><r><s n="n"/><s n="pl"/></r></p></e>
  </pardef>
 </pardefs>
 <section id="id" type="standard">
  <e><i>dream</i><par n="book__n"/></e>
  <e><i>hug</i><par n="book__n"/></e>
 </section>
</dictionary>
```

Bilingual dictionary

A simplified bidix looks like:

```
<?xml version="1.0" encoding="UTF-8"?>
<dictionary>
 <alphabet>abcdefghijklmnopqrstuvwxyz</alphabet>
 <sdefs>
  <sdef n="n" c="noun"/>
  <sdef n="sg" c="singular"/>
  <sdef n="pl" c="plural"/>
 </sdefs>
 <section id="id" type="standard">
  <e><l>dream</l><s n="n"/><r>sueño</r><s n="n"/><s
n="m"/></e>
  <e><i>hug</i><s n="n"/><r>abrazo</r><s n="n"/><s
n="m"/></e>
 </section>
</dictionary>
```

Spanish dictionary

Another monodix:

```
<?xml version="1.0" encoding="UTF-8"?>
<dictionary>
 <alphabet>abcdefghijklmnopqrstuvwxyz</alphabet>
 <sdefs>
  <sdef n="n" c="noun"/>
  <sdef n="sg" c="singular"/>
  <sdef n="pl" c="plural"/>
  <sdef n="m" c="masculino"/>
 </sdefs>
 <pardefs>
  <pardef n="libro__n">
   <e><p><l></l><r><s n="n"/><s n="m"/><s n="sg"/></r></p></e>
   <e><p><l>s</l><r><s n="n"/><s n="m"/><s n="pl"/></r></p></e>
  </pardef>
 </pardefs>
 <section id="id" type="standard">
  <e><i>sueño</i><par n="libro__n"/></e>
  <e><i>abrazo</i><par n="libro__n"/></e>
 </section>
</dictionary>
```

Practice 3: Paradigm association tool (to increase HBS dixes) [40 min.]

Hvala!