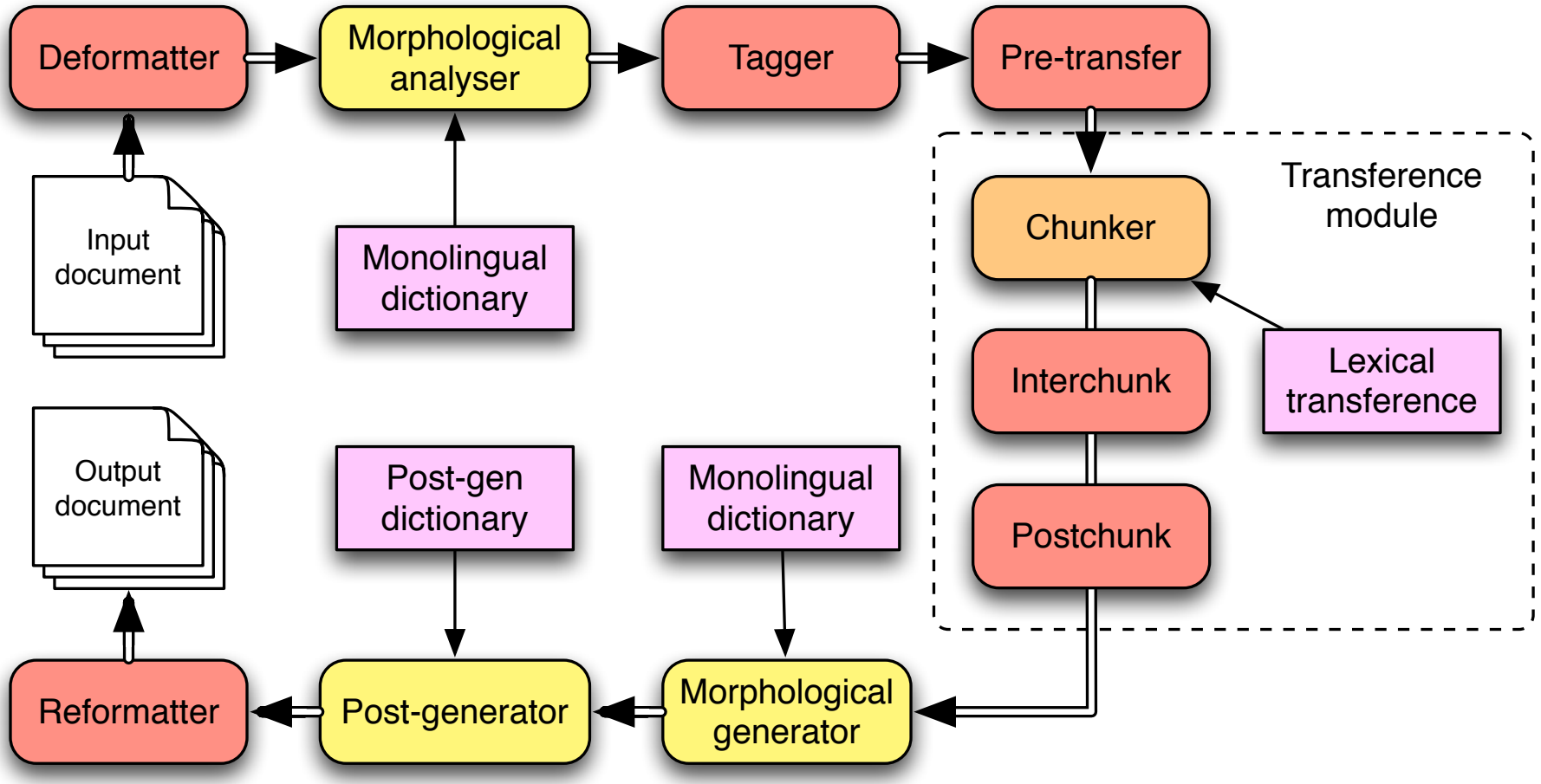
Workshop on the Apertium free/open-source machine translation platform: basics on how to control the engine through linguistics

5th/6th November 2014

Session 4: lexical transfer [15 min.]

Transfer stage /1



Transfer stage /2

The transfer module is where the *magic* happens: the intermediate representation in source language (SL) is converted into an intermediate representation in target language (TL).

Transfer in Apertium consists of two submodules:

- Lexical transfer:
  - selects the most suitable equivalent in TL for a SL word;
  - marks some lexical features which will be used by the structural transfer.
- Structural transfer: performs syntactic operations involving groups of words

Lexical transfer

The lexical transfer module reads each SL lexical form and delivers the corresponding TL lexical form by looking it up in a bilingual dictionary.

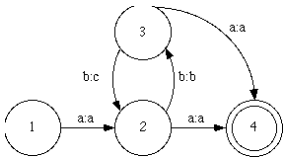
Bilingual dictionary

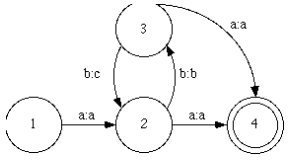
- No surface forms in this stage: input and output are lexical forms consisting of lemma, part-of-speech and inflection information.
- The dictionary contains a list of equivalent lexical forms.
- A single bilingual dictionary is used for both directions of translation.
- XML syntax similar (but simpler) to monolingual dictionaries.
- Paradigms are usually not necessary.

Translation equivalents /1

A simple task... apparently:

transducteur ⟷ transductor






```
[fr]                                [es]
transducteur<n><m><s>    ⟷    transductor<n><m><s>
transducteur<n><m><pl>   ⟷    transductor<n><m><pl>
```

Translation equivalents /2

A shorter representation

Only lemma and part-of-speech are mandatory if the rest of tags do not change:

`transducteur<n>` $\longleftrightarrow$ `transductor<n>`

XML encoding in the bilingual dictionary

```
<e><p>
  <l>transducteur<s="n"></l>
  <r>transductor<s="n"></r>
</p></e>
```

These can be used for `fr` $\rightarrow$ `es` , and `es` $\rightarrow$ `fr`.

Change of gender

Only the tags until the last change need to be indicated:

vallée<n><f> ⟷ valle<n><m>

XML encoding in the bilingual dictionary

```
<e><p>
  <l>vallée<s="n"><s="f"></l>
  <r>valle<s="n"><s="m"></r>
</p></e>
```

Lexical ambiguity

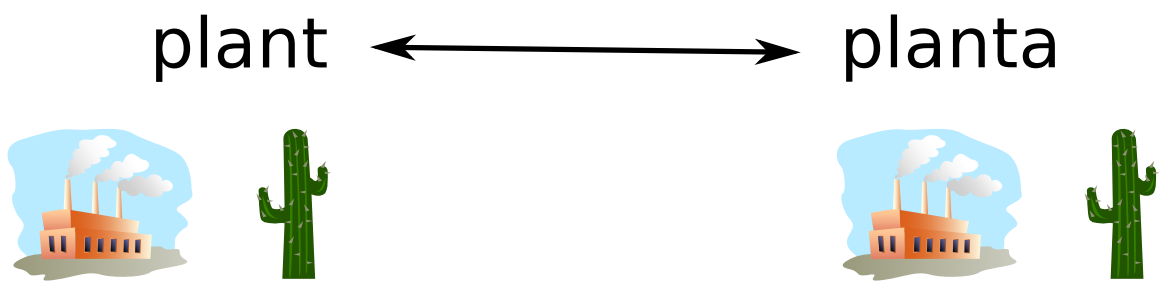
Real life is a bit more complex...

Homography

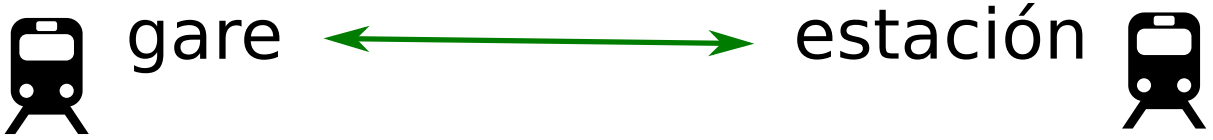
English *book* (`noun` or `verb`) translates into French *livre* (`noun`) or *réserver* (`verb`).

Polysemy

English *bank* (`noun`) translates into Spanish *banco* or *ribera*.
Free-rides do not pose any problem: English *plant* is *planta* in Spanish both the living organism or a kind of factory/installation.



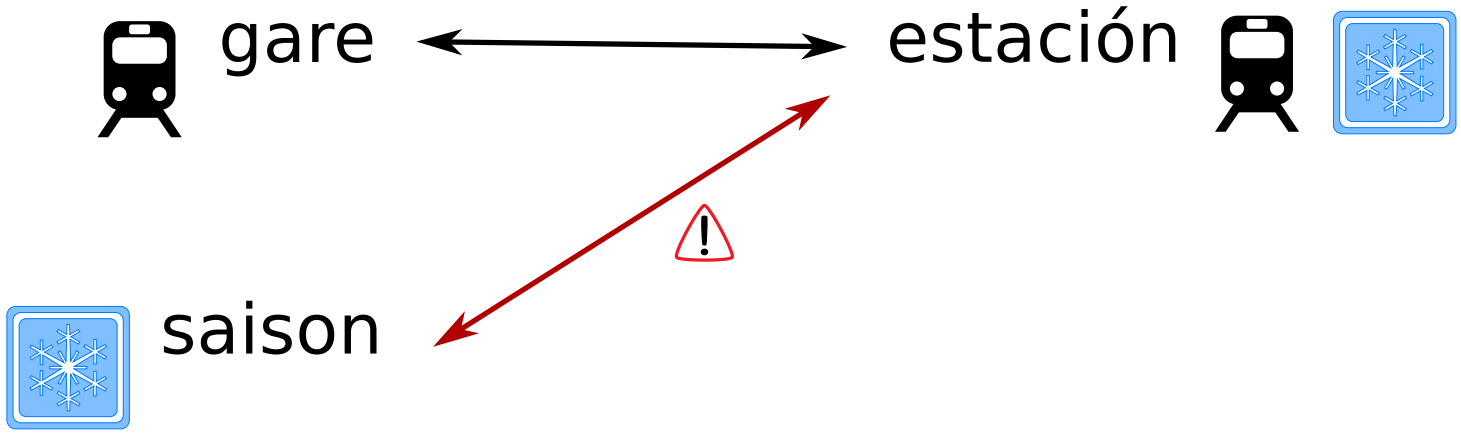
Adding entries to the dictionary /1



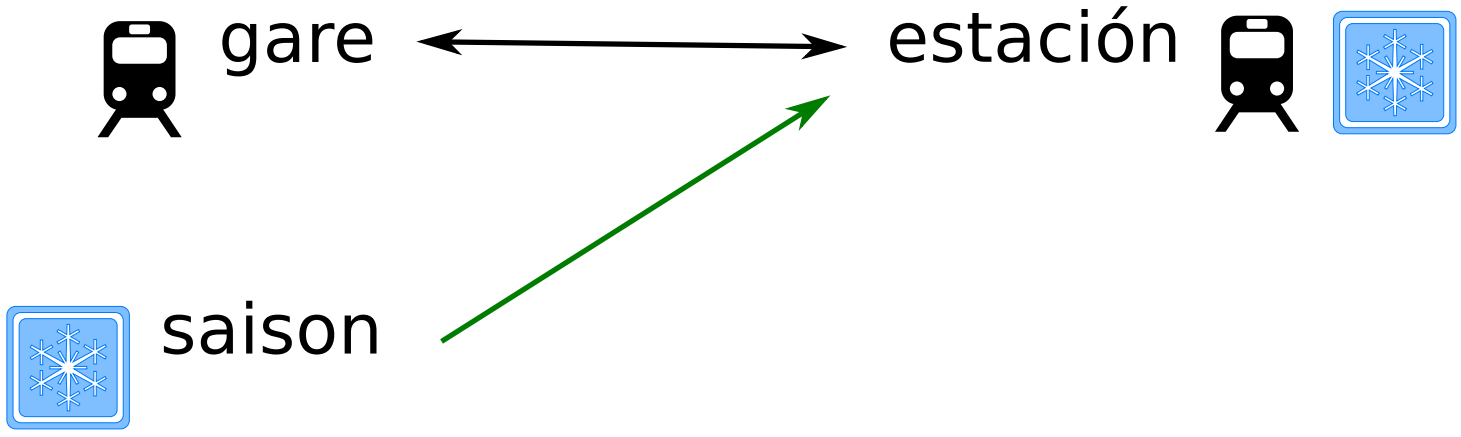
Adding entries to the dictionary /2

```
<e><p>
  <l>gare<s n="n"/></l> <r>estación<s n="n"/></r>
</p></e>
```

Adding entries to the dictionary /3



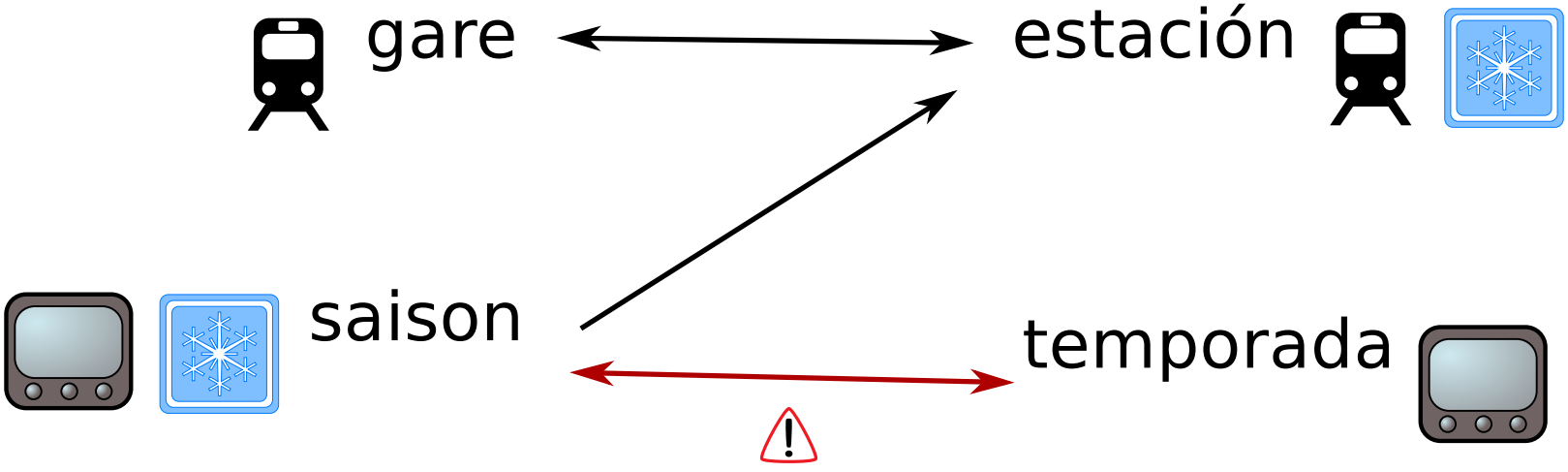
Adding entries to the dictionary /4



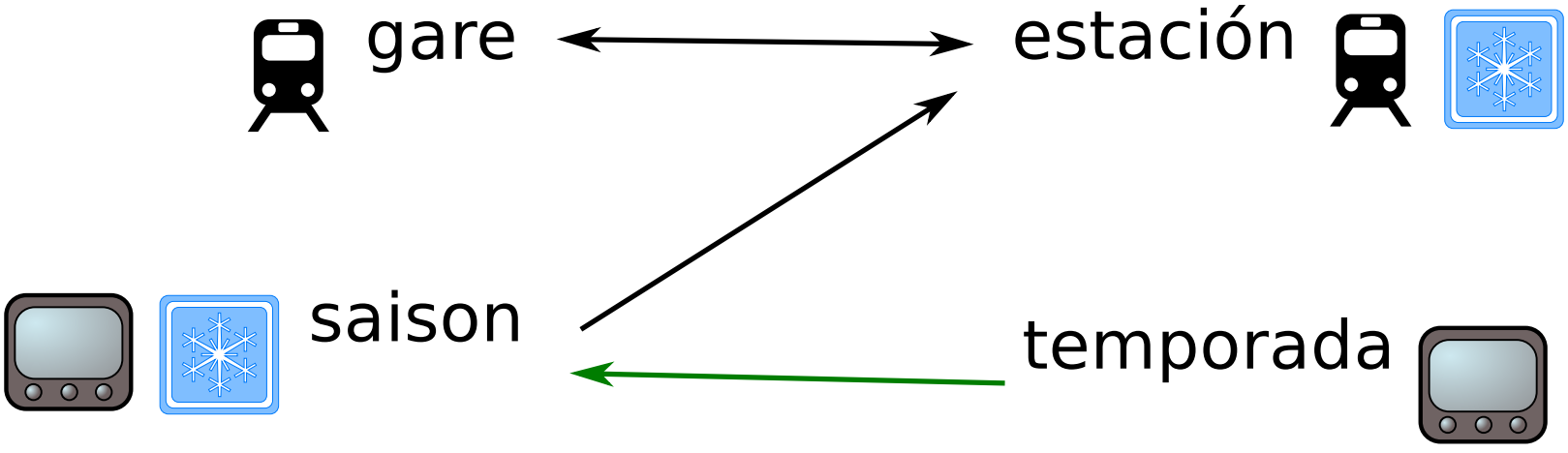
Adding entries to the dictionary /5

```
<e><p>
  <l>gare<s n="n"/></l> <r>estación<s n="n"/></r>
</p></e>
<e r="LR"><p>
  <l>saison<s n="n"/></l> <r>estación<s n="n"/></r>
</p></e>
```

Adding entries to the dictionary /6



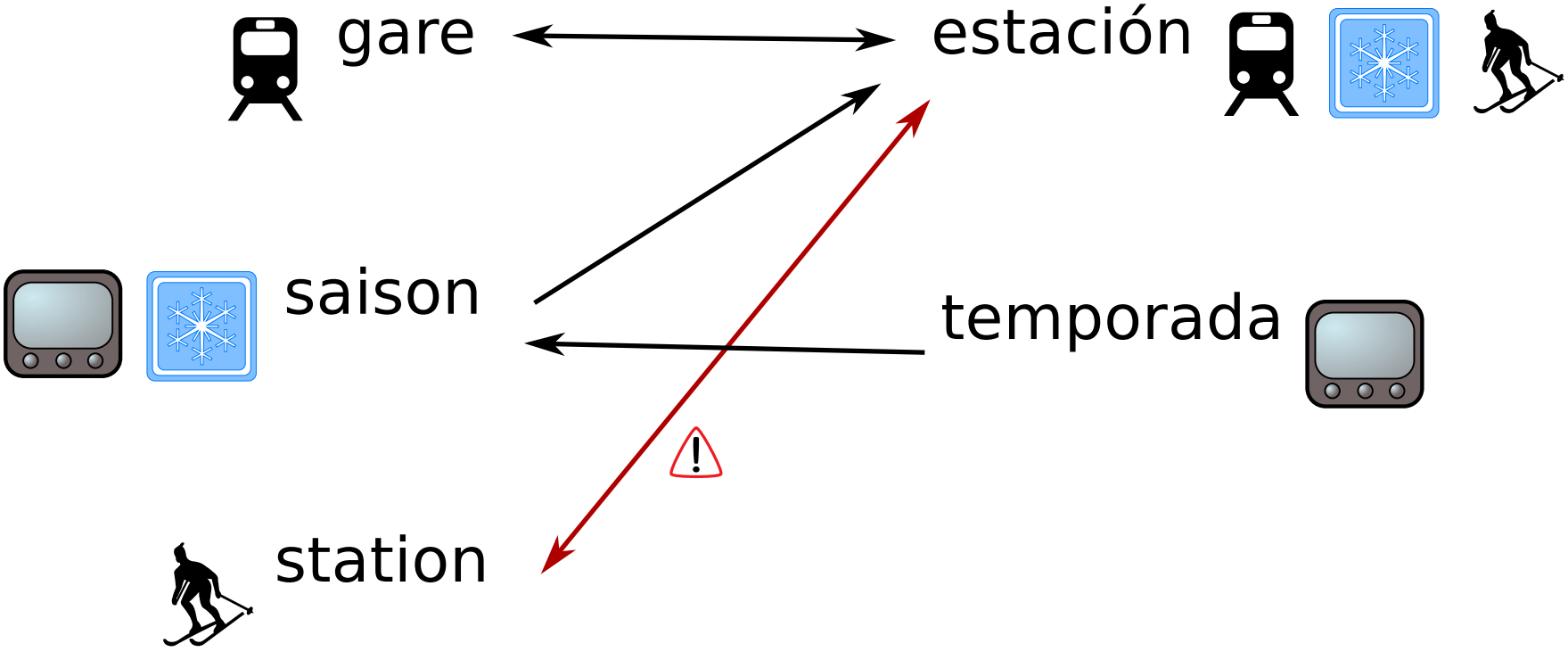
Adding entries to the dictionary /7



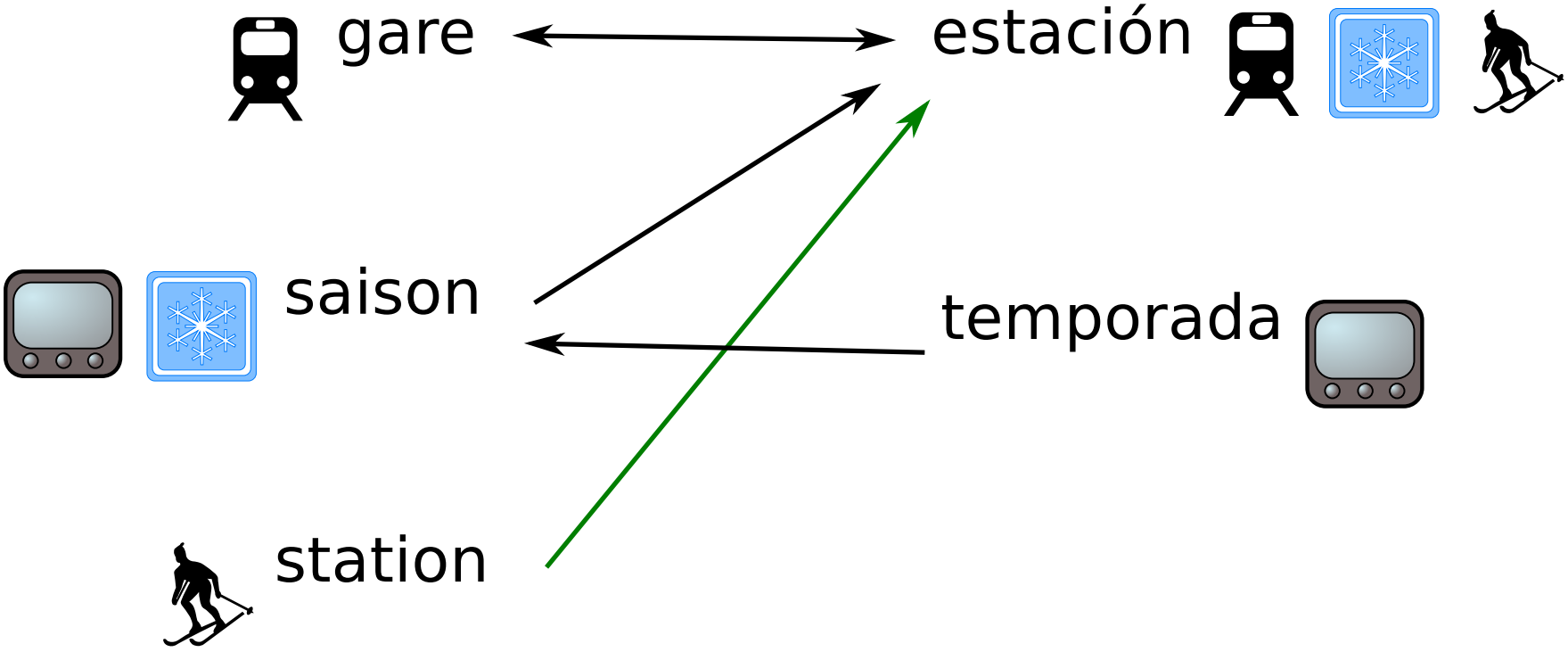
Adding entries to the dictionary /8

```
<e><p>
  <l>gare<s n="n"/></l> <r>estación<s n="n"/></r>
</p></e>
<e r="LR"><p>
  <l>saison<s n="n"/></l> <r>estación<s n="n"/></r>
</p></e>
<e r="RL"><p>
  <l>saison<s n="n"/></l> <r>temporada<s n="n"/></r>
</p></e>
```

Adding entries to the dictionary /9



Adding entries to the dictionary /10



Adding entries to the dictionary /11

```
<e><p>
  <l>gare<s n="n"/></l> <r>estación<s n="n"/></r>
</p></e>
<e r="LR"><p>
  <l>saison<s n="n"/></l> <r>estación<s n="n"/></r>
</p></e>
<e r="RL"><p>
  <l>saison<s n="n"/></l> <r>temporada<s n="n"/></r>
</p></e>
<e r="LR"><p>
  <l>station<s n="n"/></l> <r>estación<s n="n"/></r>
</p></e>
```

Disambiguation of polysemy

We may cope with lexical selection of polysemous terms by using multiwords:

```
gare<n>                          ⟷   estación<n>
station <g>de ski</g><n>         ⟷   estación <g>de esquí</g><n>
```

- Apertium also includes an optional module for lexical selection.

Marking lexical features for the structural transfer /1

- The lexical transfer also marks some lexical features which will be used by the structural transfer.

- For instance, a noun with the same surface form for its two genders.
- Spanish monolingual dictionary:

  estudiante $\longrightarrow$ `estudiante<n><mf><sg>`
  estudiantes $\longrightarrow$ `estudiante<n><mf><pl>`

- The structural transfer will choose the gender by looking at the surrounding context.
- The lexical transfer simply marks this issue with the tag `GD`.

- Similar things hold for number (`ND`).

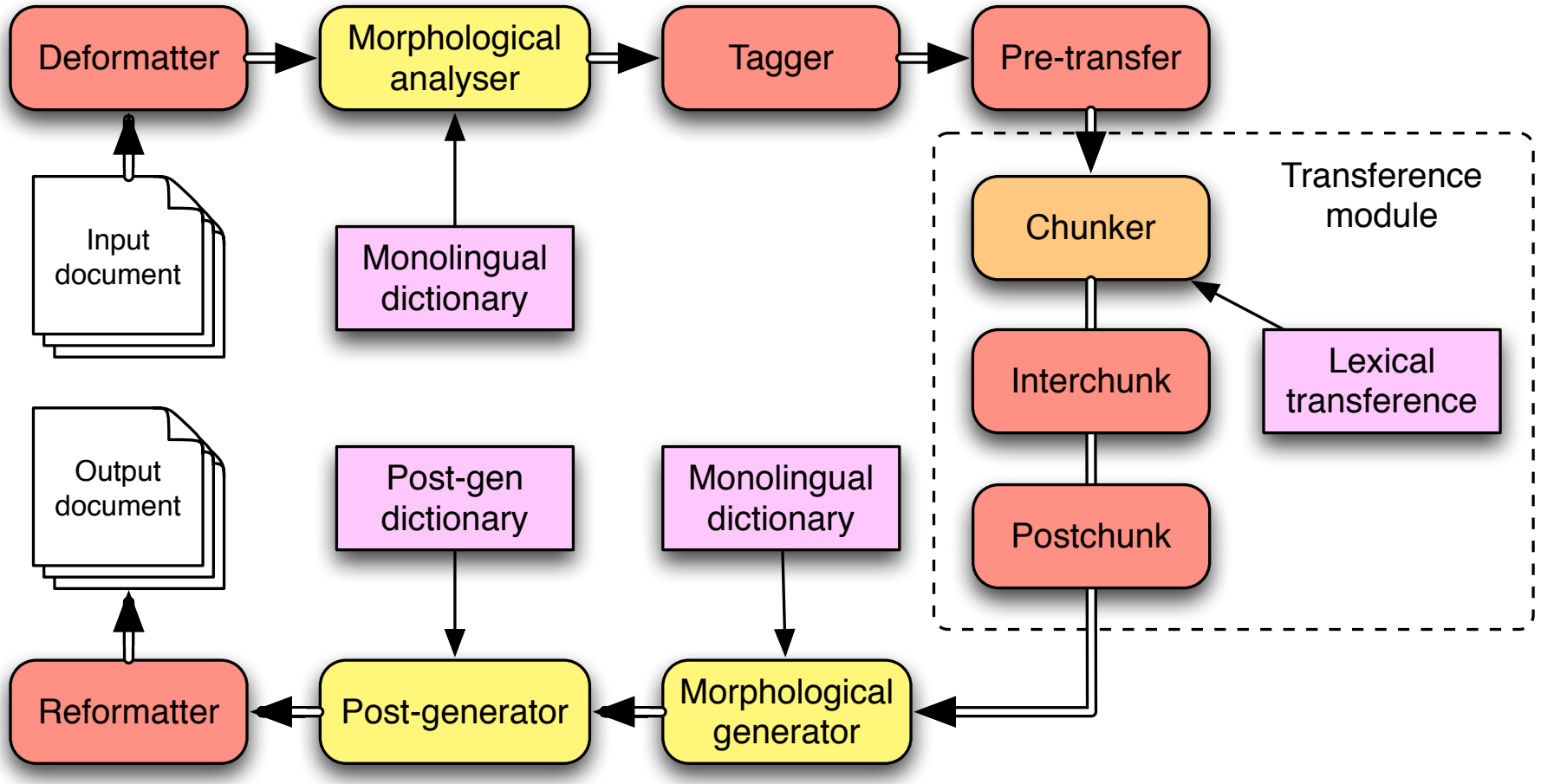
Marking lexical features for the structural transfer /2

```
<e r="LR"><p>
  <l>étudiant<s n="n"/><s n="m"/></l>
  <r>estudiante<s n="n"/><s n="mf"/></r>
</p></e>

<e r="LR"><p>
  <l>étudiant<s n="n"/><s n="f"/></l>
  <r>estudiante<s n="n"/><s n="mf"/></r>
</p></e>

<e r="RL"><p>
  <l>étudiant<s n="n"/><s n="GD"/></l>
  <r>estudiante<s n="n"/><s n="mf"/></r>
</p></e>
```

Practice 4: Translation equivalents [40 min.]

Session 5: Morphological disambiguation [15 min.]

Part-of-speech tagger: where are we?



Lexical ambiguity and part-of-speech tagging /1

Lexical ambiguity

A surface form with more than one possible morphological analysis

Ex. `[en]` *book* (`noun` or `verb`)

→ `[fr]` *livre* (`noun`)

→ `[fr]` *réserver* (`verb`)

Lexical ambiguity and part-of-speech tagging /1

Lexical ambiguity

A surface form with more than one possible morphological analysis

Ex. `[en]` *book* (`noun` or `verb`)

→ `[fr]` *livre* (`noun`)

→ `[fr]` *réserver* (`verb`)

This is not polysemy!

A lemma and part-of-speech tag that have several meanings

Ex. `[en]` *bank* (`noun`)

→ `[es]` *banco* (institution that provides financial services)

→ `[es]` *ribera* (slope of land adjoining a river)

Lexical ambiguity and part-of-speech tagging /2

Ambiguity between part-of-speech:

I (`acr`)
work (`vblex.pres` or `n.sg`)

Ambiguity within part-of-speech:

I (`prn`)
see (`vblex.inf` or `vblex.pres`

Statistical disambiguation /1

Statistics about the context in which each tag appears help to solve the part-of-speech ambiguity

These statistics are collected

- from hand-tagged texts (more accurate), or
- from untagged texts (less accurate)

Tagged text

I	(`prn.subj.p1.pl`)
see	(`vblex.pres`)
my	(`det.pos.1.sg`)
screen	(`n.sg`)

Statistical disambiguation /2

Apertium statistical tagger is based on first-order hidden Markov models

It chooses the combination of tags with the highest probability:

Book (verb) a (prep) calm (adj) room (noun)
Book (verb) a (prep) calm (vblex) room (noun)
Book (verb) a (prep) calm (noun) room (noun)
Book (noun) a (prep) calm (adj) room (noun)
Book (noun) a (prep) calm (vblex) room (noun)
Book (noun) a (prep) calm (noun) room (noun)

Practice 5: Annotating a corpus [20 min.]

Tagset definition /1

To alleviate the problem of data sparseness the sequences of morphological tags are grouped into coarse tags (called Labels)

Sequence of tags	Coarse tag
`noun.m.sg`	NOUN
...	...
`noun.f.pl`	NOUN
`verb.pres.1p.sg`	VERB.PRESENT
...	...
`verb.pres.3p.pl`	VERB.PRESENT
`prn.1p.sg`	PRONOUN
`prn.2p.sg`	PRONOUN
`prn.3p.sg`	PRONOUN.3P.SG
...	...
`prn.3p.pl`	PRONOUN

Tagset definition /2

How to design a tagset:

Rules of thumb

- Group sequences of tags having the same syntactic role and appearing in the same contexts under the same coarse tag
- Do not group under the same coarse tag those sequences of tags among which the disambiguator needs to distinguish

Starting with a tagset borrowed from a similar language might help

Tagset definition /3

Example of tagset:

```
<?xml version="1.0" encoding="iso-8859-1"?>
<tagger name="English">
 <tagset>
  ...
  <def-label name="ADJ">
   <tags-item tags="adj"/>
   <tags-item tags="adj.comp"/>
   <tags-item tags="adj.sup"/>
   <tags-item tags="adj.sint"/>
   <tags-item tags="adj.sint.*"/>
  </def-label>
  <def-label name="PREP" closed="true">
   <tags-item tags="pr"/>
  </def-label>
  ...
 </tagset>
  ...
</tagger>
```

Rule-based disambiguation /1

Statistical disambiguator

- Guarantees that a sentences is completely disambiguated
- May make mistakes because it uses a limited context window

Constraint grammar rules [optional]

- Do not guarantee that a sentences is always completely disambiguated
  - They must be applied before the statistical disambiguator
- Can reduce (or even solve) the ambiguity
- Can use a variable-length context window

Rule-based disambiguation /2

Este (prn.dem **and** *det.dem) día (n.m.sg)* (Spanish)

- `This (det.dem) day (n.sg)` (English)
- `This one (prn.dem) day (n.sg)` (English)

Example of constraint grammar rule:

```
LIST DET-DEM = (det dem);
LIST PRON-DEM = (prn dem);

REMOVE PRON-DEM IF (0 PRON-DEM) (0 DET-DEM) (1C N);
```

Remove a reading of demonstrative pronoun IF

- current word can be a demonstrative pronoun, AND
- current word can also be a demonstrative determiner, AND
- first word to the right can ONLY be a noun

Practice 6: Taking a look to a tagset [20 min.]

Hvala!