Workshop on the Apertium free/open-source machine translation platform

Gema Ramírez-Sánchez
Prompsit Language Engineering, S.L.
www.prompsit.com
Campus UMH. Edifici Quórum III.
Av. de la Universitat, s/n. 03203. Elx (Alacant). Spain

5/6 November 2014. Zagreb.

Contents

About this workshop

The materials of this workshop on the "Apertium free/open-source machine translation platform" have been created by Prompsit Language Engineering, S.L., as part of the Abu-MaTran (Automatic Building of Machine Translation) project[1] funded by European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number PIAP-GA-2012-324414. Special thanks to Nikola Ljubešić for his help.

[1] www.abumatran.eu

Overview

This guide is intended to be your best friend during this workshop for the hands-on and hands-up practical exercises you'll be working on to meet the following objectives:

1. Have a general idea of how machine translation works: you will test some translators and understand what's going on behind them
2. Understand how Apertium works: you will see and touch the inner parts of Apertium, module by module
3. Understand Apertium monolingual dictionaries: you will help us improving Apertium monolingual dictionaries
4. Understand why ambiguity is our main problem and how do we cope with it: you will explore ambiguous sentences, annotate corpora and see some rules for disambiguation
5. Understand Apertium bilingual dictionaries: and help us improving Apertium bilingual dictionaries
6. Understand Apertium from the developer point of view: you will work by frequency estimates, defined tasks and corpora-driven knowledge

For every section there will be a basic introduction to the topic before putting our hands on it.

1 Introduction to machine translation

1.1 Rule-based vs corpora-based approaches

We've reviewed together the basics about machine machine: definition, main uses and types of machine translation. Before going on, let's take a look to some machine translated texts.

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

In the next table you are presented with 2 texts translated by 4 different machine translation systems from Croatian into English. Translations have been sorted randomly for each one of them.

For each text:

- try to guess whether the translations are from a SMT or a RBMT and motivate your answer indicating two reasons for your classification
- choose a best candidate for assimilation purposes (that is, the one that would be better for getting the meaning of the original sentence)
- choose a best candidate for dissemination purposes, more specifically, for post-editing (that is, the one that would be more useful to produce an adequate translation by applying the minimum number of changes to it)

Text: 2. The Raveonnetes!

Croatian	**Prvi put u Zagreb na samostalan koncert stižu The Raveonettes, danski indie rock dvojac u kojem su basistica i pjevačica Sharin Foo i gitarist Sune Rose Wagner.**
MT1	That's the first time in Zagreb to act independently concert are The Raveonettes, Danish indie rock pair in which they basistica and singer Sharin Foo and guitarist Sune Rose Wagner.
MT2	First time in Zagreb on solo concert arrive The Raveonettes, Danish indie rock dvojac in which are basistica and singer Sharin Foo and gitarist Sune Dew Wagner.
MT3	For the first time in Zagreb on standalone concert coming The Raveonettes, a Danish indie rock in which the bass player and singer Sharin Foo and guitarist Sune Rose Wagner.
MT4	First into a Zagreb at an substantive concert stižu The Raveonettes , Danish indie rock brace into a which have been basistica plus songstress Sharin Foo plus guitar Sune Scarlet-rash Wagner.

	RBMT or SMT	**Reason**	**Best for assimilation**	**Best for dissemination**
MT1				
MT2				
MT3				
MT4				

Text: 3. Family Fazlinović!

Croatian	**Nova, osma sezona kultne serije "Lud, zbunjen, normalan" kreće od ponedjeljka na Novoj TV! Ne propustite nove zgode u legendarnoj obitelji Fazlinović!**
MT1	Nova, the eighth seasons screen cult series "Lud, zbunjen, normalan" ranges from Monday on Nova TV! There miss new convenience in legendary family Fazlinović!
MT2	Money , eight high season cult serial " witless , confused , unexceptional kree with Monday at an Learner TV! Does not let off nove time into a legendary families Fazlinovi!
MT3	New, eighth season kultne series "Lud, zbunjen, normal" moves from Monday on New TV! Not propustite new zgode in legendary family Fazlinović!
MT4	The new, eighth season of the cult series "Lud, normal" ranges from Monday on Nova TV! Do not miss the next game in the legendary family Fazlinović!

	RBMT or SMT	Reason	Best for assimilation	Best for dissemination
MT1				
MT2				
MT3				
MT4				

2 Apertium

2.1 How does Apertium work?

We've seen that Apertium is an engine with a modular architecture. Each module performs an action to the input it receives from the precedent module. Let's see how the ouput of each module looks like.

Task 4. Taking a look to Apertium with apertium-viewer [30 min.]

Apertium-viewer[2] is a tool that shows the translation process in Apertium module by module. To access it:

- Open a browser and copy/paste the following URL or click on it: `http://tinyurl.com/nwj97bl`
- A menu to *Open* or *Save* a file will appear. Let's just open it. Click on `Accept`.
- A security warning window will pop out. Click on `Run`.
- A confirmation window will then pop out. Click on `Yes`.
- A final reconfirmation window will pop out. Click again on `Yes`.

You'll finally see an interface like the one shown below.

Please, follow these instructions:

- First of all, make sure you have the option `Online` (and not `Local`) on the right top of the screen selected. Otherwise click on `Online` and wait for some seconds.

[2]Further reading: `http://wiki.apertium.org/wiki/Apertium-viewer`

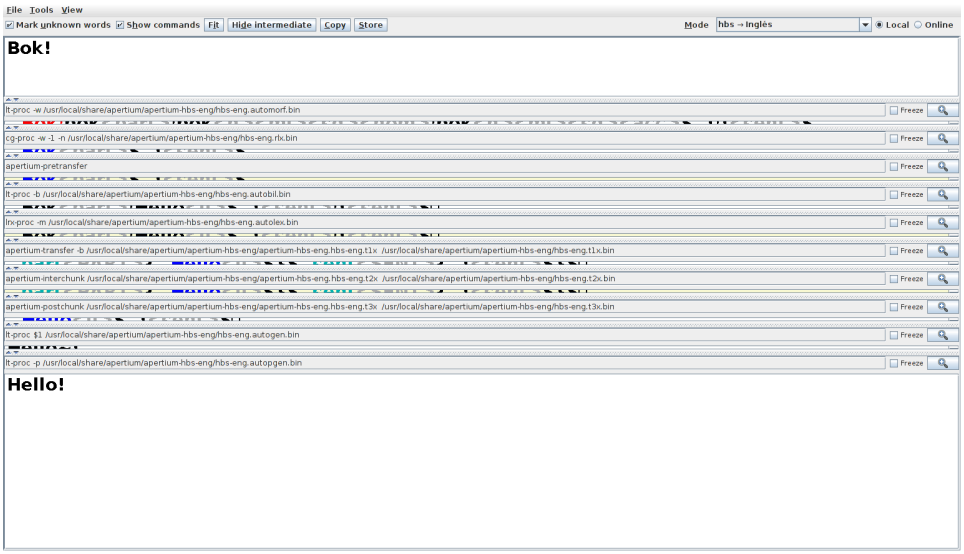

Figure 1: apertium-viewer

- Next to it you have a menu called `Mode` which says `SELECT A MODE`. In Apertium language a mode is a translation direction. Open the menu and select mode `inglés-español`. Wait for some seconds, it takes a bit to load all dictionaries...
- For better user experience, let's change the font of the user interface. Go to the left menu and click on `View`. Go to `Font` and set it to `Dialog-Bold-28`. Click on `Done`.

We are ready for testing! Let's start:

- Write a simple sentence: *Hello world.*
- You'll see the translation appearing as you type and the final translation at the end: *Hola mundo.*
- Easy, isn't it? But even the most simple sentences can be ambiguous. That's why before you type the final dot, your translation is *Hola mundial* and not *Hola mundo.* Note that *world* can be and *noun* or an *adjective*.
- So, how does Apertium know what to do? If you click on any of the bars appearing in the screen and you swipe it down you'll start to see all intermediate modules output in Apertium.

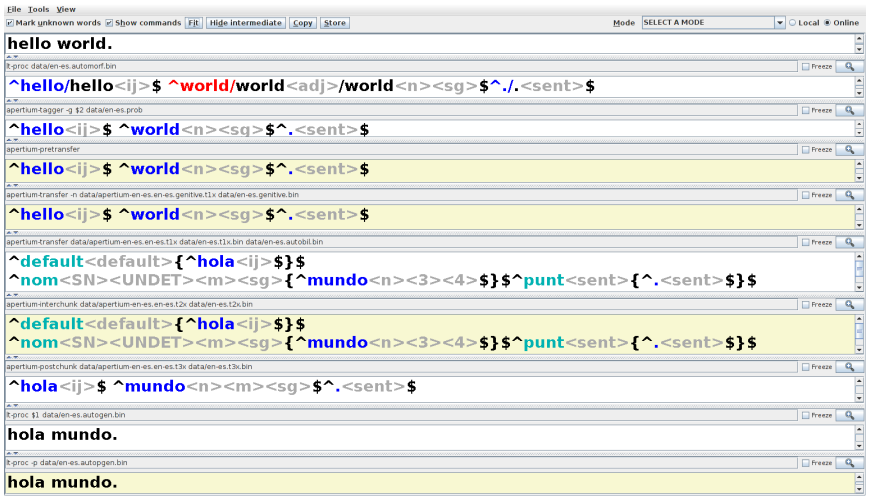

Figure 2: View of apertium-viewer modules

- Don't be scared, here is the reading for those strange name commands. You'll also find useful to open in a separate tab the wiki page which specifies how part-of-speech and other morphological features[3] are denoted in Apertium:

    1. **Morphological analyser output**: *lt-proc data/en-es.automorf.bin*
    2. **Part-of-speech tagger**: *apertium-tagger -g $2 data/en-es.prob*
    3. **Multiple-word unit handler**: *apertium-pretransfer*
    4. **Saxon genitive handler**: *apertium-transfer -n data/apertium-en-es.en-es.genitive.t1x data/en-es.genitive.bin*
    5. **First transfer step**: *apertium-transfer data/apertium-en-es.en-es.t1x data/en-es.t1x.bin data/en-es.autobil.bin*)
    6. **Second transfer step**: *apertium-interchunk data/apertium-en-es.en-es.t2x data/en-es.t2x.bin*
    7. **Third transfer step**: *apertium-postchunk data/apertium-en-es.en-es.t3x data/en-es.t3x.bin*
    8. **Morphological generator output**: *lt-proc $1 data/en-es.autogen.bin*
    9. **Post-generator output**: *lt-proc -p data/en-es.autopgen.bin*

[3] `http://wiki.apertium.org/wiki/List_of_symbols`

- Let's take a look to some sentences.

  Hello world travelers

  - Inspect module 2 to see ambiguity.
  - Inspect module 5 to see a rule *ADJECTIVE + NOUN = NOUN + ADJECTIVE*. Note that we don't know the gender yet (<GD>).
  - Inspect module 6 to see how agreement between *NOUN + ADJECTIVE* is propagated.
  - Inspect module 7 to see the sequence of lexical forms in the target language.
  - Inspect module 9 for final translation!!!

  I saw Lily's shoes

  - Inspect module 2 to see ambiguity: all words are ambiguous!!!
  - Inspect module 3 to see how ambiguity was solved: well done in this case...
  - Inspect modules 5 and 6 comparatively to see: that the pronoun disappears because it is not needed in Spanish how the Saxon genitive rule is applied: *PROPER NOUN+'S + NOUN = NOUN + DE + PROPER NOUN*
  - Inspect module 7 to see the sequence of lexical forms in the target language.
  - Inspect module 9 for final translation!!!

  In the end, I'll take the soup of the day

  - Inspect module 2 to see a multiple-word unit (in the end)
  - Inspect module 8 to see the output of the morphological generator: note the marks for the postgenerator ( ).
  - Inspect module 9 for final translation where a contraction applied *de + el = del*

As you have seen, when a user clicks on the `Translate` button of a rule-based machine translation system, a number of linguistic-motivated processings are applied before delivering the machine translated output. But even if the information is accurate, rule-based machine translation is not fully capable of solving the four big problems already reviewed in this session: analysis, synthesis, transfer and description.

3 Dictionaries

3.1 Monolingual entries

Apertium monolingual dictionaries contain information about words needed through all the translation process. Correspondences between surface forms (toe, toes) and lexical forms (toe, singular noun and toe, plural noun) are defined in Apertium's dictionaries in a synthetic way: by associating words to an inflection paradigm.

To ease this task, we've created a user interface to work with nouns, verbs and adjectives which are not inside the dictionaries yet And now we need your help to choose the correct paradigm for them.

Let's think about this task as a real situation. Take a look a this text translated from Serbian into Croatian.[4]

```
*MIROSLAV Raduljica, CENTAR REPREZENTACIJE SRBIJE:

Do *juče sam bio bradati majmun, a sada sam car!

*Miroslav Raduljica se prije nekoliko nedjelja *otisnuo
u novu avanturu u Kinu, a srpski centar koji trenutno
brani boje *Šandogana, ne krije da mu je ovo *leto bilo
jedno od najzanimljivijih.

Raduljica trenutno brani boje kineskoga *Šandogana

Do srebra na *SP u *Španiji, Raduljica je bio poznat
kao super *talentovani centar, ali na koga se ne može
uvijek *računati, pa ga je tako i nekadašnji *selektor
'orlova' Duda Ivković *precrtao sa spiska.
Sada je situacija potpuno drukčija:

- Vrlo mi je zanimljivo kako sam sad car, bog, legenda,
a do *juče sam bio *istetovirani bradati majmun i splavar.
Pa, ja sam isti taj Raduljica, koji sam bio i 2010.
Dobro, malo sam unaprijeđen što se tiče karaktera, stabilniji
sam, ali sam potpuno isti momak, istih *rezonovanja,
iste ličnosti i percepcije - rekao je Raduljica u intervjuu
za *novembarsko izdanje srpskoga '*Eskvajera'.
```

[4] Text from the Serbian `www.alo.rs` portal at at `http://tinyurl.com/ob4yrbc`

There are still many problems on it: some words are not in our dictionaries (the ones preceded with a star such as **juče*) and some of them should have a different translation (which should be *jučer* in Croatian). After this session and the one related to bilingual dictionaries, this text should look much better...

So, let's start working for this purpose.

Task 5. Paradigm association tool [40 min.]

Open a browser and navigate to `http://paradigm.abumatran.eu`. Log in with the user/password corresponding to you surname without diacritics and you'll be facing the Overview tab of this tool as shown below.

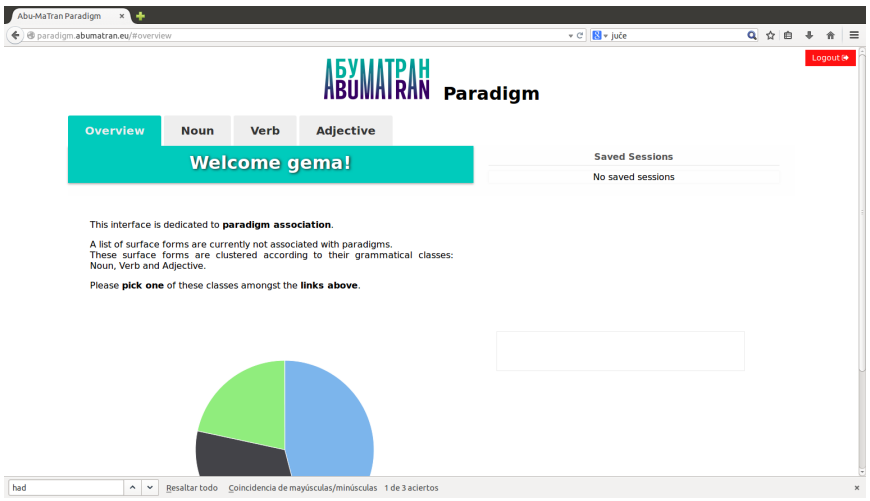

Figure 3: Paradigm association tool overview

In this tab you'll find now just a description of the tool and some stats about tasks already completed to be associated with a paradigm in Serbian. Later on, you'll find your completed sessions (every 10 words for a given category) to be able to review them.

To get started, please, go to tab *Noun*.

- You'll see a word below the tab name. This is the surface form of the word you will be working on. Below, a set of probable paradigms is shown.
- For each paradigm we show 4 things: the *lemma* or base form that the surface form could have according to this paradigm, the *paradigm name* as in Apertium dictionaries, all the *surface forms* this paradigm would generate, and some *morphological information*.

- The goal is to choose the correct paradigm for the surface form given. Once you've clicked on it, you can go to the next surface form by clicking on the *Next* button. Please, select the first paradigm that fits you in case of doubt.
- Note that there is a menu – *Change Category* – in the upper-right side of the interface, next to the surface form, that will allow you to re-assign the surface form you are working on to another category if needed.
- Note also that regarding the morphological information provided for each entrance, you can identify 4/5 different sources of information in this order:
    1. Main category: Nc denotes *Noun, common*
    2. Gender: n denotes *neuter*, m *masculine*, f *feminine*
    3. Number: s denotes *singular*, p *plural*
    4. Case: n denotes *nominative*, a *accusative*, v *vocative*, g *genitive*, d *dative*, l *locative*, i *instrumental*
    5. y denotes *animacy*
- When you complete 10 entries you'll a session will be saved for you and you will be able to access it from the Overview page.

Once you've completed the 10 noun entries, please go to tab *Verb*.

- This tab is very similar to the *Noun* tab. The only specifics of this tab are the check boxes next to the – *Change Category* – menu. Using them you you can indicate additional information for a verb besides the paradigm to fully cover the information we need for a verb that can be, for example, transitive and intransitive. We don't need you to provide this information for the purposes of this workshop. Just do it if you feel like.
- Regarding the morphological information, you should be able to read it as follows:
    1. Main category: Vm denotes *Verb main*
    2. Tense: n denotes*infinitive* , m *imperative*, a *aorist*, r *present*, e *imperfect*, f *future*, p *participle*
    3. Person: 1 denotes *first person*, 2 *second person*, 3 *third person*
    4. Number: s or -s denotes *singular*, p or -p *plural*
    5. Gender: m *masculine*, f *feminine*

- Note that all forms beginning by *App* and *Rr* are for the adjectives and adverb forms that can be derived from the verb.
- Note also that the name of the paradigm gives you information about transitivity (tv - transitive, iv - intransitive) and aspect (perf - perfective, imperf - imperfective).

Once you've completed the 10 Verb entries, please go to tab *Adjective*.

- This tab is very similar to the *Noun* and *Verb* tabs. In this case, the check boxes next to the – *Change Category* – menu will help us know whether the adjective has a comparative and superlative form or if the form contains the *yat* variant. Again, we don't need you to provide this information for the purposes of this workshop. Just do it if you feel like.
- Regarding the morphological information, for adjectives should read as follows:
  1. Main category: Agp denotes *Adjective general positive*
  2. Gender: n denotes *neuter*, m *masculine*, f *feminine*
  3. Number: s denotes *singular*, p *plural*
  4. Case: n denotes *nominative*, a *accusative*, v *vocative*, g *genitive*, d *dative*, l *locative*, i *instrumental*
  5. y denotes *animacy*

If you have completed 10 forms for each category, congratulations, you helped the coverage of Apertium HBS dictionaries a lot!!!

You can always recheck your work through the *Overview* tab or go on a bit assigning paradigms to your preferred category.

Any ideas for improvement? Let's discuss them together.

3.2 Bilingual entries

To complete the work we started by adding entries to monolingual dictionaries, bilingual equivalents should be defined now. So, where is the user interface? I'm afraid that we still don't have one.

To replace it, we've created a spreadsheet that will help you giving Apertium the info needed to perform lexical transfer: *lemma equivalence*,

translation direction, *part-of-speech* and *changes from source to target* for the rest of morphological information, e.g. gender change.

We will collaboratively define translation equivalents for a bunch of words, some of them for entries added yesterday to the monolingual dictionaries and some other for already existing entries that do not have a translation yet. We will work in our Serbian to Croatian language pair in a list of frequent unknown words created from Serbian corpora. Let's start!

Task 6. Translation equivalents [40 min.]

Open the shared spreadsheet by clicking on `http://tinyurl.com/mk2fgey`. Take a moment to understand it through the examples provided for user `gramirez`:

- column A: contains information about the user
- column B: contains the unknown surface form in Serbian (green)
- column C: is for the lemma in Croatian (red)
- column D: is for the lemma in Serbian (green)
- column E : is for the part-of-speech or main category of the equivalents. Please indicate: *n* - for nouns, *adj* - for adjectives, *vblex* - for verbs, *adv* - for adverbs, *pr* - for prepositions, *num* - for numbers, and *np* for proper names.
- column F: is to indicate if the translation works in both directions or just in one of them. Please, indicate: *yes* - for both directions, *HR-SR* - for only from Croatian to Serbian, *SR-HR* - for only from Serbian to Croatian.
- column G: is a free comment area to clarify or specify information about the entry.

If you scroll down, you'll see blocks of 30 entries and your name assigned to one of them. Please work on the 30 entries to provide the missing information taking the following instructions into account:

- **When lemmas are the same in both languages**: as Croatian and Serbian share a big portion of vocabulary, we don't need translation equivalents for all words, only when there is a difference. When no translation equivalent is needed, just leave the row empty.
  A special case are the words that differ only in the *yat* phoneme. We

A	B	C	D	E	F	G
User	**Surface form in Serbian**	**Lemma in Croatian**	**Lemma in Serbian**	**Part-of-speech**	**Translation apply in both directions**	**Comment (default translations, change in gender, etc.)**
gramirez	voza	vlak	voz	n	yes	
gramirez	kritikuju	kritizirati	kritikovati	vblex	yes	
gramirez	bezbednosti	sigurnost	bezbednost	n	yes	
gramirez	hiljade	tisuća	hiljada	num	yes	
gramirez	Kipar	Cipar	Kipar	np	yes	
gramirez	vazdušna	zračan	vazdušan	adj	yes	
gramirez	tati	ćaća	tata	n	HR-SR	Default translation is tata-tata, but in HR there is also ćaća
gramirez	paradajza	rajčica	paradajz	n	yes	Gender change: f (HR) – mi (SR)
gramirez	paradajza	pomidor	paradajz	n	HR-SR	Default translation is rajčica (HR)- paradajz (SR) but pomidor and paradajz are also possible in HR
gramirez	paradajza	paradajz	paradajz	n	HR-SR	

Figure 4: Bilingual equivalent definition

consider them as different words sharing the same lemma, so in this case, please just indicate it in *Comments* column by writting *YAT*.

- **When multiple translations are possible**: first, recheck the example provided in the spreadsheet for *tomato*). When more than one translation is possible, we will choose the most general and frequent one[5] to work in both directions (Column F set to *yes* and add the other(s) indicating the appropriate translation direction: *HR-SR* - from Croatian to Serbian, *SR-HR* - from Serbian to Croatian.

- **When equivalents have different gender or number**: please indicate it in the Comments column as in the example for *tomato*: Gender change: *f* (HR) – *mi* (SR)

Congrats! You've completed your first bilingual task in Apertium. We should now check whether the lemmas you've provided in Croatian are also in the Croatian monolingual dictionaries and add the ones missed. But this will be not done now as other topics and tasks are waiting for us.

[5]To check frequency and contexts you can use the concordance tool developed by the Natural Language Processing group at the Department of Information Sciences in the Faculty of Humanities and Social Sciences at the University of Zagreb: `http://tinyurl.com/nw96nsy` (thanks Tomaž Erjavec and Nikola Ljubešić!). You'll find there the hrWaC (web-based corpus for Croatian) and srWac (web-based corpus for Serbian) corpora to perform searches.

But let's take a look to our working text from yesterday. If we did things properly, it should look like:

MIROSLAV Raduljica, CENTAR REPREZENTACIJE SRBIJE:

Do **jučer** sam bio bradati majmun, a sada sam car!

Miroslav Raduljica se prije nekoliko **nedjelja** ⋆otisnuo u novu avanturu u Kinu, a srpski centar koji trenutno brani boje **Šandogana**, ne krije da mu je ovo **ljeto** bilo jedno od najzanimljivijih.

Raduljica trenutno brani boje kineskoga **Šandogana**

Do srebra na SP u **Španjolskoj**, **Raduljica** je bio poznat kao super **talentirani** centar, ali na koga se ne može uvijek **računati**, pa ga je tako i nekadašnji **izbornik** 'orlova' Duda Ivković ⋆precrtao sa spiska.
Sada je situacija potpuno drukčija:

– Vrlo mi je zanimljivo kako sam sad car, bog, legenda, a do **jučer** sam bio ⋆istetovirani bradati majmun i splavar. Pa, ja sam isti taj **Raduljica**, koji sam bio i 2010. Dobro, malo sam unaprijeđen što se tiče karaktera, stabilniji sam, ali sam potpuno isti momak, istih ⋆rezonovanja, iste ličnosti i percepcije – rekao je **Raduljica** u intervjuu za **studensko** izdanje srpskoga '⋆Eskvajera'.

We added: *Raduljica*, *Šandogan* and *Miroslav* which appeared 6, 2 and 2 times in the text: not a big deal, they remain the same.

We also added: *juče*, appearing 2 times and translated differently into Croatian (*jučer*) and words appearing only once in this text, but quite frequent in our list of monolingual unknown words: *nedjelja* (as *nedjelja* and not *tjedan*), *leto* (becomes *ljeto*), *talentovan* (becomes *talentiran*), *selektor* (becomes *izbornik*), *novembarski* (becomes *studenski*) and *računati* (remains the same). And, of course, *Španija* becomes *Španjolska*!!

Remaining problems: some unknown words (**otisnuo*, **precrtao*, **istetovirani*, **Eskvajera*, **rezonovanja*) and maybe some rules (next workshop!).

But overall, much better now (for post-editing purposes), isn't it? That's all thanks to you all and to the frequency strategy, of course!

4 Part-of-speech tagger data

One of the most valuable resources to build a disambiguation module for Apertium is an annotated corpus. Just a few language pairs have one because it is expensive to build and rare to find. Compatibility between tagsets makes reusability also difficult.

Last year we developed *Annotatrix*, a tool for annotating corpora according to Apertium dictionaries and to inspect and improve tagsets. We will use it during this practice to annotate a brief corpus and to see a tagset definition file.

4.1 Annotated corpora

Annotating corpora can be really time-consuming but with Apertium we can at least semi-automate the task: only the ambiguous words will need your help (provided that we have accurate dictionaries!). Let's see!

Task 7. Annotate an HBS corpus [25 min.]

We are going to work on a Croatian corpus, just a paragraph to have an idea of how annotators work. Follow these steps:

- Open a browser and go to `http://abumatran.eu:28000/accounts/login/`. Log in with username and password *uzguest*. Once you are logged in, you'll see the main dashboard for Annotatrix.

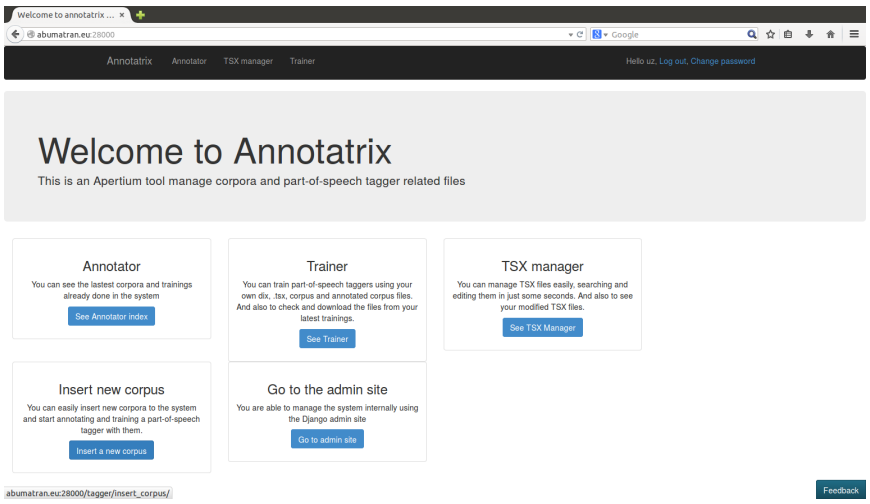

Figure 5: Bilingual equivalent definition

- Click on *Insert a new corpus*, you'll go to an interface to upload or copy/paste a corpus to be annotated.
- Paste the following text[6] (or another) into the text box *Corpus text*:

  ```
  "Kada je fotografija dvoje mladih umotanih u hrvatsku
  i srpsku zastavu izazvala veliku pozornost medija, kako
  pozitivnu tako i negativnu, shvatili smo kako je ova
  tema relevantna i aktualna.  Problem tolerancije izražen
  je na ulicama, u medijima, na sportskim priredbama i
  u svakodnevnom životu, pa smo se odlučili na okupljanje
  mladih iz cijelog svijeta kako bismo pokazali da, neovisno
  o tome iz koje zemlje dolaze, mogu ostvariti zajednički
  cilj ako se ujedine", izjavio je za SETimes Petar
  Antanasovski, predstavnik AISEC-a Srbije.
  ```

- In the box named *Corpus title*, add your name followed by a hyphen and word *mycorpus*, e.g. gramirez-mycorpus.
- On *Select the corpus language* on the drop-down menu choose *HBS* and click on the *Annotate & Train* option below that menu. You'll be transferred to the a new screen.
- In the drop down menu called *Language pair mode*, select *HBS–>NONE* and finally click on the *Start annotating button*. You'll be transferred to the *Corpus annotator* screen.

You'll see your *Corpus title* and *Corpus language* and below it the text you pasted in the first screen with some words in bold. These are the ambiguous ones according to the data encoded in Apertium: the ones that have more than one lexical form for a given surface form.

Your starting point is the first ambigous word (*Kada* in case you chose the sample text given). A set of *World alternatives* is shown on the right upper side of the screen, each one having a number. There are four in our example:

1. *Kada, cnjsub*
2. *Kada, adv*
3. *Kada, n, f, sg, nom*
4. *Kada, n, f, pl, gen*

[6] From SETtimes: `http://www.setimes.com/cocoon/setimes/xhtml/hr/features/setimes/audio_story/2013/07/25/audio_story-04`)

To select an option, you have to use the numbers in your keyboard to say whether you choose alternative 1, 2, 3 or 4 as the correct one taking into account the context of the sentence. Once you press on the option chosen, e.g. 1, you'll be transferred to the next ambiguous word. You can also use the left/right arrows in your keyboard to move from word to word.

Please, go on disambiguating your text. In case of doubt or when you don't find the right option, just press 1 and go to the next word. Sometimes, for adjectives specially, you'll have a hard time to see all options in the screen. Sorry about that, we are improving the design.

Finally, don't forget to press button *Save and Train* once you are done with the text.

Any ideas for improvement? We will discuss them together.

4.2 Tagsets

Many of the PoS taggers in Apertium rely on statistical disambiguation. Tagsets definition files are defined to help the statistical disambiguation module calculate probabilites to choose the correct part-of-speech. Tagsets contain mappings between all the morphological information delivered by the morphological analyser grouped in supra sets that have the same behaviour in a text.

To get started, we create groups for almost all main categories, separating closed and open, and we distinguish then between those that have a special role in disambiguation. Lemmas are not taken into account, only in special cases.

Let's take English as an example. In the English tagset:

- auxiliary verbs as "to be" (*VSER* or "to have" *VHAVE* are not grouped with the rest of verbs (*VLEX*).
- modal verbs as "can" also have a separate group (*VMOD*).
- we also distinguish between tenses: infinitives (*INF*), past participles (*PP*) and gerunds (*GER*) are separated from present (*PRES*) and past (*PAST*).
- singular and plural adjectives are grouped together (*ADJ*) but we distinguish between singular and plural nouns (*NOMSG*, *NOMPL*).

- and, when a noun shares ambiguitiy with other lexical form that is highly frequent, we put it in a separate special category and we take into account the lemma: this is the case of noun "can" (*CANNOM*).

Inside the tagsets, we may also define some rules to forbid sequences of categories, enforce a category after another one of set a preference for a category after another one. In this section, we work with the groups previously defined.

Following our English tagset:

- we **forbid** the sequence verb "to have" in past participle (*VHAVEPP*) followed by a verb in past tense (*PAST*), so we avoid bad reads of sentences like: *They've **had baked** potatoes in their set menu for years*.
- we **enforce** Saxon genitive (*GEN*) after proper nouns (*ANTROPONIM*, *TOPONIM*, *NPALTRES*) and others to avoid bad reads of ('s) as a form of verb "to be" in many sentences: *Jame**'s father**. Cat**'s eyes***.
- we give **preference** to acronyms (*n.acr.sg*) to help appropriate readings of sentences like: *I've been working in **IT deparments** for a long time*.

Let's take a look at the English tagset to discover other interesting groups, forbid and enforce rules:

Task 8. Inspecting a tagset [25 min.]

Having the list of Apertium symbols opened is going to be highly helpful for this practice too. Open it in a separate tab or window: `http://wiki.apertium.org/wiki/List_of_symbols`

Go to the dashboard of Annotatrix by clicking on Annotatrix in the upper menu or typing/clicking in `http://abumatran.eu:28000/` again.

Click on TSX Manager. You'll be seeing a screen to upload a TSX file or consult a previous uploaded one. To avoid uploading the tsx file for English, you'll find it available under *Your latest TSX* as a link to *apertium-en-es.en.tsx*. Click on this link and you'll see a view of the TSX having:

- **Labels**: on the right side of the screen
- **Tabs for Multi labels - Forbid rules - Enforce rules - Preferences**: on the left of the screen.

Take a look to *Labels* first:

- **Explore some of the Labels and try to understand why are they defined for**: you are seeing *Labels* given to grouped categories. If you click on any of them, e.g. *ADJ*, you'll see all the categories included in it: *adj* (beautiful), *adj.comp* (more beautiful), *adj.sup* (the most beautiful), *adj.sint* (long), *adj.sint.comp* (longer), *adj.sint.sup* (the longest).
  - What is *VDO closed* and *NOT closed*?
  - Why a special group for some determiners under the *DETQNT_ORD closed* label (much, many, enough, first, second)?
  - Are *ADJPOS closed* so different from other adjectives (mine, yours, hers, his, etc.)?
- **Try to understand some rules**: if you click on them you'll be able to see the sequence of forbidden, enforced and preferred. Explore some of the rules and try to find an example in which the rule should be applied:
  - Forbid:
    1. ADJPOS {+ NOMSG + NOMPL + NOMCAN + NOMWILL}
    2. SENT {+ RELAN + RELNN + RELADV}
  - Enforce-after:
    1. PREDET {+ NOMSG, + NOMPL, + CANNOM, + WILLNOM, + ADJ, + DET}

Congrats! Now you are an expert reader of Apertium tagsets!

Any improvement to this view? Ideas for new rules? They are more than welcome!

Recap and useful info

In this workshop we've introduced you to the basics of machine translation systems and Apertium dictionaries and part-of-speech tagger.

We thank you for your participation in this workshop and encourage you to join the Apertium community to help us improving. To do so, just subscribe to our mailing list or show up in the chat: we will help you to come in. You'll find how to contact us in our wiki page called Contact.[7]

User interfaces have come to Apertium to last. During the Abu-MaTran project, we will go on improving the ones you've been testing today and producing others. We want to hear about you if you have your say about them. Please contact us through the Abu-MaTran website form.[8]

This workshop is *to be continued*: we will be running another one for rules and advanced dictionaries next year around May. Stay tuned!

License

Please contact Gema Ramírez-Sánchez (gramirez at prompsit dot com) for a copy of the source files.

[7] `http://wiki.apertium.org/wiki/Contact`
[8] `http://www.abumatran.eu/?page_id=48`
[9] © Prompsit Language Engineering.