Using Apertium in a typical localization scenario

Spanish → Brazilian-Portuguese

EAMT 2010, Saint-Raphaël

Gema Ramírez-Sanchez, *Prompsit*
François Masselot, Petra Ribiczey, *Autodesk*

prompsit

Autodesk

This talk is about…

- A success story for an odd localization scenario
  - Why use a pivot language?
- Customizing a rule-based engine
  - How to proceed?
  - Where to stop?
  - How to keep costs under control?
- Integrating an MT service provider into a localization workflow

About Autodesk

- Autodesk is a software publisher
  - Design software (AutoCAD, Revit, Inventor…)
  - Rendering, animation software
  - For engineers, architects, animation films
- Localization Services (~100 employees worldwide)
  - manages processes, localization programs, systems (Corporate Terminology, CMS, TMS, MT…)
- Localization projects
  - Software UI, documentation
  - (user manuals, online help…)

About Prompsit

- Prompsit is a solution integrator in Machine Translation & Language Technologies
- Specialised with the Apertium open-source rule-based platform (involved since the beginning in 2004)
- Mixed group of software engineers, translators, and linguists (~5 employees worldwide)
- Academic background (Prompsit = spin-off of Transducens research group from the Universitat d'Alacant)

Project facts

- Translate for the first time, one of the company's flagship products in

  English > Brazilian Portuguese

- ~300,000 words software UI
- ~110,000 words Getting Started manuals
- Timing not critical
- Publishing quality expected
  (no damage to brand image)
- Post-edit whole content by human translators
- Need immediate ROI
  (can't amortize investments beyond the current project)
- Small bilingual in-domain corpus for this pair

Non-explicit objectives

Test adoption of MT

- Internally: sales, marketing, regional offices
- Externally: corporate reputation, public, end users

Validate MT integration

- in the localization workflow
- with LSPs partners

Acquire internal experience in MT

Setting expectations for MT vendors

"The MT output must be so that post-editors can reach processing 6,000 words per man.day"

Note: Usually admitted metric for regular translation: 2,500 words per man.day

Means we're asking to multiply throughput by 2

Figure estimated to:

cover customization costs

other process adaptation, learning curve

+ still some contengency

- If no MT solution could approach this "financial" goal, for this language pair, then it was better to do no MT at all, and do normal translation instead.

Vendor selection

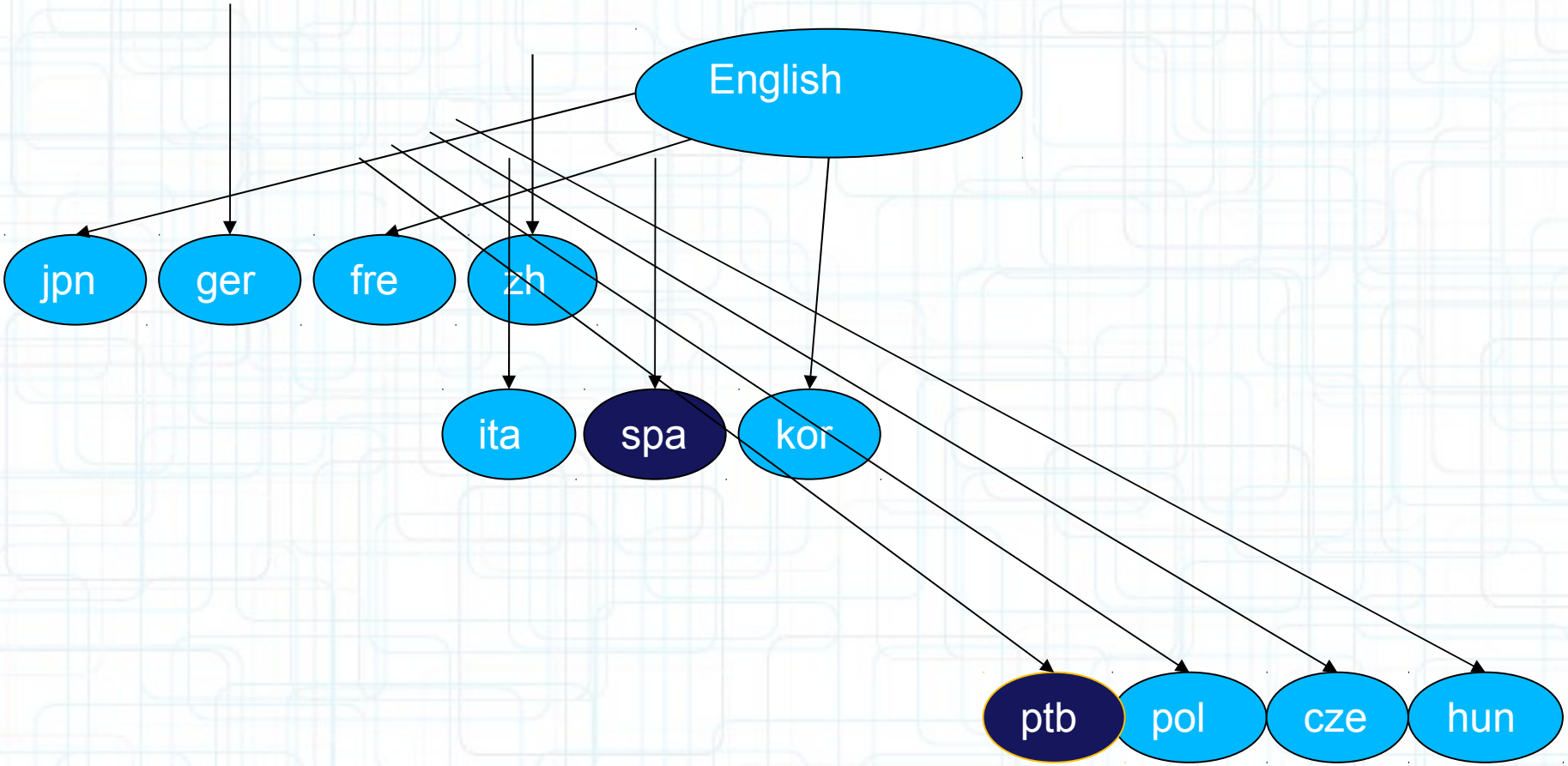
8 commercial proposals

- Some ruled-out right away:
  high license or customization costs
- English> Portuguese poor
  could hope to post-edit twice as fast as translating
- Stat MT discarded because of too small corpus

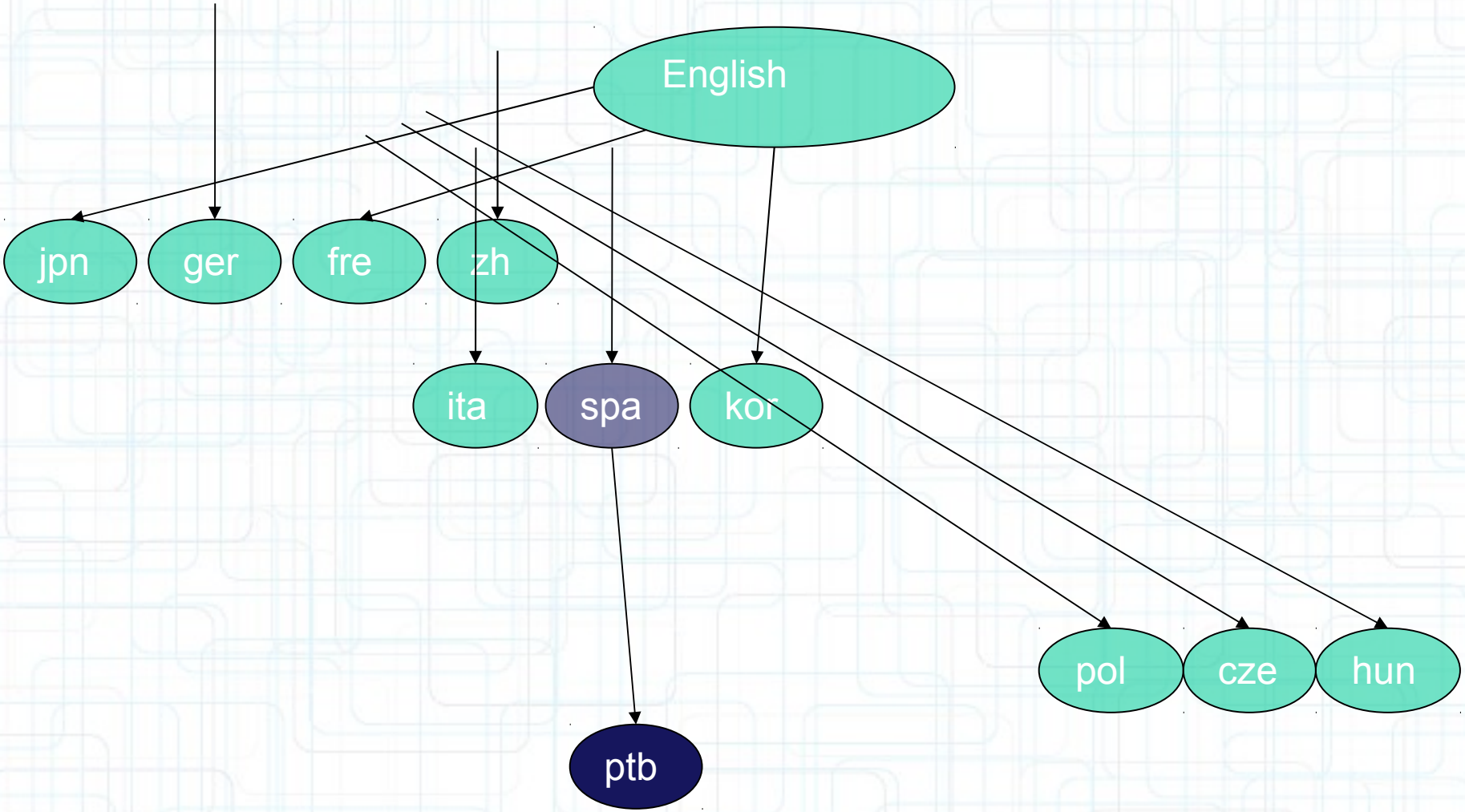
One proposal stands out: Prompsit

- Open-source
- MT service provider
- Pivot translation: English>Spanish>Portuguese

Pivot language



Pivot language

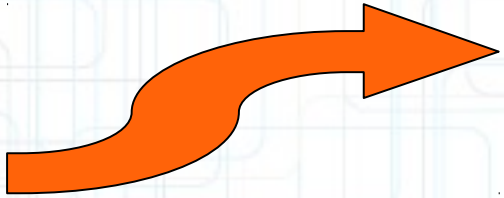


Pivot language

- Spanish translation was already under way
- Prompsit proposes an efficient shallow-transfer solution with Apertium
- Output without customization surpasses other solutions
- Some obvious areas for customization:
  - Usage of passive voice
  - New orthography
  - Domain terminology
- Good confidence that a proper customization would answer project objectives

The Apertium platform

Framework for rule-based MT systems



ENGINE

DATA

TOOLS

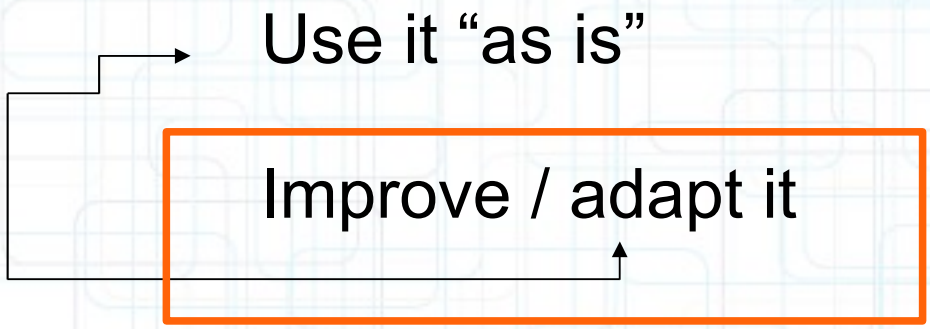
Free/Open-source resources

GNU General Public License

Apertium makes possible

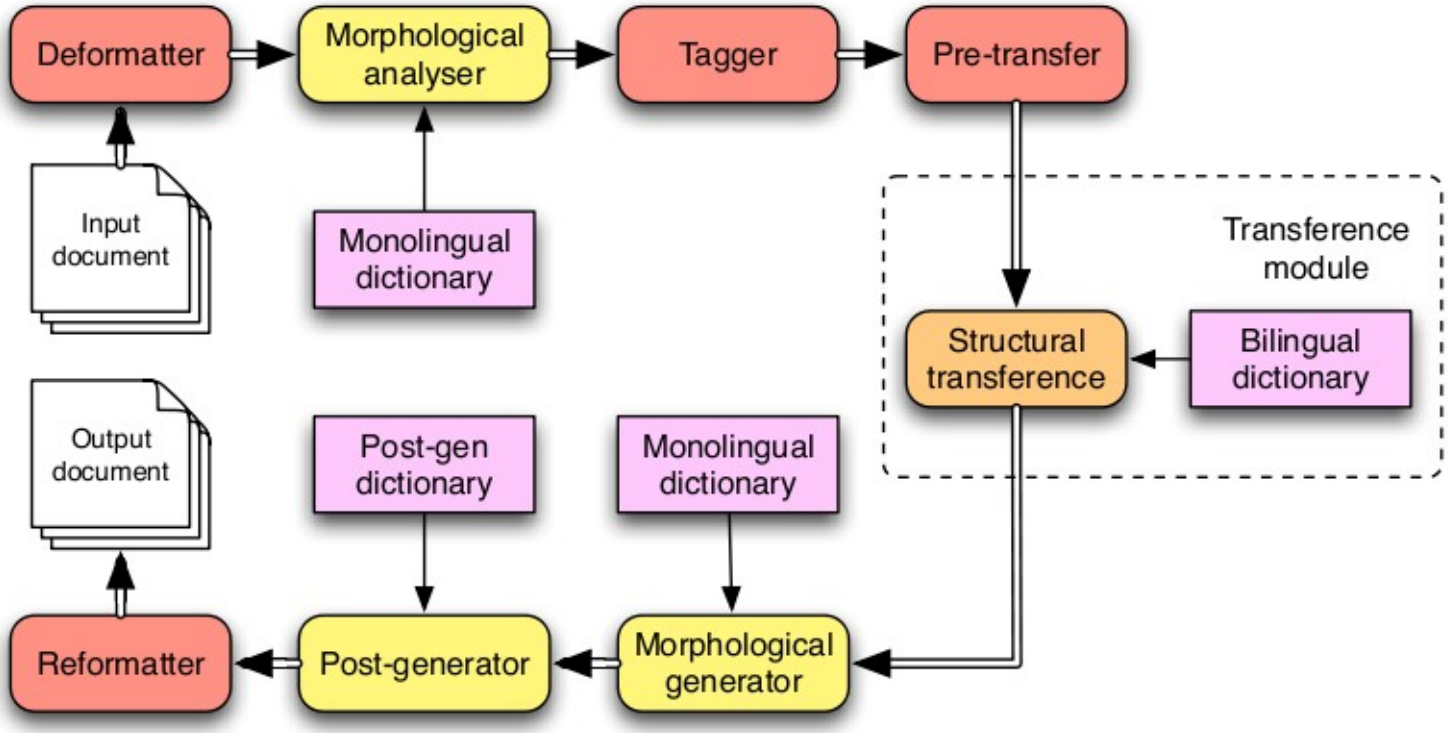
- Testing: how adequate for...
- Developing: I want a new...
- Adapting: could I have a customised...
  - engine?
  - data?
- Integrating: same workflow, new tools

Two possibilities



Where to start?

Modules and linguistic data in Apertium:



Customizing Apertium

Engine:

- unknown words: * → @@@
- encoding: utf-16 → utf-8
- special format filters:
  - CSV (comma separated value)
  - TMX (translation memory exchange)

Workflow adaptation:

- for software engineers: web service
- for post-editors: en→pt_BR translation units

Customizing *apertium-es-pt_BR* (I)

- **Expected:** publication quality output at 4000-6000 words/day for Brazilian Portuguese “2009”
- **Already in the box:** 10,000 lemmata, 100 transfer rules, Brazilian Portuguese variant
- **Missing:** new orthography for Portuguese, domain-adapted vocabulary and style
- **Decisions based on:**
  - expected results
  - available resources
  - time-cost-impact

Customizing *apertium-es-pt_BR* (II)

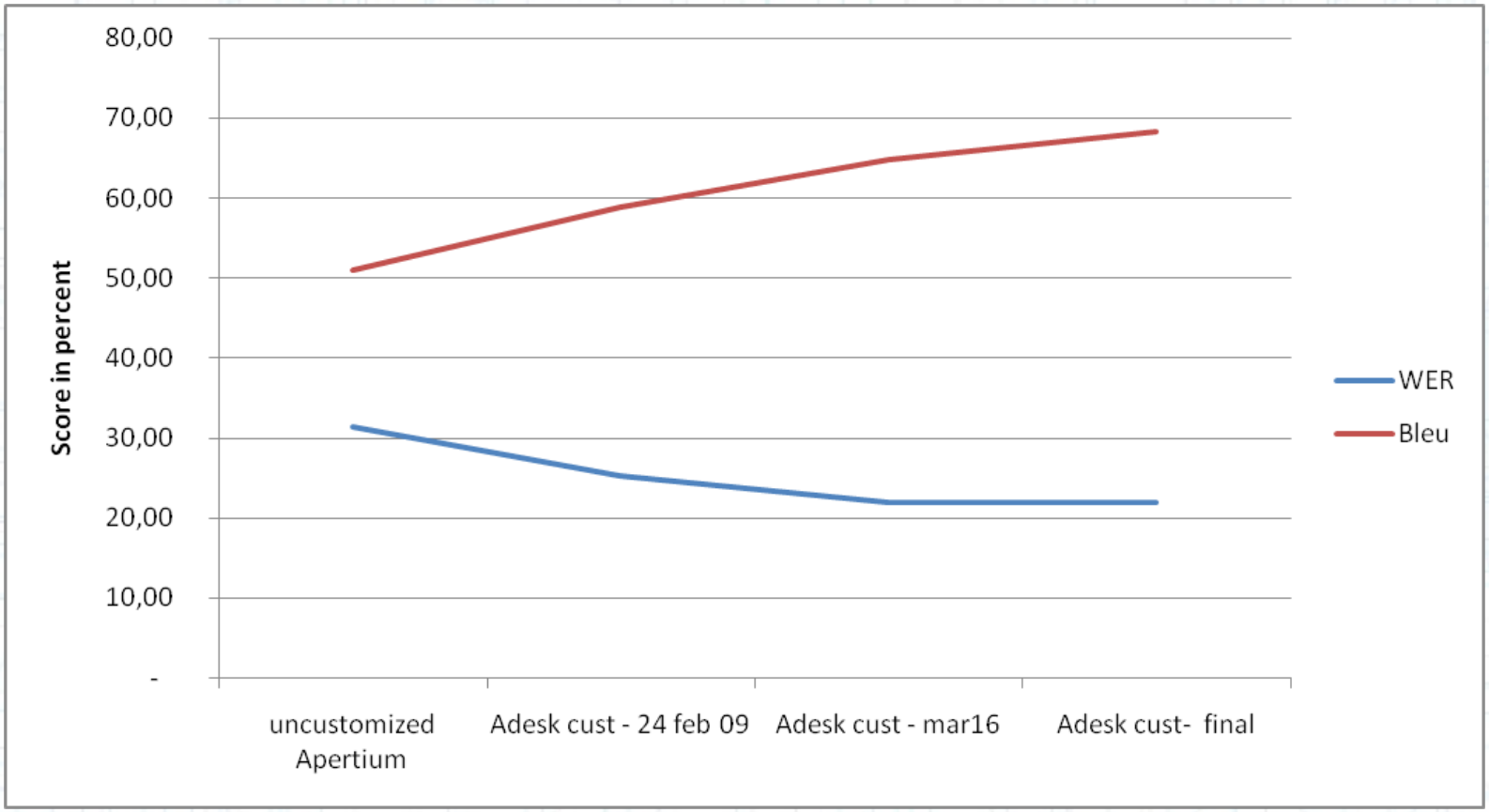
- **Compilation of resources and actions:**
  - multi-lingual glossaries (surface forms, not based on frequency) = Apertium-like entries in dictionaries
  - bilingual translation memories (*en-es* and *en-pt_BR*) = es-pt_BR parallel text = style checker to extract new transfer rules
  - the source language text to be translated = trilingual glossary turned into Apertium es-pt_BR Apertium dictionaries entries
  - new orthographical agreement = orthographical adaptation

Customizing *apertium-es-pt_BR* (III)

Some details:

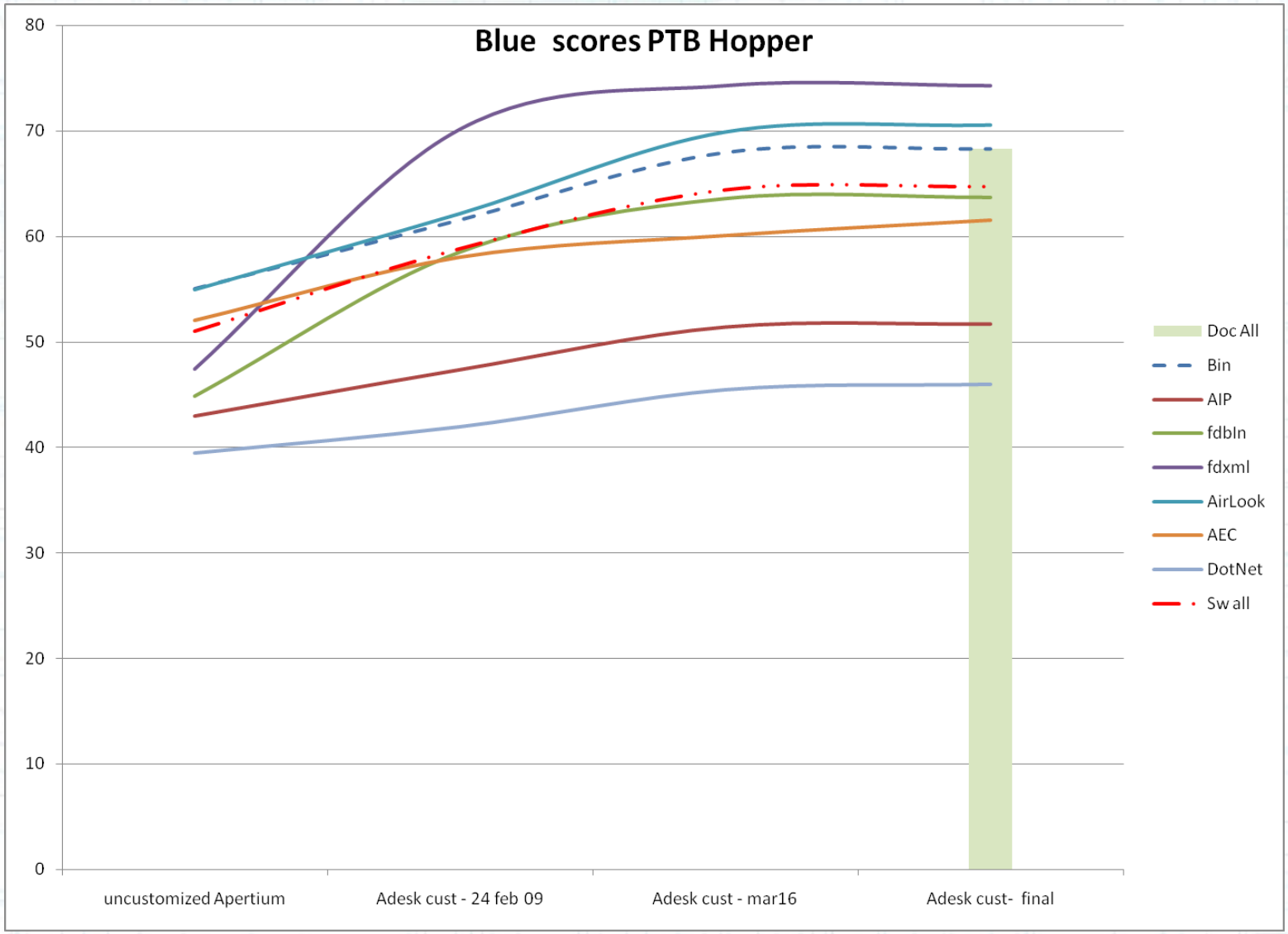
- two phases:
  - around 5 + 2 weeks
  - 2,285 new terms
  - 6 new transfer rules
- quality checks inside Autodesk term approval workflow and inside Apertium
- post-edition team feedback support
- after post-edition: proposal and agreement to contribute to the free version of apertium-es-pt
- evaluation: François will tell you...

Evaluation WER, Bleu



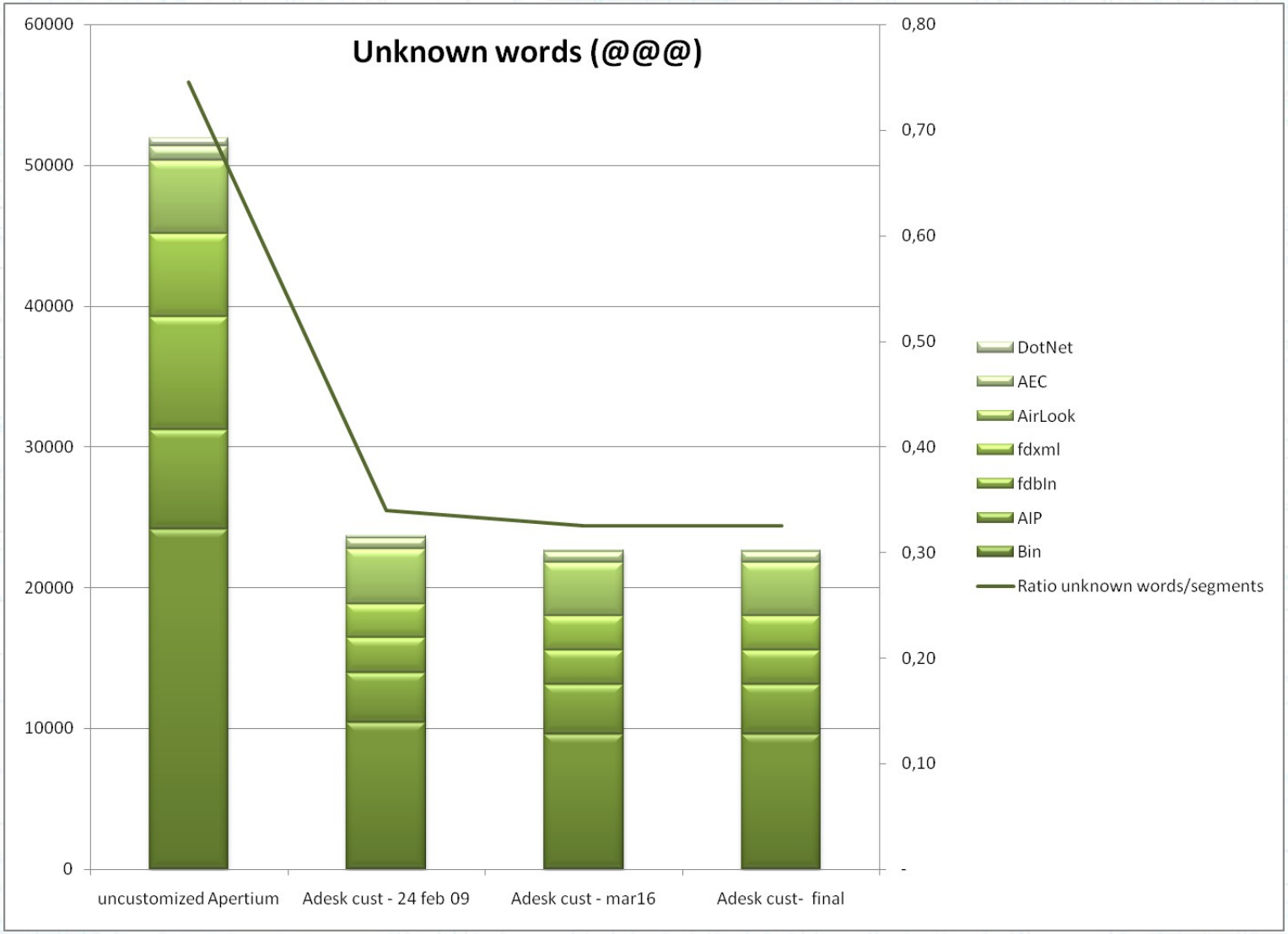
Edit distance between raw and post-edited

Edit distance



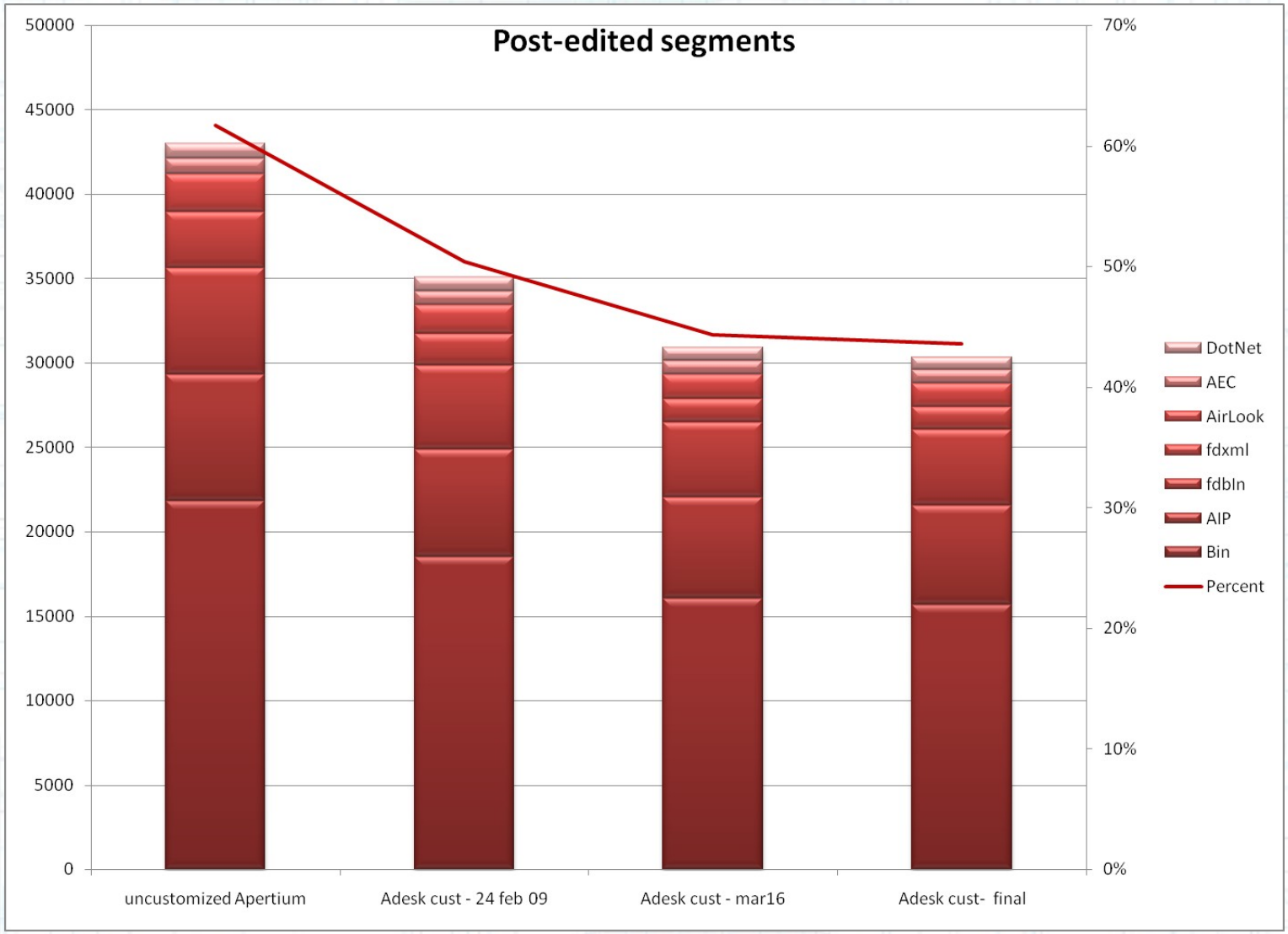
Edit distance between raw mt and post-edited (BLEU)

Unknown words



Coverage (proportion of unknown words – note: most of them are free-rides)

Post-edited TUs



Proportion of post-edited segments (vs those that didn't require post-edition)

This is it…

Questions ☺

prompsit

Autodesk