Language Factory

Fourth MT Marathon

Dublin, 28 January 2010

M.T. Carrasco Benitez

European Commission

Final product

- Multilingual parallel texts
- Camera ready
- Quality, speed, cost (QSC)

- Language factories

Phases

- Authoring, Translation, Publishing chain (ATP-chain)
- Our case: translation
- Translation: human, machine, cyborg
- Our case: machine translation

Machine translation

- Translation memories (TM)
- Example based
- Statistical based
- Rule based
- Black magic based

Interoperability

- Rail gauge
- Programs: Unix command
- Data: data structure

- Internet aware: URI

Programs: linguistic pipe

- Machine translation

```
tm source.med | smt | rbmt |
  chooser
```

- Computer-aided translation

```
tm source.med | smt | rbmt |
  cat
```

Data

- Directory structure
- File format - SQLite
- Large volumes of text
- Every character counts: save, compress
- Unicode: One character boundaries
- Segments: word, sentence, paragraph

More on data

- Multilingual Corpus Format
  - One package: we know the content
  - Full docs, segments,
- Multilingual Electronic Dossier (MED)
  - ATP-chain

Show biz

- No Ph.D. for show biz
- Write **standards**: e.g., RFC
- Lower the entrance barrier
  - The Sun: vulgarisation
  - Package: easy to install (train, translate)

End

- http://dragoman.org/langfab.pdf