Finnish and MT

Austin Matthews

2015/01/13

Sample Text

Kaikki	ihmiset	syntyvät	vapaina	ja	tasavertaisina	arvoltaan	ja	oikeuksiltaan.
All	people.NOM	born.3PL	free.PL	and	equal.PL	dignity.INS	and	rights.INS.
All human beings are born free and equal in dignity and rights.								

Heille	on	annettu	järki	ja	omatunto,
They.ALL	are	give.PASS	reason.NOM	and	conscience.NOM
They are endowed with reason and conscience					

ja	heidän	on	toimittava	toisiaan	kohtaan	veljeyden	hengessä.
and	they.PART	are	act.PASSPART	each:other	towards	brotherhood.GEN	spirit.INES.
and should act towards one another in a spirit of brotherhood.							

Why is Finnish important?

- ～5M native speakers
- Spoken primarily in Finland with minority language status in Sweden and Norway
- One of the more wealthy/developed member states of the EU
- Related to Estonian and Hungarian as part of the Uralic language family

Why is Finnish interesting?

- Because it’s not Indo-European!
- Aggulatinative/Fusional
- Extreme number of cases (16)
  - Reduces the need for adpositions
- Free word order
  - Neutral order: SVO
- Vowel Harmony

Kolmivaihekilowattituntimittari
three phase kilowatt hour meter

puhua + nut → puhunut
speak + PASTPART → was:speaking
syöda + nut → syönyt
eat + PASTPART → was:eating

MT Challenges

- Rich morphology leads to sparsity and difficulty choosing forms to output
  - If we see English “they” do we output *he*, *heidän*, *heitä*, *heissä*, or one of the other 12 forms?
- Agglutination compounds* generation problems
  - If we see “parking meter” is that *pysäköinti mittari* or *pysäköintimittari*
- Free word order means there’s no single “correct” answer
  - “the dog bit the man” might become *koira puri miestä* or *miestä puri koira*
- Vowel harmony breaks simple concatenative models of morphology

* Pun intended

NLP Resources and MT history

- There exist a Finnish TreeBank and Propbank
- Dependency parser
- Surprisingly much parallel data
  - EU parliamentary meetings
  - EU laws
  - Movie subtitles
  - Software translations
  - Total of ～400M words from OPUS and ～200M from TAUS
- Has occasionally ben used for MT research dating back to ‘93
- Supported by most translation companies (Google, Microsoft, Systran, AsiaOnline, Safaba)
- Finnish is this year’s WMT surprise language!