Hungarian / Magyar

Kazuya Kawakami

Why Hungarian??

- Hungary was beautiful!!





- Agota Kristof (*Kristóf Ágota*)
  - Hungarian writer lived in Switzerland and wrote in French.
  - Left Hungary because of the anti-communist revolution
  - “The Notebook” / “The Proof” / “The Third Lie”

Sample Text

- Parallel sentence from Bible

Chapter 1

1 In the beginning God created the heaven and the earth.
2 And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters.
3 And God said, Let there be light: and there was light.
4 And God saw the light, that it was good: and God divided the light from the darkness.
5 And God called the light Day, and the darkness he called Night. And the evening and the morning were the first day.

Fejezet 1

1 Kezdetben teremté Isten az eget és a földet.
2 A föld pedig kietlen és puszta vala, és setétség vala a mélység színén, és az Isten Lelke lebeg vala a vizek felett.
3 És monda Isten: Legyen világosság: és lõn világosság.
4 És látá Isten, hogy jó a világosság; és elválasztá Isten a világosságot a setétségtõl.
5 És nevezé Isten a világosságot nappalnak, és a setétséget nevezé éjszakának: és lõn este és lõn reggel, elsõ nap.

Overview

- 10m native speakers in Hungary, 2m in Romania, 700k in Slovakia
- Minority in Austria, North/South America, Israel, Australia
- Uralic ⊃ Hungarian, Estonian, Finish
- Script
  Past : written in right-to-left Old Hungarian runes
  Now: left-to-right order Latin alphabet.
- Phonology
  14 vowel phonemes and 25 consonant phonemes.
- Grammar
  - Affixed word and compound words (Buda+Pest)
  - Vowel harmony
  - 18 Case markers
    ház (house)+ -am (my), -ad (your),-a (his),-unk(our),-atok(your), -uk(their) ...

Latin(above) and old Hungarian script (bottom)

Syntax

- Free word order language
  Janos Keresi Marit.
  John seeks Mary(-ACC).

János keresi Marit.	Marit János keresi.
János Marit keresi.	Keresi János Marit.
Marit keresi János.	Keresi Marit János.

- Topic part + Predicate part = sentence
  Unlike English, topic role is independent of the function “grammatical subjects”.

a. [$_{Top}$ János] [$_{Pred}$ fel hívta Marit][1]
   John up called Mary-ACC[2]
   ‘John called up Mary.’

b. [$_{Top}$ Marit] [$_{Pred}$ fel hívta János]
   Mary-ACC up called John-NOM
   ‘Mary was called up by John.’

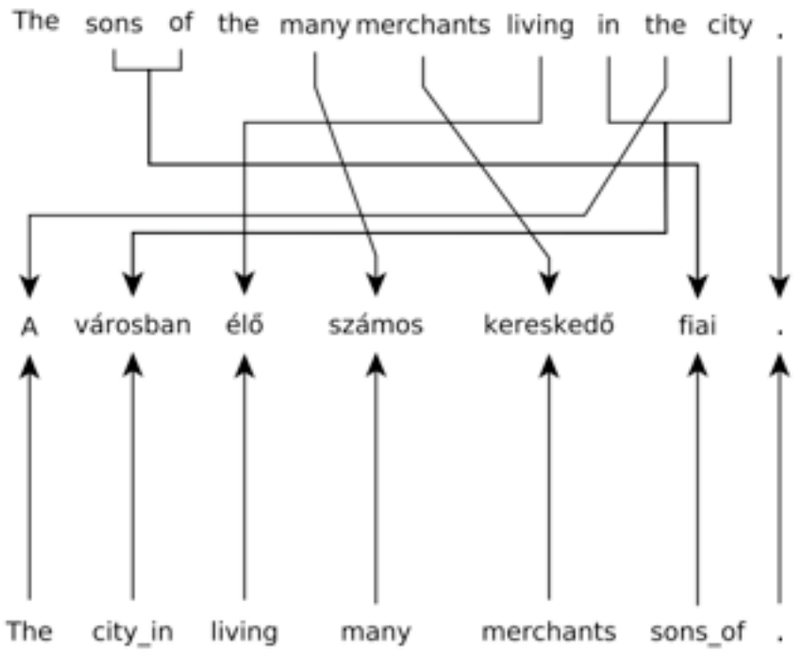
MT perspective

- Traditional Phrase-based SMT system does not work. [1]
  - Hungarian, like Finish or Turkish, is an agglutinating and compounding language, which morphological processes yield a huge number of different word forms.
  => many function words in English are expressed as suffixes in Hungarian.
  - Free word order of main grammatical constituents.

- Hybrid morpheme-based machine translation system with reordering rules works.
  - Dependency based Reordering: make more appropriate sentence structure
  - Morpheme-based restructuring: “hazamban” => “haz” + “am” + “ban”

Before reord

After reord



	mm-BLEU
Google	55.86%
Bing	53.05%
m-based reord	64.94%

[1] An English-to-Hungarian Morpheme-based Statistical Machine Translation System with Reordering Rules, Laszlo et al.

NLP Resources

- Parallel text exists. (Europal, LDC2008T01)
- Szeged Treebank is the largest hungarian TreeBank. [1]
  — 82,000 sentences, 1.2 million words, 250,000 punctuation marks.
- Malt / MST / Mate parsers are applied. [2]

corpus		Malt		MST		Mate	
		ULA	LAS	ULA	LAS	ULA	LAS
Hungarian	dev	88.3 (89.9)	85.7 (87.9)	86.9 (88.5)	80.9 (82.9)	89.7 (91.1)	86.8 (89.0)
	test	88.7 (90.2)	86.1 (88.2)	87.5 (89.0)	81.6 (83.5)	90.1 (91.5)	87.2 (89.4)
English	dev	87.8 (89.1)	84.5 (86.1)	89.4 (91.2)	86.1 (87.7)	91.6 (92.7)	88.5 (90.0)
	test	88.8 (89.9)	86.2 (87.6)	90.7 (91.8)	87.7 (89.2)	92.6 (93.4)	90.3 (91.5)

Table 1: Results achieved by the three parsers on the (full) Hungarian (Szeged Dependency Treebank) and English (CoNLL-2009) datasets. The scores in brackets are achieved with gold-standard POS tagging.

- Morphological Analyzer / Annotation system exists. [3]
- POS tagging accuracy is comparable to state of the art english taggers.
  - POS disambiguation with morphological analyzer got 98% in accuracy. [4]

[1] TreeBank: http://www.inf.u-szeged.hu/projectdirs/hlt/en/Szeged%20Treebank%202.0_en.html
[2] Dependency Parsing of Hungarian: Baseline results and Challenges, Richard Farkas et al.
[3] Hunmorph http://mokk.bme.hu/resources/hunmorph/
[4] Using a morpholigical analyzer in high precision POS tagging of Hungarian, Peter Halacsy et al.