Tutorial on Linguistic Typology

Emily M. Bender       Jeremy G. Kahn

November 24, 2003

1 The language sample

- Spanish, French, English, Russian, Arabic, Chinese (UN languages)
- Finnish, German, Swedish, Greek (selected other EU languages)

Language families:

Finno-Ugric | Sino-Tibetan | Semitic
---|---|---
Finnish | Chinese | Arabic

Indo-European

Romance | Germanic | Greek | Slavic
---|---|---|---
Spanish | French | English | German | Swedish | Greek | Russian

- The history of any particular language is complex enough that the tree metaphor isn’t unproblematic. It doesn’t account for either massive borrowing from one language or the other or for the fact that concepts like ‘English’ and ‘French’ are in some sense idealizations over lots and lots of different linguistic varieties.

- Nonetheless, the ‘genetic’ classification of languages gives some information as to their typological similarity.

- For a very extensive classification of lots and lots of languages, see the *Ethnologue*:

  http://www.ethnologue.com/family_index.asp

2 Basic word order

- In order of frequency: SOV, SVO, VSO, VOS, OVS, (OSV)
• That only tells part of the story. Even in English (which is considered very strict in its SVO word order) you also get:

  – Heavy NP shift:
    (1) a. Kim found [yesterday] [the first new species of frog to be discovered in a decade.]
        b. Kim found [a frog] [yesterday].
        c. *Kim found [yesterday] [a frog].
  – Dative alternation
    (2) a. Kim gave a book to Sandy.
        b. Kim gave Sandy a book.
        c. Kim donated a book to Sandy.
  – Cleft & pseudocleft sentences
    (3) a. It was a book that Kim bought.
        b. What Kim bought was a book.
  – Only VO, V or SV:
    (4) a. Eat food!
        b. Eat!
        c. Kim ate.
  – "Wh-movement"
    (5) a. What did Kim eat?
        b. The food which Kim ate was spoiled.
  – Focus movement:
    (6) a. Bagels Kim likes.
  – Verb-particle construction:
    (7) a. Kim looked up the number.
        b. Kim looked the number up.

• There’s also other things in a sentence besides the S, V and O (notably adverbs)...
• In some languages, it’s hard to establish a ‘basic word order’ of S, V and O (e.g., V2 languages, Warlpiri, etc.).
• V2: in main clauses, (finite) verb is second with a ‘topic’ slot before it; in subordinate clauses, finite verb is at the end.
• Some other aspects of word order roughly correlate with the order of SVO:
- OV goes with postpositions, VO with prepositions
- SVO goes with post-nominal relative clauses, SOV with pre-nominal
- Not perfect: Chinese is SVO with pre-nominal relative clauses

- Discontinuous constituents:

  (8) a. Deine Schwester habe ich gestern in der Stadt gesehen
      your sister have I yesterday in the town seen
      ‘I saw your sister in town yesterday.’
  b. Which book did you read of Asimov’s?

- SVO: English, Spanish, French, Chinese, Finnish (was SOV), Arabic, Russian (though very loose)
- V2: Swedish, German
- SOV: Hindi, Japanese

3 **Morphological type**

- Two dimensional contrast:

  - Polysynthetic
  - Fusional
  - Agglutinating
  - Isolating

- Isolating: one morpheme per word
- Polysynthetic: many morphemes per word
- Fusional: morpheme boundaries unclear/many functions per morpheme (*portmanteau* morphemes)
• Agglutinating: clear morpheme boundaries, one function per morpheme

• The polysynthetic-isolating dimension varies along the number of morphemes-per-word.

• The agglutinating-fusional dimension varies along the number of meanings-per-morpheme.

• This 2D presentation is correct in that the two dimensions are partially independent. It is misleading in that
  
  – the origin is not meaningful
  
  – very isolating languages are neither fusional nor agglutinating

• While we’re at it: Most languages with any polysynthetic tendencies string morphemes together or (at worst) signal inflection with ablaut (changes to the vowel in the stem). Semitic languages (including Arabic), instead have root-and-pattern (“interdigitated”) systems where the roots are sequences of 3 (sometimes 4) consonants, and the inflectional information is expressed through patterns of with slots for those consonants and particular vowels.

**Ramifications for MT**

• Implications for word segmentation and morphological analysis (both necessity and difficulty of). Consider the IBM models when the concept of “word” is radically different.

  (9) KIRJA +sto -i -ssa -mme -kin
      book +coll -pl -INE -POSS -CLIT
      ‘In our libraries, too’

4 **Case and Agreement: Head v. dependent marking**

• Case and agreement serve as syntactic cues to the role each piece of the sentence plays in its predicate-argument structure.

• For these purposes, syntactic position in a rigid word-order language can be seen as a type of ‘case’.

• In dependent-marking, the cue is encoded on the argument (mainly, noun phrases), e.g., via position, case marking, prepositions, etc.

• In head-marking, the cue is encoded on the predicate (mainly, verbs) with morphemes that signal agreement with the relevant NP in the sentence.

• Subject-verb agreement is a (weak) kind of head-marking.

• To the extent that Spanish clitics have in fact become agreement markers, Spanish is exhibiting further head-marking.
(10) a. Le entregó el libro a alguien.
   3sg-ind-obj gave the book to someone.
   ‘He/she gave the book to someone.’

b. Mi madre quiere conocerlo al novio de Blanca.
   My mother wants to-know-him the boyfriend of Blanca
   ‘My mother wants to know Blanca’s boyfriend.’

c. Juan la vio a ella.
   Juan her saw to her.
   ‘Juan saw her.’

- Freedom of word-order is supposed to correlate with how well predicate-argument structure can be determined by other means, but there are counter-examples.

4.1 Case

- Languages with case systems vary in the number of case distinctions they have from 2 (English, marked on pronouns only) to roughly 15 (Finnish).

- In many languages, one case will be “unmarked”, that is, correspond to the lack of any special morphology.

4.2 Agreement

- Agreement systems use features like person, number, and noun class.

- Person: 1 (speaker) 2 (hearer) 3 (other), sometimes 4 (far away other). 1st person plural may include an ‘inclusive’ v. ‘exclusive’ contrast.

- Number: singular/plural; singular/dual/plural, singular/dual/trial/plural (Fijian)

- Noun class includes (grammatical) gender systems (e.g., French, Spanish, Russian, German) with 2-3 classes, as well as more elaborated systems (Bantu) with roughly 17-18 classes.

- Gender, number, and also case can be used for agreement within noun phrases.

- English 3rd person singular pronouns have a gender distinction, but the language in general doesn’t.

- Russian counting does something funny with case on the associated (counted) noun for numbers up to 10.
Ramifications for MT

- Consider the sentence:
  
  (11) a. The police denied the students the permit because they feared violence.
  
  b. The police denied the students the permit because they advocated violence.

  Translating this into (e.g.) French, we would expect they to have a different agreement
  (in French, police is feminine and student is masculine). But the only difference here
  is the verb!

- Any given pair of languages is likely to mark different categories with its morphology
  (if any), leading to many-to-many mappings between word forms even if there were
  one-to-one mappings between lemmata. Language models for the target language can
  probably take care of some of this (e.g., gender agreement within noun phrases, since
  the gender of any given noun is fixed).

5 Pro-drop

- All arguments must be expressed (usually) (English, Dutch – these languages have
  two different expletive subjects).

  (12) a. It’s raining.
  
  b. There’s a unicorn in the garden.
  
  c. *Is raining.
  
  d. *Is a unicorn in the garden.
  
  e. They found it.
  
  f. *They found.
  
  g. They ate it.
  
  h. They ate.
  
  i. My neighbor opened his door.
  
  j. *My neighbor opened.
  
  k. What time do you open your store?
  
  l. What time do you open?

- The subject can be dropped if it is clear from context (subject-verb agreement is
  supposed to help) (Spanish)

  (13) a. (Nosotros) necesitamos estudiar.
  
  (We) need-1pl study-inf
  ‘We need to study.’
b. (Él) necesita estudiar.
   (He) need-3sg study-inf
   ‘He needs to study.’

- Just about any argument can be dropped, if it is clear from context (Chinese, Japanese — oddly enough, languages without subject-verb agreement)

(14) Mandarin:
   吃le
   eat perf
   ‘(I) ate (it)’

Ramifications for MT

- In one direction, pro-drop will provide cases where the null element has non-zero fertility. Typically, the corresponding target language morpheme should be a pronoun (expletive or contentful). To the extent that the target language requires various kinds of morphosyntactic information on the pronoun (number, gender, case, etc), there will be issues similar to the feared violence cases discussed above.

- In the other direction, an MT system with good fluency would need to know when to give source-language pronouns (or sometimes even full noun phrases, but this would probably be less common) zero fertility, as using pronouns when pro-drop is an option tends to have pragmatic effects which could lead to awkward translations.

6 Topic- v. subject-oriented languages

- Topic-oriented languages (Chinese, Japanese, Korean, ...) have topics (directly translated as “as for X”) in most sentences.

- Subject-oriented languages (English, French, most IE) have subjects in most sentences.

- In the extreme case, topic-orientation gives sentences like:

(15) Japanese:
   Amerika-ha ie-ga ookii
   America-TOP house-NOM big
   ‘As for America, houses are big.’
   ‘America has big houses.’
   ‘The houses in America are big.’
   ‘In America, the houses are big.’

This can also include double-subject constructions, such as:

(16) Jon -ga oji -ga isha -da
   John NOM uncle NOM doctor is
   ‘It is John whose uncle is a doctor.’
• Subject-oriented languages have:
  – Subjects syntactically prominent
  – Passive constructions relatively robust (has grammatical, not topical, forms)
  – “As-for” topics relatively marked
  – Presence of expletive subjects
  – Subject is typical controller of missing constituents
  – Absence of double subject constructions

Ramifications for MT
• Informally, it seems that these kinds of differences between languages lead human translators to create translations that have very different structure from the source sentence.

7 Tense, Aspect, and Mood
• Tense: Location of the event described with respect to the time of utterance and/or the reference time of the sentence.

  (17) a. Before traveling to Seattle, Kim bought a raincoat.
       b. Before traveling to Seattle, Kim will have bought a raincoat.
       c. Before traveling to Seattle, Kim will buy a raincoat.

• Mood: contrasts signaling modality (irrealis/realis, necessity, possibility, ...)

  (18) a. Go to the store! (imperative)
       b. He went to the store. (indicative)
       c. I suggest that he go to the store. (subjunctive)

English makes mood distinctions quite weakly, but it is robust in Romance languages. Mood can be marked on verb form (Romance languages), with separate (auxiliary) verbs (English would that, for optative mood), or with particles (Chinese ba for imperative mood, ma for interrogative).

• Aspect: Temporal properties of the event itself, inherent in the action (Aktionsart) or superimposed, as the perspective taken on the event (aspect).

  – Some kinds of aspect: completive, inchoative, progressive, perfective, imperfective, iterative, punctual, habitual, experiential ...
– Particular linguistic systems grammaticize only some kinds of tense and/or aspect. (Sometimes primarily only tense, or primarily only aspect, sometimes only certain tense/aspect combinations, e.g. English *use*ta, which marks past (tense) habitual (aspect).)

**perfective** sees the event as a whole (*I swam, I have swum*)

**imperfective** sees the event as a process (*I used to swim, I am swimming*)

**inchoative** sees the event as a beginning (*I began to swim*)

**cessative** sees the event as an ending (*I finished swimming*).

### Ramifications for MT

- Different languages grammaticize different kinds of semantic information in the field of tense-aspect-mood. Some languages focus more on tense, others more on aspect, leaving the rest of the interpretation either underspecified or to be determined by context. Thus the correct translation from one system into another can vary greatly from token to token.

### 8 Writing systems/written language

- Sound-based (Roman, Greek, Cyrillic alphabets, kana, hangul, Arabic, ...) v. meaning-based (hanzi/kanji)

- Within sound-based: based on syllable (Japanese), segment (English), or feature (Korean).

- May or may not introduce homographs (Arabic) or distinguish homophones (Chinese).

- The written and spoken languages of a culture may be relatively similar (English) or different (Cantonese, Arabic) in grammar & lexis.

- When translating between different writing systems, a further problem of transliterated names arises. While the system seems pretty straightforward for Japanese and Cyrillic, Chinese and Arabic (for both phonological and orthographical reasons) raise thornier problems in this regard.

- Writing systems may or may not indicate word boundaries.

### Ramifications for MT

- Segmentation issues (what’s a word?)

- Frequency of homographs can complicate the alignment problem.

- Named entity recognition (especially of transliterated foreign person and place names)
References
