Language Technology: Research and Development

Introduction

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se
Course Content

<table>
<thead>
<tr>
<th>Theory</th>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philosophy of science</td>
<td>Survey a research field</td>
</tr>
<tr>
<td>Research methods in LT</td>
<td>Plan and implement a project</td>
</tr>
<tr>
<td>Scientific writing</td>
<td>Write and review scientific papers</td>
</tr>
</tbody>
</table>

- Two old courses merged into one for synergy
- Lectures (theory) and seminars (practice) in parallel
- Individual projects with a common research theme
Research Theme

- Multiword expressions (MWE)
  - Idiosyncratic interpretations that cross word boundaries (spaces) (Sag et al., 2002).

- Examples:
  - Idioms: spill the beans, kick the bucket
  - Nominal compounds: rubber duck, natural language processing
  - Light verb constructions: take a photo, make a claim
  - Particle verbs: put off, cut short
  - Complex prepositions: in spite of, because of
Course Structure

1. Background part:
   ▶ Philosophy of science and research methods [lectures]
   ▶ Survey of the state of the art in MWE research [seminars]
   ▶ Planning an R&D project [lecture, seminar]

2. Project part:
   ▶ Implementing an R&D project [seminars]
   ▶ Writing a scientific paper [lecture, seminar]
   ▶ Reviewing scientific papers [lecture]
Reading List

▶ Textbooks:

▶ Papers:
  ▶ Available on line from the course home page http://stp.lingfil.uu.se/~nivre/master/fou.html
Assignments and Examination

1. Take home exam on philosophy of science (15%) [written]
2. Research paper presentation (15%) [oral]
3. Project proposal (15%) [written, oral]
4. Review of term papers (15%) [written]
5. Term paper (40%) [written, oral × 2]

- Pass (G) = all assignments passed
- Distinction (VG) = at least 50% passed with distinction
Going for the Real Thing

- The goal is to do real research resulting in real publications
- Guidelines for submission and reviews:
  - Transactions of the Association for Computational Linguistics
    http://www.transacl.org/submission/
- Some possible venues:
  - ACL, Beijing (Feb 27, 2015)
  - MWE Workshop (March?)
  - TACL (1st of every month)
- Actual submission is not a course requirement
Questions?
Multiword Expressions

- MWEs transcend linguistic levels
  - They look like syntactic constructions
  - They behave like lexical units (to some degree)
  - They do not fit well into standard NLP models
- Examples from Google Translate (en → sv):
  - He spilled the beans inadvertently. → Han spillde bönor oavsiktligt.
  - She made a quick decision. → Hon gjorde ett snabbt beslut.
  - She put up with him. → Hon satte upp med honom.
Definition

- Baldwin and Kim (2010): MWEs are lexical items that
  1. can be decomposed into multiple lexemes,
  2. display lexical, syntactic, semantic, or pragmatic idiomaticity.

- Types of idiomaticity:
  1. Lexical: One or more components not in lexicon (ad hoc).
  2. Syntactic: Anomalous syntactic construction (by and large).
  3. Semantic: Meaning not derivable from parts (kick the bucket).
  4. Pragmatic: Tied to specific situation (good morning).
  5. Statistical: Markedly high frequency (salt and pepper).
Nominal Compounds (NC)

- Compounds consisting of noun + modifier:
  1. Noun: orange juice, golf club
  2. Deverbal noun: investor hesitation, stress avoidance
  3. Adjective/participle: connecting flight, attorney general
  4. Prepositional phrase: part of speech, jus d’orange
  5. Fused compounds: paperclip apelsinjuice

- Issues:
  1. Semantic interpretation (paper hat, paper clip)
  2. Syntactic disambiguation with 3 or more components (natural language processing, fresh orange juice)
Verbal MWEs

- **Verb-particle constructions (VPC)**
  - Limited variability: she *put* a hat *on*; *she put* quickly *on* a hat
  - Cross-linguistic variation: hon *tog på* den vs. hun *tog* den *på*

- **Prepositional verbs (PV)**
  - Variable or fixed: it was *referred to*; *it was come across*

- **Light-verb constructions (LVC)**
  - Verb semantically thin: *take a photo*, *make a decision*
  - Very important in some languages (Hindi, Persian, Japanese)

- **Verb-noun idiom combinations (VNICT)**
  - Decomposable or not: *spill the beans* vs. *kick the bucket*
Function Words

- **Determined prepositional phrases**
  1. Adverbial: stay *on top*, go *by car*, go *out of town*
  2. Compound modifiers: *jus d’orange* (vs. *jus de l’orange*)

- **Complex function words:**
  1. Prepositions: *in spite of*, thanks to
  2. Conjunctions: *both . . . and*, so that
  3. Adverbs: *by and large*, *in any case*
Classification of MWEs

- Institutionalized phrases (statistical idiomaticity only)
  - Collocations: salt and pepper; sur mjölk vs. härsket smör

- Lexicalized phrases (lexical, syntactic or semantic idiomaticity)
  1. Fixed expressions (no variability): ad hoc, by and large
  2. Semi-fixed (inflection): kick(s/ed) the bucket, part(s) of speech
  3. Flexible (syntax): put up, look for, take a walk, spill the beans
Research Problems

- **Representation:**
  - Linguistic treatment of MWEs
  - Construction of lexical resources, grammars and ontologies

- **Processing:**
  - Extract MWEs (types) for resource building
  - Identify MWEs (tokens) for interpretation
  - Syntactic disambiguation
  - Semantic interpretation

- **Improving applications:**
  - Machine translation
  - Syntactic parsing
  - Word sense disambiguation
  - Information retrieval
A Short History of MWE Research

▶ 90s:
  ▶ Collocation extraction (Church and Hanks, 1990; Smadja 1993)
  ▶ Mostly statistical methods
▶ Early 00s:
  ▶ Grammar and lexicon (Sag et al., 2002; Calzolari et al., 2002)
  ▶ Stanford project ("MWE a pain in the neck for NLP")
▶ More recently:
  ▶ Extraction and identification
  ▶ Disambiguation and interpretation
  ▶ Application studies (MT, parsing, etc.)
▶ Very active research area:
  ▶ Dedicated workshops since 2003
Papers