Translation and categorization

Diana Santos

ILOS, 22 September 2011
Sources of inspiration

• Ellis, John M. *Language, Thought, and Logic.*

• Hofstader, Douglas. *Le ton beau de Marot: Praise for the music of language.*

• Sig Johansson and Lauri Carlson

• Whorf, Catford, Sampson, Veale, Borges, Sandström, Kilgarriff, Saussure, Snell-Hornby, Nakhimovsky, Vinay & Darbelnet, Lakoff & Johnson, ...
Stig Johansson (1939-2010)
Professor of English linguistics at the University of Oslo
Responsible for the LOB corpus, and the ENPC corpus
Founder of ICAME
Lauri Carlson (1952-)
Professor of Linguistics and Translation, University of Helsinki
Game theory, semantics, machine translation, parsing, terminology
Happy ending? Or just the beginning?

Lisbon, 17 January 1997
My picture of two languages

Yellow language
Gives more attention to X, less attention to Y, does not mention Z

Blue language
A book worth reading

• Every language is a particular system of classification

• Communication requires thinking

• ‘good’ is more basic than ‘triangle’

• Categorization is a way to join different things
Another one

• What does it mean to understand?

• Language has both form and content, formal and content restrictions

• Translation has to take both in consideration
Lexical “overlap”

Jurafsky & Martin (2000:806)
Trying to make sense of language differences

• How to indicate the relationship of meaning “nuggets” in different languages?

• Snell-Hornby (1983) on the translation of German von regem Geschäftsstreibern erfällt

• Verb descriptivity: verbs have two parts of meaning
How can two sentences be translations of each other?

• … a SL and a TL text or item [being] relatable to (at least some of) the same features of substance (Catford, 1967:50)

I → speaker
female
arrival
on foot
have arrived

ja
prior event
linked to present
completed

prišla
Catford (1967:39)
The translation network model

- Independent description of two languages
- Bridges between different categories
Which figure is correct? (Santos, 1998)

Models of machine translation
The traditional (wrong) view
The proper view

Língua fonte

Análise

Transferência

Interlíngua

Directa

Língua alvo

Geração
Temporal duration: time categories

- *Kino’s people had sung of everything that happened or existed.*

- *A gente de Kino cantara tudo o que acontecera ou existira*

<table>
<thead>
<tr>
<th>past</th>
<th>present</th>
</tr>
</thead>
</table>

- *He was trapped as his people were always trapped*

- *Estava peado, como todos os da sua raça sempre tinham estado*
The culture dependence of a picture

Collections: captions are essential to give unity. Captions (small sized pieces of text) are extremely difficult to translate.
The cultural dependence of captions

- *Man reading.* This can be a good enough caption in an European museological context, but certainly not in an Asian, or African context
The language dependence of illustration

- Vi roser Anne!

- *Amor perfeito*  
  (means perfect love in Portuguese)

- *Stemorblomst*  
  (stepmother’s flower in Norwegian)
Same concept: Foxes and blue

- Feminine, related to love and friendship
- Perfect: *Ouro sobre azul*

- Masculine, related to tricks
- Depression: the blues
Colour differences in COMPARA

• Different metaphors:

  *sorriso amarelo* -> *wan smile*; *romance cor de rosa* -> ?; *blue movies* -> *filmes tristes*; *nódoas negras* -> *bruises*; *armas brancas* -> *knives*; *fazer a vida negra* -> *give a hard time*; *red herring*; *paint the town red* -> ?; *brown off* -> *maçar-se*

• Different cultures:

  *black coffee* -> *café*, *red meat* -> *carne mal passada*; *correio azul* -> *first class stamp*; *red demands* -> *intimações*; *red tape*; *Black Maria* -> *ramona*; *red-light district*;

• Vagueness: *golden* -> *dourado*, *or de ouro*, *de prata*
Colour differences attested in COMPARA

• Translator creativity:

  putty-coloured -> cor de massa de vidraceiro; civic redbrick -> novas e sem tradição: azuleleca -> tricklight

• Different conventions:

  brown paper -> papel pardo, goldfish -> peixe vermelho; claras -> egg whites; página em branco -> blank page; dark purposes -> negros propósitos;
Differences between languages

(1) “I want a different apple.” “Why? They are all the same.”

(2) They wore the same dress.

(3) I’ll have the same as her (said to a waiter).

(4) These two pens look similar, but one is more expensive than the other

• English same is ambiguous between type and token identity
• Finnish: not the same item in (1) nor (2), but in (3).
• Portuguese: not the same item in (1): são todas iguais
• Portuguese: parecem iguais in (4)
Same or similar revisited

- Similarity is relative, variable, culture dependant (Goodman, 1972); Circumstances alter similarities (Goodman, 1972); The similarity of objects is modified by the manner in which they are classified (Tversky, 1977); “similarity” is a sign that is attributed to a set of entities, attributed by someone and also interpreted by someone (Chesterman, 1998) (similarity-as-trigger vs. similarity-as-attribution); the greater the extension of the set of items assessed as being similar, the less the pertinent degree of similarity; Tension between “oneness” and “separate individuation” (Sovran, 1992)
Linguistic-cultural infrastructure for contrastive studies

Access to a lot of material fine-grained annotated that can be browsed and further investigated

1) Processing large amounts of data

2) Doing fine grained analysis on situated utterances in context
I want to avoid…

• Uninspiring counting of dubious “hits”: so what?

• Universal theories based on one (!) example

• Empirically-based theories cannot be proved right

• Deductive arguments cannot be proved wrong
Properties that define a natural language as opposed to artificial ones

1. Metaphorical nature
2. Context dependency
3. Reference to implicit knowledge
4. Vagueness
5. Dynamic character (evolution and learnability)

Slide 12 from Santos (2006)
Corpus size, AC/DC cluster

• 21 corpora, 295 million tokens (words and punc)
• Variety:
  224,254,595 PT
  65,671,800 BR
• Genre: 252 millions newspaper, 17 millions fiction, 4.5 millions technical
• When full articles: 245,490 articles
• Sentences: 12,639,914
Corpus size, AC/DC cluster

- Different wordforms: 1,435,045
- Different lemmas (excluding MWE): 872,691
- Different verb lemmas: 76,012
- Different verb forms: 333,937
- Different colour words: 2,642
- Different colour lemmas: 1,173
What does colour annotation mean?

• First we select the colour vocabulary

• Then we annotate based on that, and look at the result

• Then we create fine-grained categories and rules
  – To remove wrong cases
  – To add rare cases
  – To deal with multiword expressions

• Then we comb every case to get to 100% precision (and hopefully 100% recall as well)
What do we learn about language if we do semantic annotation?

• There are always vague cases

• There are cases where people have a hard time to pin down the meaning/classification

• Languages change, and what is “rigid”, collocation or fixed?

• Is metaphorical use, and terminological use, the same thing? Are literary and technical genres different when we come to the meaning of words?
Examples of “fine-grained” categories

• Colour as race
• Metaphorical colour
  – When coloured things become symbols for one category: cartão amarelo, luz verde
  – When a colour represents a state of mind or a moral judgement:
  – When a coloured expression takes on another meaning: red herring, black tie
• Metonymical colour: politics, sports
• Conventional colour: white wine, gullfisk
• Absence of colour; indefiniteness of colour; multiple colours
All sorts of categories

• From specific concepts at the word level (lexical semantics) to part-of-speech (what is an adjective, what is a noun, what is a verb), to discourse (what is a clause, what is a sentence) to interaction (what is a conversation, a turn) to all sorts of “linguistic features”

• For example: foreign words
  – is a marker of sloppy discourse in Italian (Santini)
  – is a marker of educated discourse in English (Biber)
Depending on the language...

The same category for

- Hunger, anger, missing someone, pain: *Estou cheia de fome, de raiva, de saudades, de dores*

- “feeling about the future”: *grue seg til, glede seg til*

Different category for

- *Puxar o autoclismo, trekke ned (på do)*
Different metaphors

- *Estar sem pé*: not attaining the ground in water
- *Ao pé de*: near
- *Um pé de vento*: a sudden and wild breeze
- *Um pé de hortelã*: a plant that can reproduce
- *À mão*: at hand

- Bare bra! Não tenho nada!
Concluding remarks

- It is hardly to be found ONE distinction that is common across all natural languages
- Languages tend to evolve and age and innovate continuously
- The comparison of languages is arguably the best mirror into language ... and the comparison itself is best done through translation data
Concluding remarks (cont.)

• Words carve different domains in different languages, words are different in different languages, the differences between inter-translatable words (and not only) are a wonderful mirror to differences in systematic organization of the languages (systematicity includes creativity)
Questions
& Comments