An Uphill Battle: Achieving pragmatic congruency in multilingual texts

Donia Scott

Multilinguality: an early goal of AI/CL

- Using computers to aid in the process of producing different language versions of the same document.

- Strong practical imperatives:
  - Legal requirements (e.g., parliamentary procedures, technical manuals, pharma leaflets)
  - Gisting, translating web pages, Facebook comments ...
  - Spoken translation, subtitling

- Interesting theoretical issues

Text quality and multilingual equivalents

Quality language versions convey the same meaning in a fluent manner that is appropriate to the

- Genre
- Register
- Discourse context

These affect text in a myriad of ways, including

- Lexical choice
- Syntax
- Use of discourse connectives

Equivalence in multilingual documents

- Same message, different form

    Semantic content is constant but delivery will vary widely, depending on language, genre, register and discourse context.

- Multilinguality as *paraphrase*, not *imitation*

Same semantics, different discourse relations, different discourse connectives, different syntax

generation   (loosen locking lever, alter stepping load)

🇩🇪 *Nach* Lockern der Klemmhebel kaan die Tretbelastung verändert werden.

'<u>After</u> (the) loosening of the locking lever the stepping load can be altered.'

SEQUENCE

🇬🇧 The stepping load can be altered <u>*by*</u> loosening the locking lever.

MEANS

🇫🇷 *Pour* modifier la change d'appui, desserrer le levier.

'<u>In order to</u> alter the stepping load, loosen the locking lever.'

PURPOSE

The bottom line

Good multilingual versions must involve considerations of discourse, since this drives choices at almost all levels.

Semantics → Discourse

Discourse →

Discourse relation
↓ ↓ ↓
Discourse connective
↓ ↓ ↓
Syntax

....

Discourse relation
↓ ↓ ↓
Discourse connective
↓ ↓ ↓
Syntax
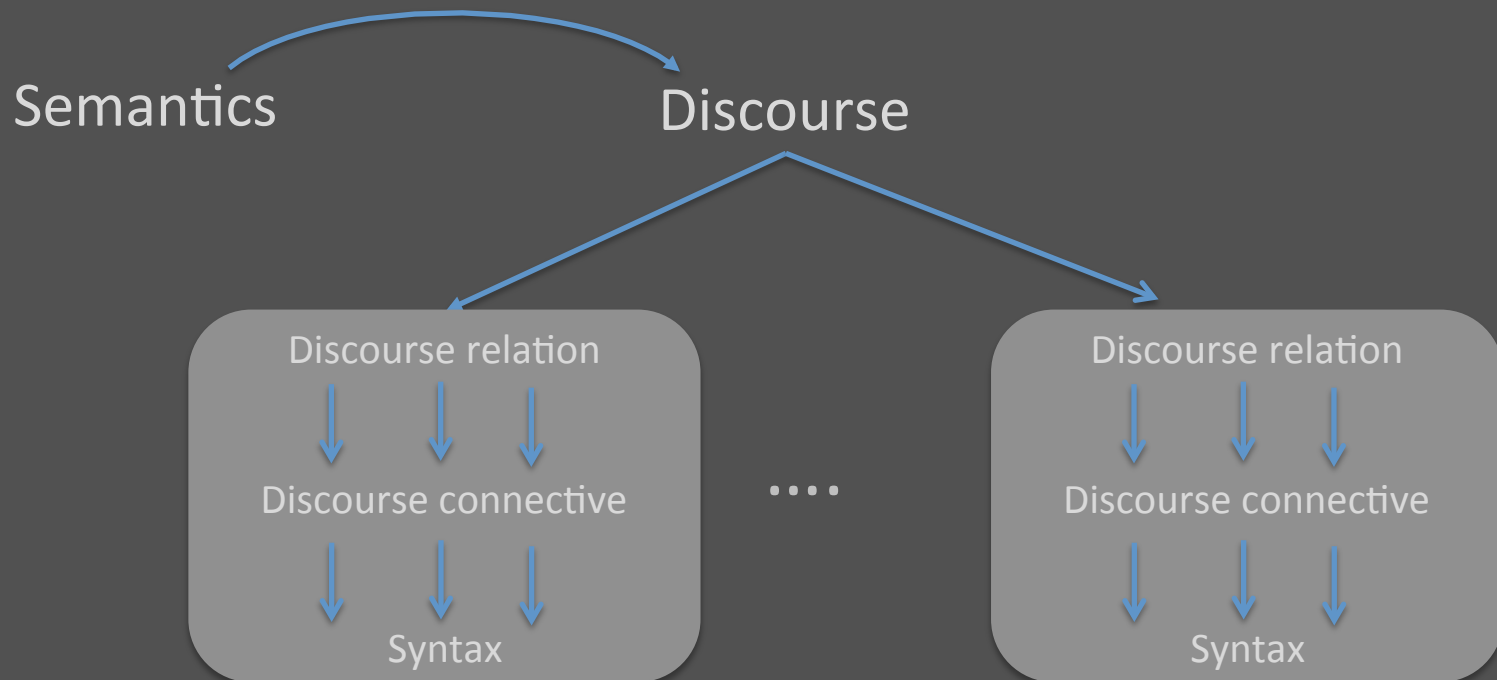
Think globally, act locally…

- Corpus-based studies on instructional texts show that

    - discourse preferences (and thus syntactic and lexical choices), within a given genre, vary across languages.

    - Achieving a satisfactory level of congruence requires the use of syntactic and pragmatic features appropriate to each language version with respect to a single underlying task model.

        (Delin et al, 1996; Grote, 1995; Scott, Delin and Hartley,1998)

- Other points to note:

    - discourse-based dependencies can operate over wide spans of text (e.g., Mann and Thompson, 1998)

    - semantically equivalent discourse structures can show quite different organisations (e.g., Scott & Souza, 1990; Knott, 1996)

The situation up to now …

- Multilingual versions are produced by MT, in particular, SMT, which does not rely on strong linguistic models, where decisions are made locally, and where the sentence is the maximal unit for both training and decoding.

- Long-distance effects cannot be properly dealt with, e.g., anaphora, discourse connectives, lexical choice.

- Text quality, in this context, is often considered to be how close the target sentence is to an *imitation* of the source sentence.

Some hope …

- Through renewed interest in
    - Conducting experiments that explore how and where language versions produced by machines *vs* humans vary wrt discourse-related phenomena, such as connectives and anaphora (e.g., Meyer and Webber, 2013; Li et al, 2014)

    - Designing decoders with feature models that have access to the complete discourse context (e.g., Stymne et al, 2013; Hardmeier et al, 2012; Zhang and Ittycheriah, 2015)

    - Extending traditional evaluation metrics to include discourse features (e.g., Wong and Kit, 2002; Guzman et al, 2014)

    - Designing quality estimation algorithms for MT that consider the document level (Scarton and Specia, 2014)
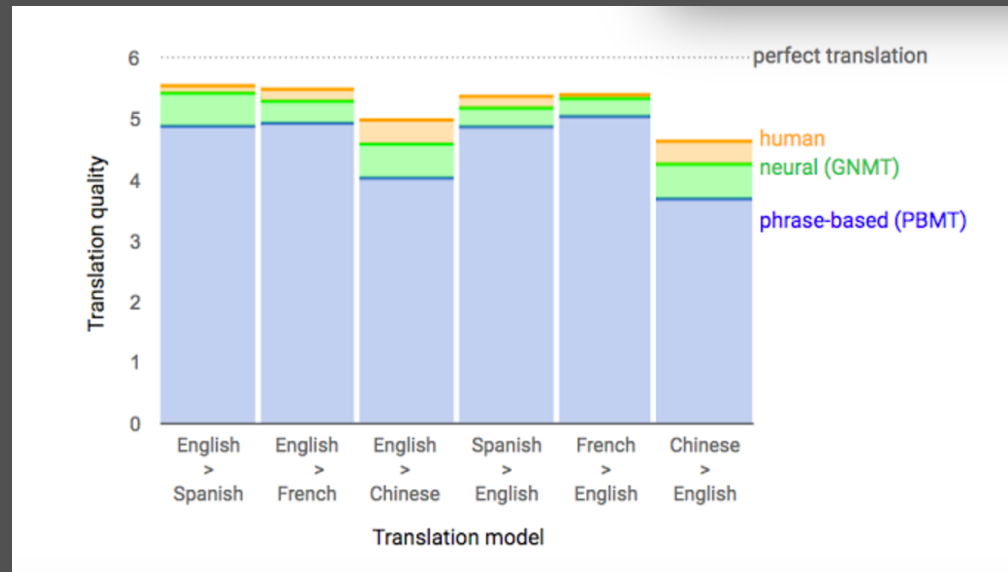
Hope ?

Move away from phrase-based SMT to NN-MT

- easier to maintain
- better adaptability over domains and languages
- *achieves close to human quality translation*

BUT
still focuses on individual sentences.

e.g., DFKI character-based neural machine translation system; newly-launched version of GoogleTranslate

Time to resurrect Multilingual NLG??

NLG engines for
selected languages

language
versions

Document
semantics