

COLIBRI: Constructions as Linguistic Bridges

Maarten van Gompel, Radboud University Nijmegen

FLARN - Tilburg University - March 2012



Introduction

Research objective

Research objective: Identification and extraction of “good” constructions in parallel corpus data.

About the research

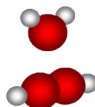
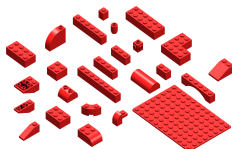
- **Implicit Linguistics:** Deducing linguistic knowledge from corpus data
- Beyond n-gram models
- What is “good”?

Constructions

What is a construction?

What is a construction?

- A *pattern* of words, not necessarily consecutive, which *in some way* forms an entity
- Constructions emerge from the data rather than linguistic theory
- Intuitive “building blocks” for various NLP tasks
- A level above n-gram models and below syntactic level.



Emerging constructions

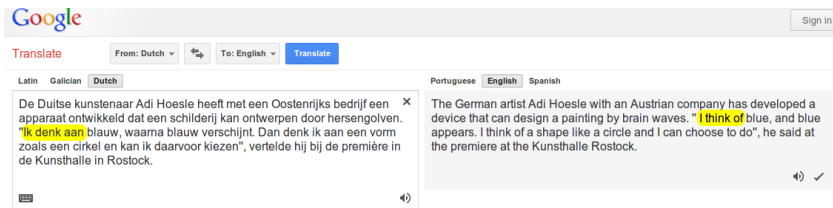
Constructions emerge from the corpus data rather than from linguistic theory:

- 1 by frequency
- 2 by context
- 3 **by multilingual alignment**

Counting patterns

- Iterative counting, first unigrams, then bigrams
- ... discarding pattern candidates if a sub-part was not found
- Counting constructions with gaps: “*to be * to be*”
- ...by punching all combinations “holes” in found consecutive constructions

Phrases as constructions in Machine Translation



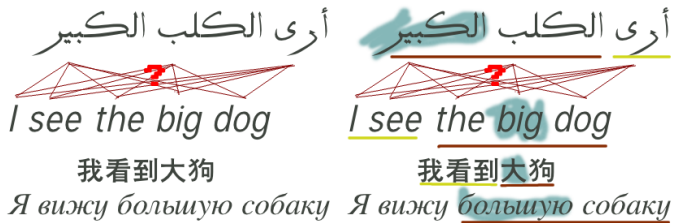
The screenshot shows the Google Translate interface. At the top is the Google logo and a 'Sign In' button. Below the logo, the 'Translate' section is active, with 'From: Dutch' and 'To: English' selected. A blue 'Translate' button is visible. The input text in Dutch is: 'De Duitse kunstenaar Adi Hoesle heeft met een Oostenrijks bedrijf een apparaat ontwikkeld dat een schilderij kan ontwerpen door hersengolven. "Ik denk aan blauw, waarna blauw verschijnt. Dan denk ik aan een vorm zoals een cirkel en kan ik daarvoor kiezen", vertelde hij bij de première in de Kunsthalle in Rostock.' The output text in English is: 'The German artist Adi Hoesle with an Austrian company has developed a device that can design a painting by brain waves. "I think of blue, and blue appears. I think of a shape like a circle and I can choose to do", he said at the premiere at the Kunsthalle Rostock.' The phrase 'I think of blue' is highlighted in yellow in both the Dutch and English versions. The interface also includes language selection tabs for Latin, Galician, Dutch, Portuguese, English, and Spanish, and a small speaker icon for audio playback.

Phrases as constructions in Machine Translation

Google Translate interface showing a Dutch sentence and its English translation. The Dutch sentence is: "De Duitse kunstenaar Adi Hoesle heeft met een Oostenrijks bedrijf een apparaat ontwikkeld dat een schilderij kan ontwerpen door hersengolven. "Ik denk aan blauw, waarna blauw verschijnt. Dan denk ik aan een vorm zoals een cirkel en kan ik daarvoor kiezen", vertelde hii hii de première in de Kunsthalle in Rostock". The English translation is: "The German artist Adi Hoesle with has developed a device that can waves. "I think of blue, and blue appears. I think of a shape like a circle and I can choose to do'', he sa premiere at the Kunsthalle Rostock." A tooltip is visible over the English text "I think of a shape" with the text "Drag with shift key to reorder." and a list of suggestions: "Then I think of", "Then I think", "I think about", and "I am thinking of".

How are these relations accross languages found?

How? *Measures of co-occurrence*



Example

'Uraa al-kalba al-kabira	I see the big dog
'Uraa al-qitta al-saghira	I see the small cat
'Uraa al-qitta al-kabira	I see the big cat
akala al-rajul	The man ate
Yuhabbu al-rajul al-qitta	The man loves the cat

Try!

'uraa?

al-kalba?

al-qitta al-saghira?

al-rajul?

yuhabbu?

Excerpts of automatic alignment:

als lid	as a member	0.09
wil benadrukken dat	want to emphasise that	0.0493827
vrije verkeer van personen	the free movement of persons	0.140625
tijdschema	timetable	0.237954
Dat wil zeggen dat	This means that the	0.0277778
vertrouwen van de	consumer confidence	0.16
één enkele	single Member State	0.0277778
oplossing van het	the conflict	0.0219479
goed voorstel	to get this	0.01
Er bestaan	of mobile	0.0123457

MT Approaches

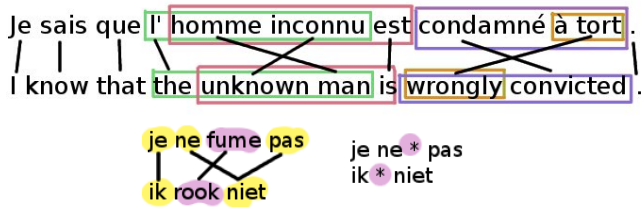
- 1 **Rule-based:** explicit linguistic knowledge
- 2 **Data-driven:** implicit linguistics
 - Statistical/machine learning models

Two important aspects of translation

- 1 Faithful conservation of meaning: good mapping between constructions
- 2 Fluent natural style: natural word-order

What are the units of translation? → constructions

- Whole sentences at once? No
- Single words, word by word? No
- Words in context? Better
- Variable-length phrases in context? Even better
- Loosening constraints: “constructions” in context? Best?



Relations

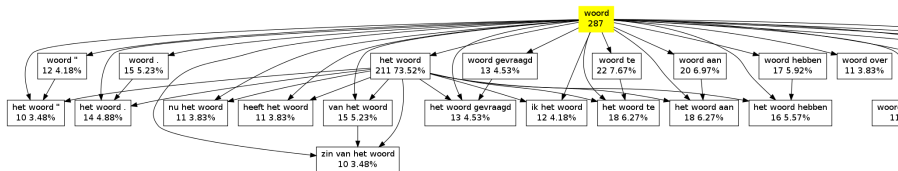
- Relations between constructions from different languages
- Relations within constructions of the same language

Example

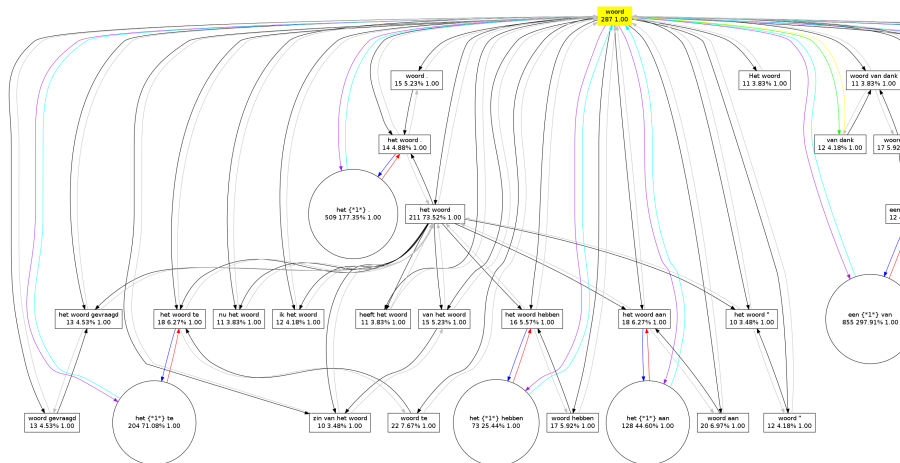
“I see the dog move”

- **Subsumption:** “I” is a sub-part of “I see”
- **Succession:** “see” is a successor of “I”
- **Instantiation:** “I see the dog move” instance of “I see * move”

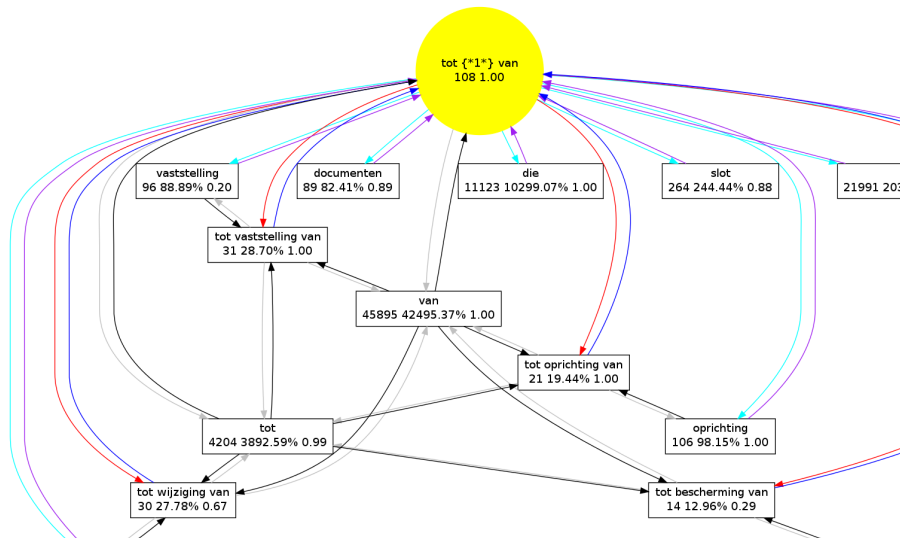
Subsumption relations



All relations: "woord"



All relations: “tot * van”



Hypotheses

- Constructions can be found efficiently in corpus data
- Graph-based relations can be used to constrain to “good” constructions
- Constructions can be aligned without resorting to word-alignments as a basis
- In MT, constructions (i.e. possibly with gaps) result in better translation than mere consecutive phrases

Empirical Evaluation

- Evaluation of constructions in an MT setting
- Evaluation based on translation quality
 - Comparison to human reference translations
- Alternative use-case: Constructions in Language Modelling

Use in linguistic studies

- Abstracting fully lexicalised constructions
 - Finding semantic subclasses in constructions: “from *time-expression* to *time-expression*”
 - Collapsing constructions with word disjunctions (“he/she/it”) or part-of-speech tags
- Correlations with experimental findings
 - Switch tasks, cloze tests, reaction times, ...
- Modelling multilinguality
 - Parallel constructions
- Software package for linguists?



Questions?