

Fourth CLIF Symposium

9 December 2009

Universiteit Antwerpen

**Alignment of grammatically divergent parses  
using interlingual MT techniques**

Tom Vanallemeersch

Lessius / KU Leuven

Centre for Computational Linguistics

# Research question

Can interlingual MT techniques improve the alignment of grammatical divergences ?

# Background (1/6)

Translational divergences (Appelo 1993, Dorr 1994):

- Structural source: general syntactic differences: *do*
- Lexical source:
  - Conflational: *zich voordoen, occur*
  - Categorical: *during their meeting, terwijl ze vergaderen*
  - Differences on conceptual level: *runway, startbaan*
- Tense, aspect, voice, mood (words, affixes): *will work, travaillera*
- Extralinguistic: target audience (Nord 1988)

## Background (2/6)

Human translation:

- Translator education on specific structures
- Example: Dutch infinitives into French (Van Baardewijk-Rességuier and Van Willigen-Sinemus 1986):

*Het berekenen van cijfers is moeilijk*

→ *La calculation de chiffres est difficile*

→ *Il est difficile de calculer des chiffres*

# Background (3/6)

## Machine Translation:

- RBMT (Van Eynde 1993):
  - Anticipation/readjustment: word order, ...
  - Normalisation: canonical form, lexical differentiation
  - Abstraction: apt for >2 languages (closed classes)
- SMT: anticipation (Yamada and Knight 2001)
- Generation-heavy MT (Vandeghinste 2008)
- Intertwining mono/bilingual steps (M.T. Rosetta 1994)

## Background (4/6)

Human alignment of translated sentences:

- Alignment guidelines (Melamed 1998)
- Resource for translation shifts (Cyrus 2006)
- Example:

En premier lieu, cette loi commerciale viserait des projets temporaires.

Deze handelswet zou nadrukkelijk doelen op tijdelijke projecten.

# Background (5/6)

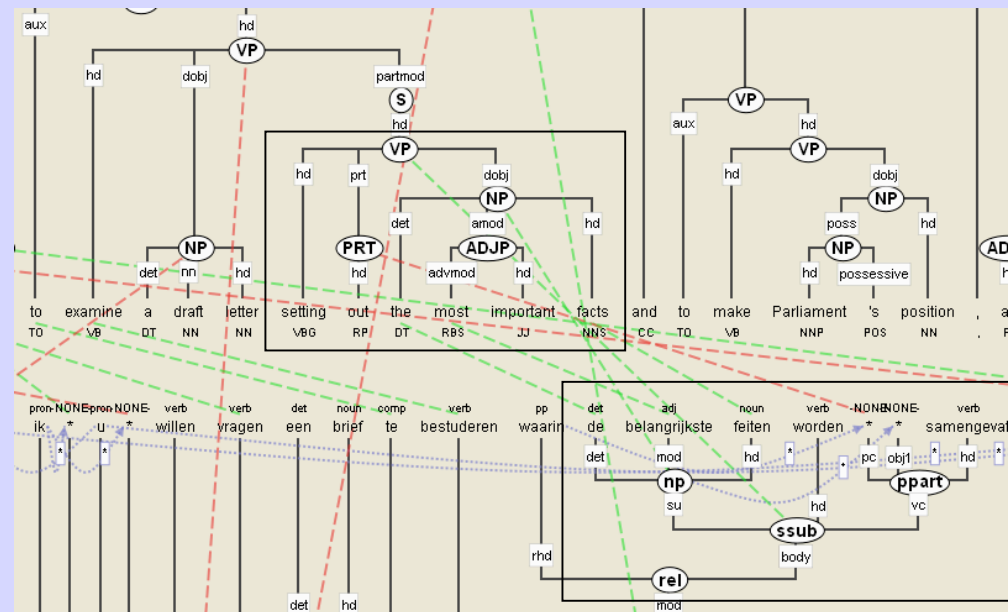
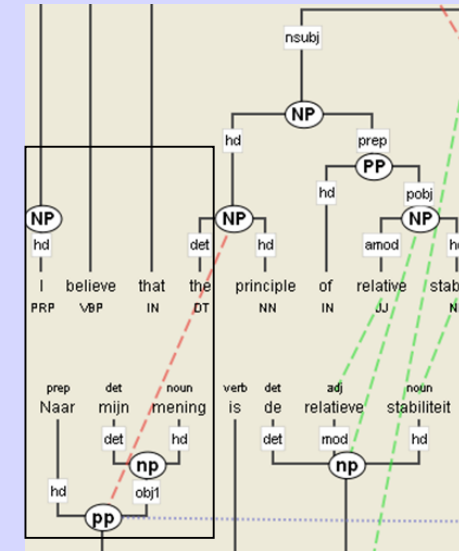
Automated alignment:

- Level: paragraphs, sentences, words, trees, ...
- Tree alignment:
  - Connect corresponding (non-)terminal nodes
  - DOT (Hearne 2005), PaCo-MT project (CCL)
- Supervised method Tiedemann and Kotzé (2009):
  - Theory-independent (labelling, ...)
  - Discriminative approach: lexical/positional equivalence, subtree level/span, node labels, context

# Background (6/6)

Problems with this method:

- Divergences
- Different tree topology (non-isomorphism)
- Partial alignment



# Approach (1/5)

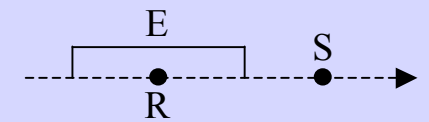
Create semantic hypotheses for subtree alignment:

- Based on interlingual MT system Eurotra (Allegranza et. al. 1991):  
constituent → relational → semantic structure
- Abstraction strategy for grammatical elements (closed classes)
- Focus on clauses and nominalizations:  
hypothesis = verb + semantic tense, aspect, subject, object

## Approach (2/5)

Tense/aspect *forms* → *meanings* (Van Eynde 1991):

- Represent meaning through timeline



“He was eating an apple”

- Example:

- *Il vit à Paris depuis 1968*

Form: présent + simple

Meaning: **simultaneous**/posterior + perfective/durative/**terminative**

- *He has lived in Paris since 1968*

Form: present + perfect

Meaning: **simultaneous** + **terminative**/retrospective

- Adverbials (disambiguation): *depuis* = terminative

## Approach (3/5)

Syntactic functions → semantic roles:

- Diathesis: active/passive voice, impersonal forms
- Example:

*Er wordt gedanst, on danse, there is dancing*

→ verb = “dance”, semantic subject = empty

## Approach (4/5)

Nominalization → verb + semantic roles:

- Deverbal noun expressing an action: *rejection of*
- Infinitival nominal: *het eten van de appel*
- Nominal gerund: *I like her singing of that song*  
(vs. *Her singing that song yesterday surprised me*)
- Ambiguity of preposition:  
*de groei van de markt*  
→ *markt* = semantic subject, not object

# Approach (5/5)

Alignment of semantic hypotheses:

- Through semantic properties
- Through probabilistic bilingual lexicon
- Compute best match

# Application (1/2)

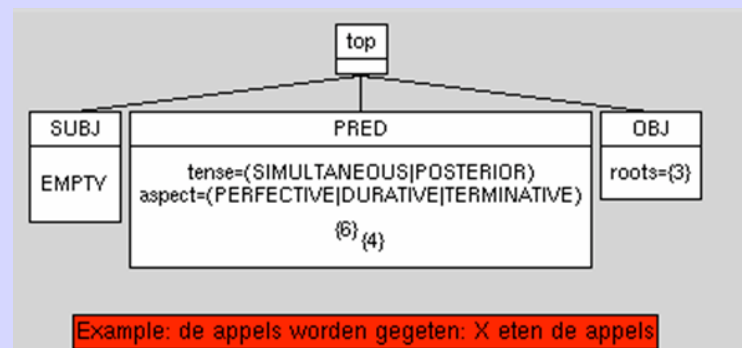
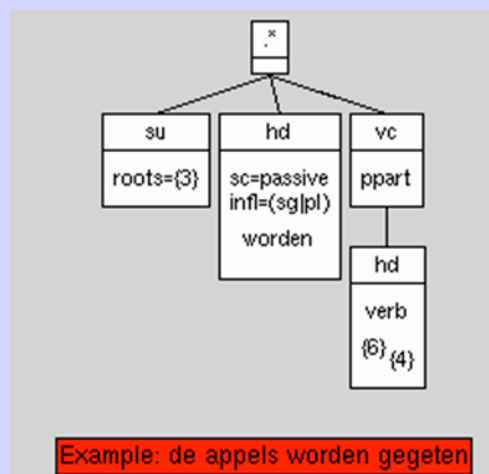
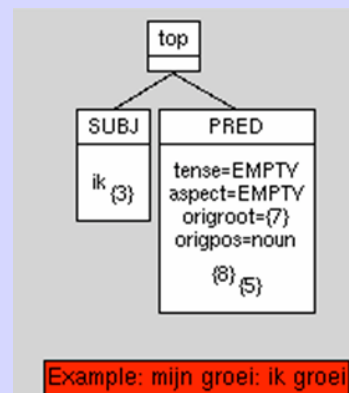
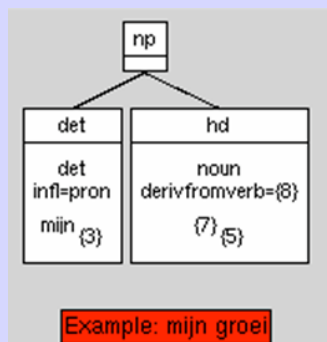
- Corpus: Europarl (Koehn et al. 2005)
- Parses:
  - PaCo-MT project (Vandeghinste and Martens 2009):
    - English: Stanford parser (Klein and Manning 2003)
    - Dutch: Alpino parser (van Noord 2006)
    - French: Malt parser (Nivre et al. 2007), trained on French Treebank (Abeillé et al. 2003)
  - French Passage corpus (de la Clergerie et al. 2008):  
combined parser output
- Pairs deverbal noun - verb: heuristically extracted from CELEX and Multext (Véronis 1998)

## Application (2/2)

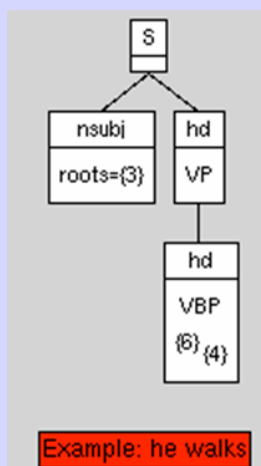
Creation of patterns with associated hypotheses:

- Verify representation of clauses/nominalizations in parse trees
- Create patterns matching subtrees
- Patterns and hypotheses contain placeholders whose values are unified during matching

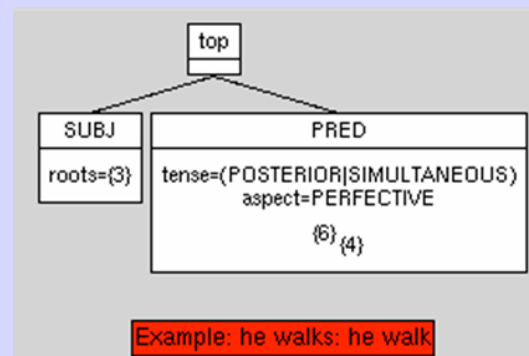
# Examples: patterns (1/2)



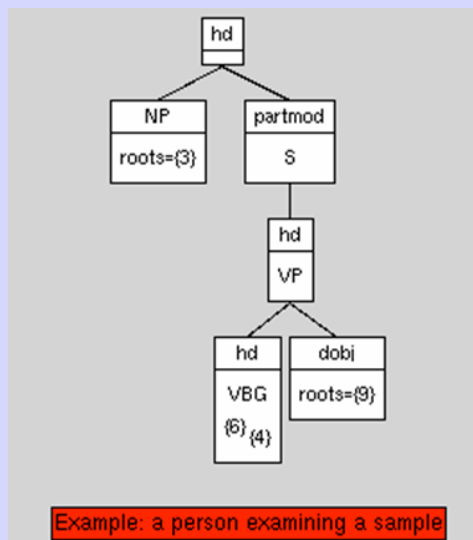
# Examples: patterns (2/2)



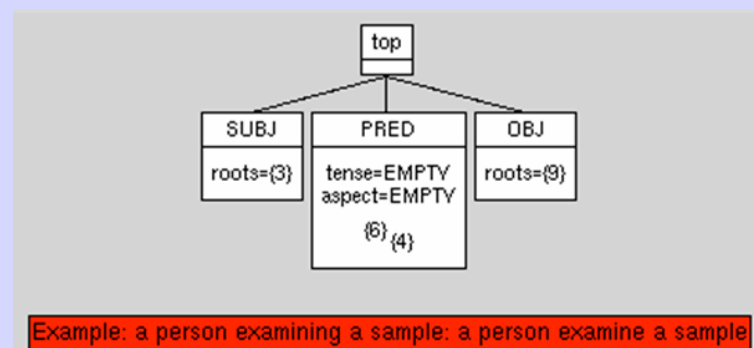
Example: he walks



Example: he walks: he walk

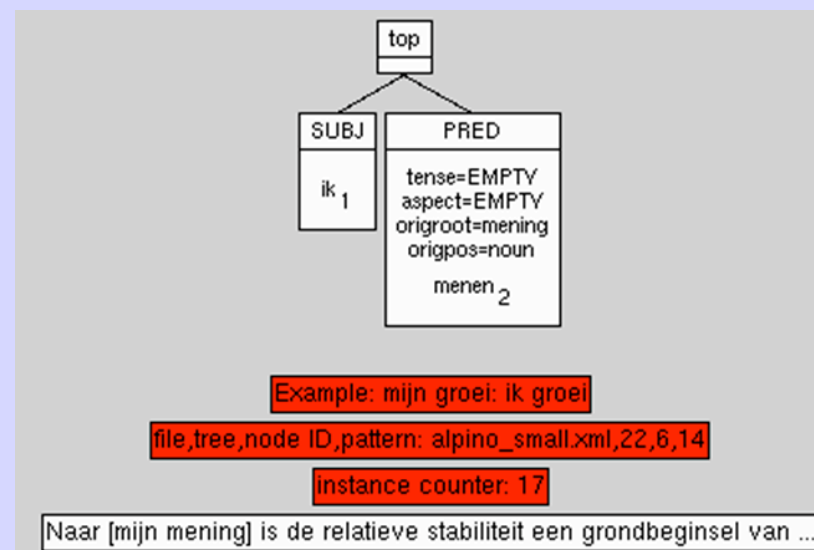
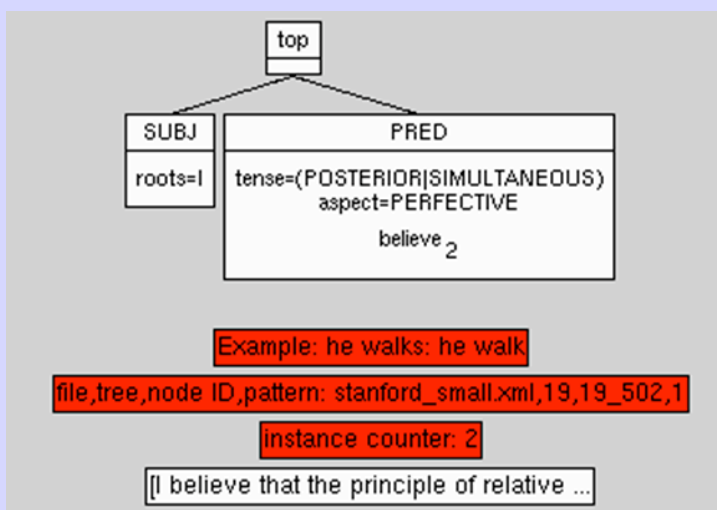


Example: a person examining a sample

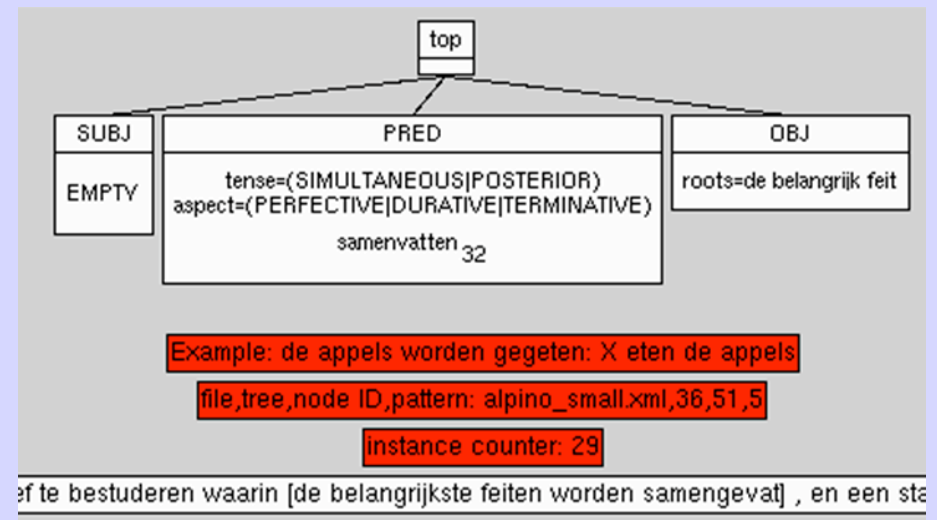
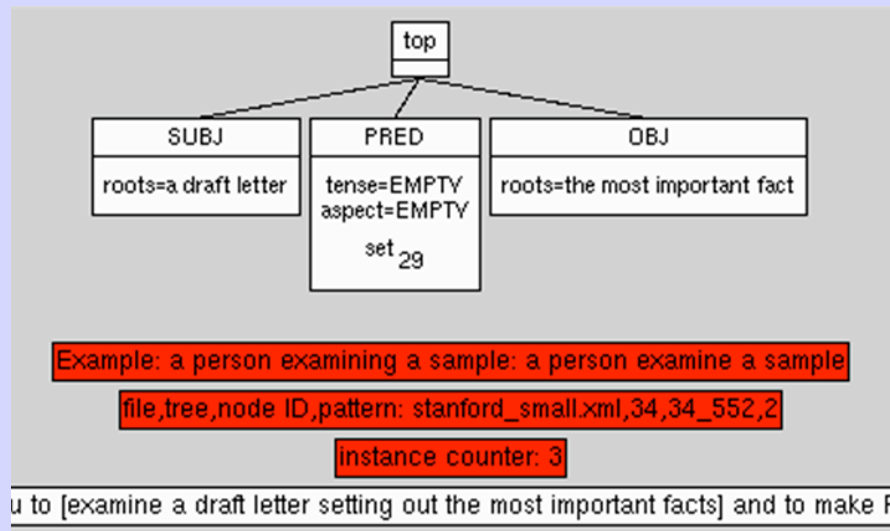


Example: a person examining a sample: a person examine a sample

# Examples: matches (1/2)



# Examples: matches (2/2)



## Future work (1/3)

- Create reference corpus with patterns to be covered
- Compare with tree alignment of Tiedemann and Kotzé (2009)
- Secondary research questions:
  - What monolingual information do we need for applying the interlingual MT approach ?
  - Can this approach be coarser for alignment than for MT (MT needs information for generation) ?

## Future work (2/3)

Improve probabilistic lexicon:

- **FragmALex**: alignment of words, word groups/parts based on bilingual lexicon (Vanallemeersch and Wermuth 2008)
- French-Dutch sentence-aligned corpus “Belgisch Staatsblad”: 2.4 million unique sentence pairs (abstract submitted to LREC 2010)
- Detection of inconsistent translation of phraseology using term extractor / SMT (Vanallemeersch and Kockaert 2010, submitted to SALALS)

## Future work (3/3)

Europarl:

- Create subcorpora English-Dutch, French-Dutch, English-French (+ vice versa), checking original language
- Cfr. Van Halteren (2008)

# References (1/2)

- Abeillé, A., L. Clément & F. Toussnel (2003). Building a Treebank for French. In A. Abeillé (ed) *Treebanks*, Kluwer, Dordrecht.
- Allegranza, V., P. Bennett, J. Durand, F. Van Eynde, L. Humphreys, P. Schmidt & E. Steiner (1991). Linguistics for Machine Translation: The Eurotra Linguistic Specifications. In C. Copeland, J. Durand, S. Krauwer & B. Maegaard (eds) *The Eurotra Linguistic Specifications*, Office for Official Publications of the Commission of the European Community, Luxembourg, pp. 15-123.
- Appelo, L. (1993). Categorical Divergences in a Compositional Translation System, Ph.D. Thesis, University of Utrecht.
- Cyrus, L. (2006). Building a resource for studying translation shifts. In *Proceedings LREC 2006*, pp. 697-702.
- de la Clergerie, E., O. Hamon, D. Mostefa, C. Ayache, P. Paroubek & A. Vilnat (2008). PASSAGE: from French Parser Evaluation to Large Sized Treebank. In *Proceedings LREC 2008*.
- Dorr, B. (1994). Machine Translation Divergences: A Formal Description and Proposed Solution. In *Computational Linguistics*, 20(4), pp. 597-633.
- Hearne, M. (2005). Data-Oriented Models of Parsing and Translation. Ph.D. Thesis, Dublin City University.
- Klein, D. & C. D. Manning (2003). Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, Cambridge, MA: MIT Press, pp. 3-10.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit 2005*, pp. 79-86.
- Melamed, D. (1998). Annotation Style Guide for the Blinker Project, IRCS Technical Report #98-06.
- M.T. Rosetta (1994). Compositional Translation. Kluwer, Dordrecht.
- Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryiğit, S. Kübler, S. Marinov & E. Marsi. (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. In *Natural Language Engineering*, 13(2), pp. 95-135.
- Nord, C. (1988). Textanalyse und Übersetzen. Theoretische Grundlagen, Methode und didaktische Anwendung einer übersetzungsrelevanten Textanalyse, Heidelberg: Groos 1988, 2. neu bearb. Auflage 1991, 3. Aufl. 1995.
- Tiedemann, J. & G. Kotzé (2009). A Discriminative Approach to Tree Alignment. In *Proceedings of RANLP*.

# References (2/2)

- Vanallemeersch, T. & C. Wermuth (2008). Linguistics-based Word Alignment for Medical Translators. In *Journal of Specialized Translation* (Jostrans), nr. 9.
- Vanallemeersch, T. & H. J. Kockaert (2010). Automated Detection of Inconsistent Phraseology Translation, submitted to Southern African Linguistics and Applied Language Studies (SALALS).
- Van Baardewijk-Rességuier, J. & M. Van Willigen-Sinemus (1986). Matériaux pour la traduction du néerlandais en français. Muiderberg: Coutinho.
- Vandeghinste, V. (2008). A Hybrid Modular Machine Translation System. LoRe-MT: Low Resources Machine Translation. Ph.D. Thesis. LOT, Utrecht, The Netherlands.
- Vandeghinste, V. & S. Martens (2009). Top-down Transfer in Example-based MT. In *Proceedings of the 3rd Workshop on Example-based Machine Translation*, pp. 69-76.
- Van Eynde, F. (1991). The Semantics of Tense and Aspect. In: M. Filgueiras, L. Damas, N. Moreira & A.P. Tomás (eds.), *Natural Language Processing*. Lecture Notes in Artificial Intelligence. Volume 476. Springer-Verlag, Berlin, pp. 158-184.
- Van Eynde, F. (1993). Machine Translation and Linguistic Motivation. In F. Van Eynde (ed.) *Linguistic Issues in Machine Translation*, Pinter, London, pp. 1-43.
- Van Halteren, H. (2008). Source language markers in Europarl translations. In *Proceedings COLING 2008*, pp. 937-944.
- van Noord, G. (2006). At Last Parsing Is Now Operational. In: Piet Mertens, Cedrick Fairon, Anne Dister, Patrick Watrin (eds): *TALN06. Verbum Ex Machina*. Actes de la 13e conference sur le traitement automatique des langues naturelles, pp. 20-42.
- Véronis, J. (1998). Multext-Lexicons. A set of Electronic Lexicons for European Languages. CD-ROM, distributed by ELRA/ELDA.
- Yamada, K. & K. Knight (2001). A syntax-based statistical translation model. In *Proceedings ACL 2001*.