The PROIEL parallel corpus of old Indo-European translations of the New Testament

Marius L. Jøhndal  Hanne M. Eckhoff  Dag Haug

University of Oslo

6 December 2008
Goals

▶ Contrastive study of the syntax of early IE languages
  ▶ Word order
  ▶ Pronouns
  ▶ Participles
  ▶ Discourse particles
  ▶ Definiteness
Annotation levels

- Morphology and lemmatisation
- Syntax
- Information structure
- Alignment
Theory choice

- Enriched dependency grammar
- Empty nodes for
  - asyndetic coordination,
  - verb gapping,
  - some pro-drop elements
- Secondary dependency relations (’slashes’)

![Diagram showing enriched dependency grammar and secondary dependency relations]
Alignment levels

- Text division alignment
Alignment levels

- Text division alignment
- Sentence alignment

The Greek text is shown alongside the corresponding English translations.
Alignment levels

- Text division alignment
- Sentence alignment
- Token alignment
Alignment levels

- Text division alignment
- Sentence alignment
- Token alignment
- Syntactic alignment
Sentence alignment

- Sentence alignment can be done automatically.
- The automatic alignment can be corrected iteratively by
  - 'anchoring' sentences or
  - 'blacklisting' sentences.
Sentence alignment

- Sentence alignment can be done automatically.
- The automatic alignment can be corrected iteratively by
  - 'anchoring' sentences or
  - 'blacklisting' sentences.

103 24 καὶ οἱ Φαρισαῖοι ἔλεγον αὐτῷ.

104 24 ἱδε τί ποιοῦσιν τοῖς σάββασιν δ οὐκ ἔξεστιν; 105 25 καὶ λέγει αὐτοῖς.

106 25 οὐδέποτε ἀνέγνωτε τί ἐποίησεν Δαυείδ, ὅτε χρείαν ἔσχεν καὶ ἐπείνασεν αὐτός καὶ οἱ μετ’ αὐτοῦ;

154 24 ἡ δὲ ἐπιστολὴ ἅπλη εἰς Μακαρίου.

155 24 виждь что творать въ сабботы егоже не достоитъ.

156 25 ἔγραψεν τῷ ὑπερήφανῳ.
Sentence alignment can be done automatically. The automatic alignment can be corrected iteratively by
- 'anchoring' sentences or
- 'blacklisting' sentences.

103 24 καὶ οἱ Φαρισαῖοι ἔλεγον αὐτῷ·

104 24 ἠδὲ τί ποιοῦσιν τοῖς σάββασιν δόσιν ἐξεστὶν;

105 25 καὶ λέγει αὐτοῖς·

106 25 οὔδέποτε ἀνέγυμπτε τί ἐποίησεν Δαυείδ, ὡς χρείαν

154 24 καὶ ψαρικῇ ἰδίῳ ἐμού.

155 24 виждъ что творатъ въ сбобртъ съйнсне не достойтъ.

156 25 и тъ глъше имъ.

157 25 нѣсте ли николиже чъли. что сътвори дѣвдъ егда
Sentence alignment

- Sentence alignment is a many-to-many relation.
- We adjust sentence divisions, which are fluid to begin with.
- Performance depends on text style and sentence length.
- Faulty sentence divisions often disrupt the alignment.
Dependency alignment

- Dependency alignment is mostly a manual process.
- We employ iterative refinement of alignments.
A graph problem

- Each dependency structure is a directed graph.
- The *root node* is the node with no incoming edges.
- A dependency relation is a labeled edge from the head to the dependent.
- A secondary dependency relation is a labeled edge from the ‘*slasher*’ to the ‘*slashee*’.
Alignments are unlabeled edges from a translation to the original.
A graph problem

- Alignments are unlabeled edges from a translation to the original.
  - Sentence alignments are edges between root nodes.
Alignments are unlabeled edges from a translation to the original.
- Sentence alignments are edges between root nodes.
- Token alignments are edges between non-empty nodes.
A graph problem

- Alignments are unlabeled edges from a translation to the original.
  - Sentence alignments are edges between root nodes.
  - Token alignments are edges between non-empty nodes.
  - Dependency alignments are edges between non-root nodes.
Inheritance

- Dependency alignments align dependency subgraphs.
- We say that dependency alignments are *inherited* in dependency subgraphs.
- *Terminations* are used to block inheritance.
Principles

- Syntactic-semantic alignments
  - Same syntactic level
  - Same semantic role and/or equivalent syntactic relation
- Groups of syntactic relations:
  - Subjects
  - Nominal arguments of verbs (OBJ, OBL)
  - Clausal arguments of verbs
  - Adjuncts
  - Nominal modifiers
- Different syntactic relations can be linked when the semantic role is the same
Matthew 16:26

773 26 τί γὰρ ὄφεληθῆσεται ἄνθρωπος, ἐὰν τὸν κόσμον ὅλον κερδήσῃ τὴν δὲ ψυχήν αὐτοῦ ζημιώθη;

653 26 καὶ ποιῆσα εστὶ πόλιν ἐπὶ τὴν ἐπιφάνειαν ἁπτε νέων ἐπὶ προσφεύγειν. αὐτὸ ἦν ὁ πόλεως ὀρθήτητι.
Principles of linking: same syntactic level
Principles of linking: semantics or syntax?

Alignment  Principles

The PROIEL corpus

Jøhndal, Eckhoff and Haug (UiO)
Principles of linking: empty nodes
Principles of linking: empty nodes and null terminations
Principles of null terminations

- Null terminations used when:
  - something is missing
  - syntactic relation is too different
  - attachment site is too different
Mark 1:4

5-4 ἐγένετο Ἰωάννης ὁ βαπτίζων ἐν τῇ ἐρήμῳ καὶ κηρύσσων βάπτισμα

58-4 δῆστε πολλῷ κρύσταλι ἐν ποισμῇ. ἢ προφοβάδα κρύσταλλον

μετανοίας εἰς ἁφεσίν ἄμαρτων.

πολλῷ ἐν ὀποίου πιστεύω γράφομεν.
Principles of null terminations
Principles of null terminations
Principles of null terminations
Conclusion

Summing up

- Alignment layer gives great potential for contrastive work
- Automatic token alignment
- Manual dependency alignment
- Mixed criteria for dependency alignment gives richer search possibilities
- Inheritance highlights differences between languages
- Terminations allow a bottom-up as well as a top-down perspective in the graphs
Completion

- **Church Slavonic**: ca. 25,000 of 58,000 words annotated
- **Greek**: ca. 110,000 of 138,000 words annotated
- The **Latin**, **Gothic** and **Armenian** versions are also available.
Availability

- The corpus is available for everyone to use.
- We publish XML files with raw data as well.
- All our data is released under a Creative Commons license.
- Visit http://www.hf.uio.no/ifikk/proiel/ for details.