# The PROIEL corpus of New Testament translations

Syntax, semantics and information structure

Dag Haug

University of Oslo

March 29, 2012
The background

- A corpus for linguists: focus on making the most of a limited data set for linguistic research
The background

- A corpus for linguists: focus on making the most of a limited data set for linguistic research
- Pragmatic Resources in Old Indo-European Languages (PROIEL)
The background

- A corpus for linguists: focus on making the most of a limited data set for linguistic research
- Pragmatic Resources in Old Indo-European Languages (PROIEL)
  - word order
The background

- A corpus for linguists: focus on making the most of a limited data set for linguistic research
- Pragmatic Resources in Old Indo-European Languages (PROIEL)
  - word order
  - anaphoric expressions
The background

- A corpus for linguists: focus on making the most of a limited data set for linguistic research
- Pragmatic Resources in Old Indo-European Languages (PROIEL)
  - word order
  - anaphoric expressions
  - definiteness
The background

- A corpus for linguists: focus on making the most of a limited data set for linguistic research
- Pragmatic Resources in Old Indo-European Languages (PROIEL)
  - word order
  - anaphoric expressions
  - definiteness
  - participles (background events)
The background

- A corpus for linguists: focus on making the most of a limited data set for linguistic research
- Pragmatic Resources in Old Indo-European Languages (PROIEL)
  - word order
  - anaphoric expressions
  - definiteness
  - participles (background events)
  - discourse particles
A naturally occurring parallel corpus

- The New Testament in its Greek original and Latin, Gothic, Classical Armenian and OCS translations
  - Ancient Greek (original, 1st century AD)
  - Gothic (4th century AD)
  - Latin (ca. 400 AD)
  - Classical Armenian (ca. 400 AD)
  - Old Church Slavic (9th century AD)
A naturally occurring parallel corpus

- The New Testament in its Greek original and Latin, Gothic, Classical Armenian and OCS translations
  - Ancient Greek (original, 1st century AD)
  - Gothic (4th century AD)
  - Latin (ca. 400 AD)
  - Classical Armenian (ca. 400 AD)
  - Old Church Slavic (9th century AD)

- The NT translations are the oldest attestations of Armenian and OCS, and virtually the only attestation of Gothic
A naturally occurring parallel corpus

- The New Testament in its Greek original and Latin, Gothic, Classical Armenian and OCS translations
  - Ancient Greek (original, 1st century AD)
  - Gothic (4th century AD)
  - Latin (ca. 400 AD)
  - Classical Armenian (ca. 400 AD)
  - Old Church Slavic (9th century AD)

- The NT translations are the oldest attestations of Armenian and OCS, and virtually the only attestation of Gothic

- The gospels constitute the core of the OCS text canon
Extensions

- Herodotus’ Histories (Greek 5th century BC)
- Caesar’s Gallic War (Latin, 1st century BC)
- Cicero’s Letters to Atticus (Latin, 1st century BC)
- Peregrinatio Aetheriae (Vulgar Latin, ca. 400 AD)
- Hagiographies (The Slavic Codex Suprasliensis, 11th century AD)
 Corpora in historical linguistics

- All historical linguistics is (or should be) corpus linguistics
Corpora in historical linguistics

- All historical linguistics is (or should be) corpus linguistics
- Statistics as a substitute for (non-native) intuitions
Corpora in historical linguistics

- All historical linguistics is (or should be) corpus linguistics
- Statistics as a substitute for (non-native) intuitions
- Sometimes we can approach negative evidence
Corpora in historical linguistics

- All historical linguistics is (or should be) corpus linguistics
- Statistics as a substitute for (non-native) intuitions
- Sometimes we can approach negative evidence
- Pioneer work in corpus linguistics by historical linguists: *Index Thomisticus* (started 1949!), *Thesaurus Linguae Graecae* (released in the 70s)
The PROIEL corpus

- Many-layered manual annotation of all the texts
  - Morphological annotation

<table>
<thead>
<tr>
<th>Introduction</th>
<th>Morpho-syntactic annotation</th>
<th>Semantics</th>
<th>Information structure</th>
<th>Discourse</th>
</tr>
</thead>
<tbody>
<tr>
<td>The PROIEL corpus of New Testament translations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The PROIEL corpus

- Many-layered manual annotation of all the texts
  - Morphological annotation
  - Syntactic annotation (dependency/LFG-based)
The PROIEL corpus

- Many-layered manual annotation of all the texts
  - Morphological annotation
  - Syntactic annotation (dependency/LFG-based)
  - Semantic and other customised annotation (e.g. animacy)
The PROIEL corpus

- Many-layered manual annotation of all the texts
  - Morphological annotation
  - Syntactic annotation (dependency/LFG-based)
  - Semantic and other customised annotation (e.g. animacy)
  - Annotation of information structure
The PROIEL corpus

- Many-layered manual annotation of all the texts
  - Morphological annotation
  - Syntactic annotation (dependency/LFG-based)
  - Semantic and other customised annotation (e.g. animacy)
  - Annotation of information structure
  - Annotation of discourse structure
Workflow for annotation

- International team of student annotators
Workflow for annotation

- International team of student annotators
- Manual disambiguation of morphology and lemmatization
Workflow for annotation

- International team of student annotators
- Manual disambiguation of morphology and lemmatization
- Syntactic annotation
Workflow for annotation

- International team of student annotators
- Manual disambiguation of morphology and lemmatization
- Syntactic annotation
- Review by project members
Workflow for annotation

- International team of student annotators
- Manual disambiguation of morphology and lemmatization
- Syntactic annotation
- Review by project members
- Advanced annotation by project members
Morphological annotation

- Detailed morphological annotation

We now have enough data to train a model and pretag texts. Accuracy (10-fold cross-validation) in The Gallic War is 84.3% for MSD-tags and 92.88% for PoS-tags (Skjærholt 2011). Manual post-correction: example DAG HAU.
Morphological annotation

- Detailed morphological annotation
- We now have enough data to train a model and pretag texts
Morphological annotation

- Detailed morphological annotation
- We now have enough data to train a model and pretag texts
- Accuracy (10-fold cross-validation) in *The Gallic War* is 84.3% for MSD-tags and 92.88% for PoS-tags (Skjærholt 2011)
Morphological annotation

- Detailed morphological annotation
- We now have enough data to train a model and pretag texts
- Accuracy (10-fold cross-validation) in *The Gallic War* is 84.3% for MSD-tags and 92.88% for PoS-tags (Skjærholt 2011)
- Manual post-correction: example
The syntactic annotation scheme: dependency grammar

- Information about syntactic relations and word order stored separately
The syntactic annotation scheme: dependency grammar

- Information about syntactic relations and word order stored separately
- Reliance on overt elements
The syntactic annotation scheme: dependency grammar

- Information about syntactic relations and word order stored separately
- Reliance on overt elements
- Inherent problems: (asynthetic) parataxis, structure sharing
The syntactic annotation scheme: dependency grammar

- Information about syntactic relations and word order stored separately
- Reliance on overt elements
- Inherent problems: (asyndetic) parataxis, structure sharing
- Dependency grammar with LFG adjustments
The syntactic annotation scheme: dependency grammar

- Information about syntactic relations and word order stored separately
- Reliance on overt elements
- Inherent problems: (asyndetic) parataxis, structure sharing
- Dependency grammar with LFG adjustments
  - Limited set of empty nodes (to handle asyndetic coordination and ellipsis)
The syntactic annotation scheme: dependency grammar

- Information about syntactic relations and word order stored separately
- Reliance on overt elements
- Inherent problems: (asyndetic) parataxis, structure sharing
- Dependency grammar with LFG adjustments
  - Limited set of empty nodes (to handle asyndetic coordination and ellipsis)
  - Secondary dependencies (for structure sharing, incl. control/raising)
The syntactic annotation scheme: dependency grammar

- Information about syntactic relations and word order stored separately
- Reliance on overt elements
- Inherent problems: (asyndetic) parataxis, structure sharing
- Dependency grammar with LFG adjustments
  - Limited set of empty nodes (to handle asyndetic coordination and ellipsis)
  - Secondary dependencies (for structure sharing, incl. control/raising)
  - More granular syntactic relations
## Syntactic relations

<table>
<thead>
<tr>
<th>Label</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRED</td>
<td>Predicate</td>
</tr>
<tr>
<td>SUB</td>
<td>Subject</td>
</tr>
<tr>
<td>OBJ</td>
<td>Object</td>
</tr>
<tr>
<td>OBL</td>
<td>Oblique</td>
</tr>
<tr>
<td>AG</td>
<td>Agent</td>
</tr>
<tr>
<td>ADV</td>
<td>Adverbial</td>
</tr>
<tr>
<td>ATR</td>
<td>Attribute</td>
</tr>
<tr>
<td>APOS</td>
<td>Apposition</td>
</tr>
<tr>
<td>NARG</td>
<td>Nominal argument</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>XADV</td>
<td>Free predicative</td>
</tr>
<tr>
<td>XOBJ</td>
<td>Open complement</td>
</tr>
<tr>
<td>Aux</td>
<td>Auxiliary</td>
</tr>
<tr>
<td>XOBJ</td>
<td>Open complement clause</td>
</tr>
<tr>
<td>COMP</td>
<td>Complement clause</td>
</tr>
<tr>
<td>PART</td>
<td>Partitive</td>
</tr>
<tr>
<td>PARPRED</td>
<td>Parenthetical</td>
</tr>
<tr>
<td>VOC</td>
<td>Vocative</td>
</tr>
</tbody>
</table>

---

Dag Haug

The PROIEL corpus of New Testament transl... March 29, 2012
Empty nodes

- Null conjunctions for asyndetic parataxis
- Null verbs for null copulas and elided verbs
Empty nodes

- root
  - PRED
  - and
    - PRED
      - drank
        - SUB
          - Peter
        - OBJ
          - beer
      - PRED
        - V
          - SUB
            - John
          - OBJ
            - wine
Empty nodes
Human processing
Human processing
Secondary dependencies

- Control
  - example
Secondary dependencies

- Control
- Raising
Secondary dependencies

- Control
- Raising
- Shared arguments
Secondary dependencies

- Control
- Raising
- Shared arguments
- Predicate identity
Storing linguistic analyses

- Theory-neutrality →
- data for larger audiences
Storing linguistic analyses

- Theory-neutrality →
  - data for larger audiences
  - widening gulf between corpus linguistics and linguistic theory
Storing linguistic analyses

- **Theory-neutrality** →
  - data for larger audiences
  - widening gulf between corpus linguistics and linguistic theory

- DG corpora (Prague, PROIEL) → DG not really in use as a linguistic theory
Storing linguistic analyses

- Theory-neutrality →
  - data for larger audiences
  - widening gulf between corpus linguistics and linguistic theory

- DG corpora (Prague, PROIEL) → DG not really in use as a linguistic theory

- PS corpora (Penn, NEGRA) typically use flatter tree structures than anyone believes in
Storing linguistic analyses

- Theory-neutrality →
  - data for larger audiences
  - widening gulf between corpus linguistics and linguistic theory

- DG corpora (Prague, PROIEL) → DG not really in use as a linguistic theory

- PS corpora (Penn, NEGRA) typically use flatter tree structures than anyone believes in

- On the other hand, LFG and HPSG corpora can be hard to use for people who do not share the theoretical assumptions of these theories
# Our take

## Principles

1. **Encode no more structure than is common to all frameworks**
Our take

Principles

1. Encode no more structure than is common to all frameworks
2. Enoded structure could be seen as derived/secondary in some frameworks
Our take

Principles

1. Encode no more structure than is common to all frameworks
2. Encoded structure could be seen as derived/secondary in some frameworks
3. Encode enough structure to allow reconstruction of theoretically motivated structures
Our take

Principles

1. Encode no more structure than is common to all frameworks
2. Encoded structure could be seen as derived/secondary in some frameworks
3. Encode enough structure to allow reconstruction of theoretically motivated structures

In the ideal situation, the information in the annotation can be (monotonically) expanded to structures conforming to a particular theory by adding information from the assumptions of that theory.
The ideal situation

- The added assumptions will typically be about phrase structure, such as various versions of $X'$ theory.
The ideal situation

- The added assumptions will typically be about phrase structure, such as various versions of X’ theory.
- Given information about what the subject is, it will be possible to create a structure where the subject has a specific position if the theory requires that (unless the data contradict the theory).
The ideal situation

- The added assumptions will typically be about phrase structure, such as various versions of X' theory.
- Given information about what the subject is, it will be possible to create a structure where the subject has a specific position if the theory requires that (unless the data contradict the theory).
- Useful for testing theories.
Relationship to LFG’s f-structures

licet  viro  dimittere  uxorem
it is allowed  man.DAT  divorce  wife.ACC
Relationship to LFG’s f-structures

licet

viro

it is allowed man.DAT

dimittere divorce

uxorem wife.ACC
Relationship to LFG’s c-structures

PROIEL

Introduction Morpho-syntactic annotation Semantics Information structure Discourse

Dag Haug The PROIEL corpus of New Testament transl March 29, 2012 19 / 64
Customised tagging

- Possibility for user-defined tagging of any phenomenon
Customised tagging

- Possibility for user-defined tagging of any phenomenon
- Individual scholars can make their classifications available to others in the interest of replicability and further research
Customised tagging

- Possibility for user-defined tagging of any phenomenon
- Individual scholars can make their classifications available to others in the interest of replicability and further research
- So far mostly semantic tagging: animacy, aktionsart, noun relationality
Economy of effort

- Annotation at lemma level whenever possible, adjustment at token level
Economy of effort

- Annotation at lemma level whenever possible, adjustment at token level
- Tag transfer from nouns to e.g. pronouns via anaphoric links
Economy of effort

- Annotation at lemma level whenever possible, adjustment at token level
- Tag transfer from nouns to e.g. pronouns via anaphoric links
- Tags that are applicable to all the languages are tagged on Greek lemmata first, then transferred via the token alignments (more on this in lecture 2!)
Semantic annotation – animacy

- HUMAN
- ORG
- ANIMAL
- VEH
- CONC
- PLACE
- NONCONC
- TIME

All Greek noun lemmata annotated for animacy
Adjustments at token level
Tag transfer to other parts of speech via anaphoric links
Tag transfer to other languages via token alignments
Semantic annotation – animacy

- All Greek noun lemmata annotated for animacy
- Adjustments at token level
- Tag transfer to other parts of speech via anaphoric links
- Tag transfer to other languages via token alignments
Animacy tagging – challenges

- The tagging is meant as token-level tagging, but is most efficiently done at lemma level.
- The lemma should therefore have the tag that the majority of its tokens indicate, not its ‘basic meaning’.
- Case in point: *kardia* ‘heart’ got the tag NONCONC, since none of the tokens referred to physical hearts, but always to minds, thoughts and opinions.
Information structure annotation

- A core interest in the PROIEL project
Information structure annotation

- A core interest in the PROIEL project
- Difficult field: We concentrate on categories where it is possible to achieve intersubjective agreement
Information structure annotation

- A core interest in the PROIEL project
- Difficult field: We concentrate on categories where it is possible to achieve intersubjective agreement
- In this stage mainly givenness and anaphoricity
The process towards the annotation scheme

- Pilot annotation of 655 NPs by core project members
The process towards the annotation scheme

- Pilot annotation of 655 NPs by core project members
- Critical evaluation of the results
### The process towards the annotation scheme

- Pilot annotation of 655 NPs by core project members
- Critical evaluation of the results
- Discussion with other groups annotating IS, including Arndt Riester
The process towards the annotation scheme

- Pilot annotation of 655 NPs by core project members
- Critical evaluation of the results
- Discussion with other groups annotating IS, including Arndt Riester
- Trial and error, including retagging of large portions of annotated text
Introduction Morpho-syntactic annotation Semantics Information structure Discourse

The process towards the annotation scheme

- Pilot annotation of 655 NPs by core project members
- Critical evaluation of the results
- Discussion with other groups annotating IS, including Arndt Riester
- Trial and error, including retagging of large portions of annotated text
- Current coverage: 58186 NPs (the Gospels + starting on Caesar)
The design

- Choice of referential expression → reference-based notion of givenness
The design

- Choice of referential expression → reference-based notion of givenness
- But there are clear effects of word repetition that we then miss
Mark 14:66-72

66 Und Petrus war unten im Hof. Da kam eine von des Hohenpriesters Mägden; 67 und da sie sah Petrus sich wärmen, schaute sie ihn an und sprach: Und du warst auch mit Jesus von Nazareth. 68 Er leugnete aber und sprach: Ich kenne ihn nicht, wei auch nicht, was du sagst. Und er ging hinaus in den Vorhof; und der Hahn krächte. 69 Und die Magd sah ihn und hob abermals an, zu sagen denen, die dabeistanden: Dieser ist deren einer. 70 Und er leugnete abermals. Und nach einer kleinen Weile sprachen abermals zu Petrus, die dabeistanden: Wahrlich, du bist deren einer; denn du bist ein Galiläer, und deine Sprache lautet gleich also. 71 Er aber fing an, sich zu verfluchen und zu schwören: Ich kenne den Menschen nicht, von dem ihr sagt. 72 Und der Hahn krächte zum andernmal. Da gedachte Petrus an das Wort, das Jesus zu ihm sagte: Ehe der Hahn zweimal kräht, wirst du mich dreimal verleugnen. Und er hob an, zu weinen.
The design

- Choice of referential expression $\rightarrow$ reference-based notion of givenness
- But there are clear effects of word repetition that we then miss
- Study object: the semantics of referential expressions
The design

- Choice of referential expression $\rightarrow$ reference-based notion of givenness
- But there are clear effects of word repetition that we then miss
- Study object: the semantics of referential expressions
- The form of referential expressions should not influence the annotation (to avoid circularity)
Non-standard article usage

Mark 4:3

Höret zu! Siehe, es ging ein Sämann aus, zu säen.

- This is definite in Greek, despite the presentation construction
Non-standard article usage

Mark 4:3

Höret zu! Siehe, es ging ein Sämann aus, zu säen.

- This is definite in Greek, despite the presentation construction
- Generally freer use of definites in the parables
Quantificational article use

Matthew 7:6

\[ m\acute{e} \ d\grave{o}t\grave{e} \quad to \quad hagion \quad tois \quad kusin \]

not give.IMP the holy.ACC the dogsDAT

‘Do not give dogs what is holy’
Quantificational article use

Matthew 7:6

*mê dôte to hagion tois kusin*
not give.IMP the holy.ACC the dogs.DAT

‘Do not give dogs what is holy’

Luke 16:18

*ho apolelumenên apo andros gamôn moikheuei*
the divorced from man.GEN marrying.NOM commit adultery

‘Whoever marries a woman divorced from a man commits adultery’
The design

- Choice of referential expression $\rightarrow$ reference-based notion of givenness
- But there are clear effects of word repetition that we then miss
- Study object: the semantics of referential expressions
- The form of referential expressions should not influence the annotation (to avoid circularity)
- Should be applicable across languages with different referential expression systems
The design

- Choice of referential expression → reference-based notion of givenness
- But there are clear effects of word repetition that we then miss
- Study object: the semantics of referential expressions
- The form of referential expressions should not influence the annotation (to avoid circularity)
- Should be applicable across languages with different referential expression systems
- An annotation to be used together with other layers (in particular syntax)
Combining layers

- Some schemes incorporate distinction such as the following in the GIVEN-category (from Riester et al.)
  - PRONOUN
  - REFLEXIVE
  - SHORT
  - REPEATED
  - EPITHET
Combining layers

- Some schemes incorporate distinction such as the following in the GIVEN-category (from Riester et al.)
  - PRONOUN
  - REFLEXIVE
  - SHORT
  - REPEATED
  - EPITHET

- In PROIEL, information about the type of anaphoric expression is available from the syntactic annotation
Combining layers

- Some schemes incorporate distinction such as the following in the GIVEN-category (from Riester et al.)
  - PRONOUN
  - REFLEXIVE
  - SHORT
  - REPEATED
  - EPITHET

- In PROIEL, information about the type of anaphoric expression is available from the syntactic annotation

- Preferrable, because it gives much more detail, but of course time-consuming
Givenness — our approach

How can the hearer establish the referent of an NP?
Givenness

Givenness — our approach

How can the hearer establish the referent of an NP?

- Primary question: does the NP have one or more referents at all?
Discourse referents

The heuristics

Does the NP establish a referent that could be picked up in the following discourse?

1. Jesus saw a leper. He was blind.
2. Jesus saw some lepers. They were blind.
3. Jesus wandered in deserted places. *They were called Sodoma and Gomorra.

In Latin and (and to some extent Greek) we cannot rely on the presence of an overt determiner to decide which NPs are referential.
Discourse referents

The heuristics

Does the NP establish a referent that could be picked up in the following discourse?

1. Jesus saw a leper. He was blind.
Discourse referents

The heuristics

Does the NP establish a referent that could be picked up in the following discourse?

1. Jesus saw a leper. He was blind.
2. Jesus saw some lepers. They were blind.
Discourse referents

The heuristics

Does the NP establish a referent that could be picked up in the following discourse?

1. Jesus saw a leper. He was blind.
2. Jesus saw some lepers. They were blind.
3. Jesus wandered in deserted places. *They were called Sodoma and Gomorra
Discourse referents

The heuristics

Does the NP establish a referent that could be picked up in the following discourse?

1. Jesus saw a leper. He was blind.
2. Jesus saw some lepers. They were blind.
3. Jesus wandered in deserted places. *They were called Sodoma and Gomorra

In Latin and (and to some extent Greek) we cannot rely on the presence of an overt determiner to decide which NPs are referential
NPs without determiners

Luke 15:14

egeneto limos iskhura kata tên khôran ekeinên become.AOR.3SG famine powerful over the country that

There was (a) great famine over that country
NPs without determiners

Caesar BG 6.35.1

Transeunt Rhenum navibus ratibus=que

cross Rhine ships barks=and

They cross the Rhine in (the/their) ships and barks
NPs without determiners

Luke 5:18

et ecce viri portantes in lecto hominem

and behold men carrying in bed man
NPs without determiners

Luke 5:18

-et ecce viri portantes in lecto hominem

and behold men carrying in bed man

Und, siehe, etliche Männer brachten einen Menschen auf seinem Bette
And behold, some men were bringing on a bed a man (English Standard Version)
And behold, men brought in a bed a man (King James)
The English indefinite plural is not the plural of the indefinite and does not normally introduce a discourse referent.
Determinerless languages

- The English indefinite plural is not the plural of the indefinite and does not normally introduce a discourse referent.
- But in Greek and Latin, the determinerless plural can but does not have to introduce a (plural) discourse referent.
Determinerless languages

- The English indefinite plural is not the plural of the indefinite and does not normally introduce a discourse referent.
- But in Greek and Latin, the determinerless plural *can but does not have to* introduce a (plural) discourse referent.
- We need heuristics.
Determinerless languages

- The English indefinite plural is not the plural of the indefinite and does not normally introduce a discourse referent
- But in Greek and Latin, the determinerless plural *can but does not have to* introduce a (plural) discourse referent
- We need heuristics
- We assume that singular count nouns introduce DRs, but not mass nouns and plurals – unless they are explicitly picked up in the contexts
Modal subordination

Luke 15:4–5

4 Welcher Mensch ist unter euch, der hundert Schafe hat und, so er der eines verliert,
Modal subordination

Luke 15:4–5

4 Welcher Mensch ist unter euch, der hundert Schafe hat und, so er der eines verliert, der nicht lasse die neunundneunzig in der Wüste und hingehe nach dem verlorenen,
Modal subordination

Luke 15:4–5

4 Welcher Mensch ist unter euch, der hundert Schafe hat und, so er der eines verliert, der nicht lasse die neunundneunzig in der Wüste und hingehe nach dem verlorenen, bis da er’s finde?
**Modal subordination**

**Luke 15:4–5**

4 Welcher Mensch ist unter euch, der hundert Schafe hat und, so er der **eines** verliert, der nicht lasse die neunundneunzig in der Wüste und hingehe nach dem **verlorenen**, bis da er’s finde? 5 Und wenn er’s gefunden hat, so legt er’s auf seine Achseln mit Freuden.
Modal subordination

Luke 15:4–5

4 Welcher Mensch ist unter euch, der hundert Schafe hat und, so er der eines verliert, der nicht lasse die neunundneunzig in der Wüste und hingehe nach dem verlorenen, bis da er’s finde? 5 Und wenn er’s gefunden hat, so legt er’s auf seine Achseln mit Freuden.

- Quite common in the NT
Modal subordination

Luke 15:4–5

4 Welcher Mensch ist unter euch, der hundert Schafe hat und, so er der eines verliert, der nicht lasse die neunundneunzig in der Wüste und hingehe nach dem verlorenen, bis da er’s finde? 5 Und wenn er’s gefunden hat, so legt er’s auf seine Achseln mit Freuden.

- Quite common in the NT
- So we want to deal with embedded/non-specific discourse referents
Modal subordination

Luke 15:4–5

4 Welcher Mensch ist unter euch, der hundert Schafe hat und, so er der eines verliert, der nicht lasse die neunundneunzig in der Wüste und hingehe nach dem verlorenen, bis da er's finde? 5 Und wenn er's gefunden hat, so legt er's auf seine Achseln mit Freuden.

- Quite common in the NT
- So we want to deal with embedded/non-specific discourse referents
- But we don’t want to mix them with other discourse referents
Modal subordination

Luke 15:4–5

4 Welcher Mensch ist unter euch, der hundert Schafe hat und, so er der eines verliert, der nicht lasse die neunundneunzig in der Wüste und hingehe nach dem verlorenen, bis da er’s finde? 5 Und wenn er’s gefunden hat, so legt er’s auf seine Achseln mit Freuden.

- Quite common in the NT
- So we want to deal with embedded/non-specific discourse referents
- But we don’t want to mix them with other discourse referents
- We will get back to these
Discourse referents

- Our definition includes most NPs (also those embedded in other NPs!), since even non-specifics can be picked up inside embeddings
Discourse referents

- Our definition includes most NPs (also those embedded in other NPs!), since even non-specifics can be picked up inside embeddings.
- But some are excluded:
  - Idiom chunks (\textit{khreian ekhousin} lit. ‘have need’)
  - NPs in temporal expressions (\textit{he stayed four days})
  - Appositions (\textit{Herod the king})
Discourse referents

Null arguments

eipen
say.IND.AOR.3SG

- Greek allows prodrop of subject and non-subject arguments
Discourse referents

Null arguments

- eipen
- say.IND.AOR.3SG

- Greek allows prodrop of subject and non-subject arguments
- Subject prodrops are easy: just insert them whenever a verb that is not impersonal lacks an overt subjects
Discourse referents

Null arguments

-eipen
say.IND.AOR.3SG

- Greek allows prodrop of subject and non-subject arguments
- Subject prodrops are easy: just insert them whenever a verb that is not impersonal lacks an overt subject
- This example clearly includes a pro subject, but is there a pro dative?
Discourse referents

Null arguments

eipen
say.IND.AOR.3SG

- Greek allows prodrop of subject and non-subject arguments
- Subject prodrops are easy: just insert them whenever a verb that is not impersonal lacks an overt subjects
- This example clearly includes a pro subject, but is there a pro dative?
- We have been restrictive with non-subject null arguments
Null objects

*kai eurôn ∅ epitithêsin ∅ epi tous ômous autou*

and finding puts on the shoulders his

Und wenn er’s gefunden hat, so legt er’s auf seine Achseln (Luke 15:5)

● Clear non-subject null arguments that must be inserted
Null objects

*kai eurôn ∅ epitithêsin ∅ epi tous ômous autou*

and finding puts on the shoulders his

Und wenn er’s gefunden hat, so legt er’s auf seine Achseln (Luke 15:5)

- Clear non-subject null arguments that must be inserted
- Lacking good valency dictionaries this is impressionistic at the moment
Null objects

\textit{kai eurôn} $\emptyset$ \textit{epitithêsin} $\emptyset$ \textit{epi tous ômous} \textit{autou}

and finding puts on the shoulders his

Und wenn er’s gefunden hat, so legt er’s auf seine Achseln (Luke 15:5)

- Clear non-subject null arguments that must be inserted
- Lacking good valency dictionaries this is impressionistic at the moment
- Can hopefully be made uniform once the corpus is big enough to create a valency dictionary
Meaning contexts

- Givenness tags based on which context the hearer uses to establish reference
  - Discourse (anaphora) → OLD
Meaning contexts

- Givenness tags based on which context the hearer uses to establish reference
  - Discourse (anaphora) → OLD
  - Situation (deixis) → ACC-sit
Meaning contexts

- Givenness tags based on which context the hearer uses to establish reference
  - Discourse (anaphora) $\rightarrow$ OLD
  - Situation (deixis) $\rightarrow$ ACC-SIT
  - Scenarios (inferences) $\rightarrow$ ACC-INF
Meaning contexts

- Givenness tags based on which context the hearer uses to establish reference
  - Discourse (anaphora) → OLD
  - Situation (deixis) → ACC-SIT
  - Scenarios (inferences) → ACC-INF
  - Encyclopedic knowledge → ACC-GEN
Meaning contexts

- Givenness tags based on which context the hearer uses to establish reference
  - Discourse (anaphora) → OLD
  - Situation (deixis) → ACC-SIT
  - Scenarios (inferences) → ACC-INF
  - Encyclopedic knowledge → ACC-GEN
  - No context (no extra-NP information) → NEW
Discourse context

- This is the most straightforward tag
- It is also the most common tag – 58.5% of referents
- It is accompanied by an anaphoric link pointed to the antecedent
- In the case of split antecedents we point to the last one

▶ sample
Situation context

- Another straightforward tag, mostly found in direct speech
Situation context

- Another straightforward tag, mostly found in direct speech
- Rare in our texts – 1.5%
Situation context

- Another straightforward tag, mostly found in direct speech
- Rare in our texts – 1.5%
- Mostly given off by explicit linguistic cues such as demonstratives
Situation context

- Another straightforward tag, mostly found in direct speech
- Rare in our texts – 1.5%
- Mostly given off by explicit linguistic cues such as demonstratives
- Can occur within direct speech with a discourse antecedent outside the direct speech
Situation context

- Another straightforward tag, mostly found in direct speech
- Rare in our texts – 1.5%
- Mostly given off by explicit linguistic cues such as demonstratives
- Can occur within direct speech with a discourse antecedent outside the direct speech
- We mark coreference in these cases
Caesar BG 1.10

*tres quae circum Aquileiam hiemabant ex hibernis*

three who around Aquieia winter.IMPF.3P from winter camp.ABL

*PRO-SUB educit*

PRO lead

‘He led the three legions that were wintering around Aquileia out of (their) winter camps’
Scenário conhecimento

Caesar BG 1.10

tres quae circum Aquileiam hiemabant ex hibernis
three who around Aquileia winter.IMPF.3P from winter camp.ABL

PRO-SUB educit

PRO lead

‘He led the three legions that were wintering around Aquileia out of (their) winter camps’

- Legions → camps
Scenario knowledge

Caesar BG 1.10

tres quae circum Aquileiam hiemabant ex hibernis
three who around Aquileia winter.IMPF.3P from winter camp.ABL
PRO-SUB educit
PRO lead

‘He led the three legions that were wintering around Aquileia out of (their) winter camps’

- Legions → camps
- When Jesus enters a town, there will typically be lepers around
Scenario knowledge

Caesar BG 1.10

tres quae circum Aquileiam hiemabant ex hibernis
three who around Aquileia winter.IMPF.3P from winter camp.ABL

PRO-SUB educit

PRO lead

‘He led the three legions that were wintering around Aquileia out of (their) winter camps’

- Legions $\rightarrow$ camps
- When Jesus enters a town, there will typically be lepers around
- But where to stop?
Restricting scenario knowledge

- We require that the inference is based on lexical evidence, not typical plots
  - The connection should be obvious and stereotypical: example
  - But not necessarily predictable in advance: example
Restricting scenario knowledge

- We require that the inference is based on lexical evidence, not typical plots
  - The connection should be obvious and stereotypical: example
  - But not necessarily predictable in advance: example
- 4.4% of tags
Encyclopaedic context

- These are things that we expect a Hellenized Jew (NT) or an educated Roman (Caesar) to know about
Encyclopaedic context

- These are things that we expect a Hellenized Jew (NT) or an educated Roman (Caesar) to know about
- Sounds difficult, but in practice not too hard
Encyclopaedic context

- These are things that we expect a Hellenized Jew (NT) or an educated Roman (Caesar) to know about
- Sounds difficult, but in practice not too hard
- We can (and must) rely on overt indicators
Encyclopaedic context

- These are things that we expect a Hellenized Jew (NT) or an educated Roman (Caesar) to know about
- Sounds difficult, but in practice not too hard
- We can (and must) rely on overt indicators
- 6.4% of referents
Encyclopedic knowledge

Caesar BG 1.6.3

*Extremum oppidum Allobrogum est proximum=que Helvetiorum furthest town Allobroges.Gen is closest=and Helvetii.Gen finibus Genava borders.DAT Geneva.NOM

‘The furthest town of the Allobroges and the closes to the borders of the Helvetii is Geneva’
Encyclopedic knowledge

Caesar BG 1.10.3

tres quae circum Aquileiam hiemabant ex hibernis
three who around Aquieia winter.IMPF.3P from winter camp.ABL

PRO-SUB educit

PRO lead

‘He led the three legions that were wintering around Aquileia out of (their) winter camps
No context

- Another straightforward tag, once it is decided that there is a taggable in the first place
- Not very common (9.8% of NPs)
Complex NPs

Luke 10:16

ho de eme athetôn athetei ton aposteilanta me
the but me.ACC refusing refuse.PRS.3SG the sending me.ME

‘Whoever refuses me, refuses the one who sent me’
Internal anchoring

- ‘Bridging’ is often explicitly marked, e.g. by a possessive
‘Bridging’ is often explicitly marked, e.g. by a possessive

In this case too, we want to mark both referents
Embedded discourse referents

- We know from Karttunen/DRT that we can have anaphoric processes inside embeddings that otherwise block anaphoric reference from the outside.

Luke 17:33

*hos ean zêtêsêi tên psukhênn autou peripoiêsasthai apolesei autên*

whoever seeks the soul his save lose it

Whoever seeks to save his soul will lose it
### DRT analysis

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>person(( x ))</td>
<td>soul(( y, z ))</td>
<td>( z = x )</td>
</tr>
</tbody>
</table>

\[ \text{try_to_save}(x, y) \]

\[ \Rightarrow \]

<table>
<thead>
<tr>
<th>( v, w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>lose(( u, w ))</td>
</tr>
<tr>
<td>( u = x )</td>
</tr>
<tr>
<td>( w = z )</td>
</tr>
</tbody>
</table>
Bridging inside embeddings

Luke 11:33

Niemand zündet ein Licht an und setzt es an einen heimlichen Ort, auch nicht unter einen Scheffel, sondern auf den Leuchter, auf da, wer hineingeht, das Licht sehe.
Modal/quantificational subordination

**Luke 5:39**

Und niemand ist, der vom alten trinkt und wolle bald den neuen; denn er spricht: Der alte ist milder.

- The translation follows the Greek
- There is no overt indication of the quantificational subordination
Our solution

- DRT allows recursive embedding of DRSs (and hence universes of discourse referents)
Our solution

- DRT allows recursive embedding of DRSs (and hence universes of discourse referents)
- This is impracticable in an annotation scheme
Our solution

- DRT allows recursive embedding of DRSs (and hence universes of discourse referents)
- This is impracticable in an annotation scheme
- But we can use a separate set of tags for all DRs outside the main DRS
  - NONSPEC (but QUANT for quantification)
  - NONSPEC_INF
  - NONSPEC_OLD
Our solution

- DRT allows recursive embedding of DRSs (and hence universes of discourse referents)
- This is impracticable in an annotation scheme
- But we can use a separate set of tags for all DRs outside the main DRS
  - NONSPEC (but QUANT for quantification)
  - NONSPEC_INF
  - NONSPEC_OLD
- No counterparts to ACC-GEN or ACC-SIT as these belong in the main DRS by definition
### Distribution of tags

<table>
<thead>
<tr>
<th>info status</th>
<th>freq</th>
<th>info status</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>new</td>
<td>5704</td>
<td>non_spec</td>
<td>4457</td>
</tr>
<tr>
<td>kind</td>
<td>1177</td>
<td>non_spec_inf</td>
<td>405</td>
</tr>
<tr>
<td>acc_gen</td>
<td>3741</td>
<td>non_spec_old</td>
<td>1795</td>
</tr>
<tr>
<td>acc_inf</td>
<td>2570</td>
<td>quant</td>
<td>2004</td>
</tr>
<tr>
<td>acc_sit</td>
<td>883</td>
<td></td>
<td></td>
</tr>
<tr>
<td>old</td>
<td>34070</td>
<td></td>
<td></td>
</tr>
<tr>
<td>old_inact</td>
<td>1373</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Interannotator agreement

Towards the end of the NT tagging projects, kappa values were around 0.8 (after long periods of weekly meetings)
Towards the end of the NT tagging projects, kappa values were around 0.8 (after long periods of weekly meetings)

New project: Caesar’s Gallic War
Interannotator agreement

- Towards the end of the NT tagging projects, kappa values were around 0.8 (after long periods of weekly meetings)
- New project: Caesar’s Gallic War
- Supervised tagging of 8 chapters (ca. 400 taggables)
Towards the end of the NT tagging projects, kappa values were around 0.8 (after long periods of weekly meetings)

New project: Caesar’s Gallic War

Supervised tagging of 8 chapters (ca. 400 taggables)

Unsupervised tagging of 5 chapters (ca. 250 taggables)

$k = 0.66$ counting divergences in taggables

$k = 0.75$ on tags set by both annotators
Towards the end of the NT tagging projects, kappa values were around 0.8 (after long periods of weekly meetings)

New project: Caesar’s Gallic War

Supervised tagging of 8 chapters (ca. 400 taggables)

Unsupervised tagging of 5 chapters (ca. 250 taggables)

κ = 0.66 counting divergences in taggables

κ = 0.75 on tags set by both annotators

Decent; but much potential for more agreement, especially in taggables
Anaphoric links

- Old and inferrable discourse referents are equipped with anaphoric links
Anaphoric links

- Old and inferrable discourse referents are equipped with anaphoric links
- Link back to the previous mention of the referent if it is within the last 13 sentences, otherwise OLD_INACT
Anaphoric links

- Old and inferrable discourse referents are equipped with anaphoric links
- Link back to the previous mention of the referent if it is within the last 13 sentences, otherwise OLD\_INACT
- Relatively ‘objective’ measure, disagreements are usually errors
Anaphoric links

- Old and inferrable discourse referents are equipped with anaphoric links
- Link back to the previous mention of the referent if it is within the last 13 sentences, otherwise OLD_INACT
- Relatively ‘objective’ measure, disagreements are usually errors
- Enables studies of how choice of referential expression is affected by
  - Exact distance between anaphor and antecedent
  - Length and density of anaphoric chains
  - Relative saliency of discourse referents
IS annotation interface

- Provides guesses for discourse referents
- Tags and anaphoric links set by simple clicking

IS interface
Discourse experiments

- We have experimented with annotation of discourse structure
Discourse experiments

- We have experimented with annotation of discourse structure
- The annotation is broadly based on segmented discourse representation theory
Discourse experiments

- We have experimented with annotation of discourse structure
- The annotation is broadly based on segmented discourse representation theory
- Basic distinction between subordinating and coordinating relations
Season's greetings from PROIEL
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.