Syntactic discontinuities in treebanks and linguistic theories

Dag T. T. Haug

Treebanks & Linguistic Theories 14
December 12

Intro

- Starting point: a classical philologist interested in Greek and Latin word order

Intro

- Starting point: a classical philologist interested in Greek and Latin word order
  - I need data → I need a treebank

Intro

- Starting point: a classical philologist interested in Greek and Latin word order
  - I need data → I need a treebank
  - I need an annotation scheme → I need linguistic theory

Intro

- Starting point: a classical philologist interested in Greek and Latin word order
  - I need data → I need a treebank
  - I need an annotation scheme → I need linguistic theory
  - What can linguistic theory do for treebanks?

Intro

- Starting point: a classical philologist interested in Greek and Latin word order
  - I need data → I need a treebank
  - I need an annotation scheme → I need linguistic theory
  - What can linguistic theory do for treebanks?
- Resulting research also framed in theoretical linguistics
  - What can treebanks do for linguistic theory?

Intro

- Starting point: a classical philologist interested in Greek and Latin word order
  - I need data → I need a treebank
  - I need an annotation scheme → I need linguistic theory
  - What can linguistic theory do for treebanks?
- Resulting research also framed in theoretical linguistics
  - What can treebanks do for linguistic theory?
- Especially acute in a historical linguistic setting:
  - No other data sources

Intro

- Starting point: a classical philologist interested in Greek and Latin word order
  - I need data → I need a treebank
  - I need an annotation scheme → I need linguistic theory
  - What can linguistic theory do for treebanks?
- Resulting research also framed in theoretical linguistics
  - What can treebanks do for linguistic theory?
- Especially acute in a historical linguistic setting:
  - No other data sources
  - No other use for treebanks

Greek and Latin word order: state of the art

- The characteristic feature of languages like Greek and Latin are their free word order:
  - All permutations of S, V and O found with reasonable frequency

Greek and Latin word order: state of the art

- The characteristic feature of languages like Greek and Latin are their free word order:
  - All permutations of S, V and O found with reasonable frequency
  - Discontinuous constituency is common

Greek and Latin word order: state of the art

- The characteristic feature of languages like Greek and Latin are their free word order:
  - All permutations of S, V and O found with reasonable frequency
  - Discontinuous constituency is common
- The agreement stops there. . .

Word order in Luke/Acts

	Luke + Acts		Luke only	
	Rife(1933)	Davison(1989)	Rife(1933)	Kirk(2012)
VSO	15	20	9	14
SVO	50	56	19	13
SOV	9	8	8	5
VOS	3	4	2	3
OVS	6	6	1	1
OSV	1	1	0	1

Word order in Luke/Acts

	Luke + Acts		Luke only	
	Rife(1933)	Davison(1989)	Rife(1933)	Kirk(2012)
VSO	15	20	9	14
SVO	50	56	19	13
SOV	9	8	8	5
VOS	3	4	2	3
OVS	6	6	1	1
OSV	1	1	0	1

- All the authors claim to report the same basic fact: word order in declarative main clauses where there is an NP subject and object

Word order in Luke/Acts

	Luke + Acts		Luke only	
	Rife(1933)	Davison(1989)	Rife(1933)	Kirk(2012)
VSO	15	20	9	14
SVO	50	56	19	13
SOV	9	8	8	5
VOS	3	4	2	3
OVS	6	6	1	1
OSV	1	1	0	1

- All the authors claim to report the same basic fact: word order in declarative main clauses where there is an NP subject and object
- How can we agree on higher level analyses if the facts are so unclear?

What went wrong?

- Rife (1933): "The investigation was limited to main declarative clauses where both subject and object are substantives."

What went wrong?

- Rife (1933): "The investigation was limited to main declarative clauses where both subject and object are substantives."
  - Which text?

What went wrong?

- Rife (1933): "The investigation was limited to main declarative clauses where both subject and object are substantives."
  - Which text?
  - How are terms defined?

What went wrong?

- Rife (1933): "The investigation was limited to main declarative clauses where both subject and object are substantives."
  - Which text?
  - How are terms defined?
- Davison (1989): "clauses . . . which contained at least one nominative noun, one accusative noun and one indicative verb . . . Verbs normally followed by a genitive or a dative were traced using a concordance"
  - Same problem, although the text is at least available electronically

Being specific about it

- The criteria of Kirk (2012)

Being specific about it

- The criteria of Kirk (2012)
  - The clause contains at least an S(ubject), V(erb) and O(bject)

Being specific about it

- The criteria of Kirk (2012)
  - The clause contains at least an S(ubject), V(erb) and O(bject)
  - The clause is continuous

Being specific about it

- The criteria of Kirk (2012)
  - The clause contains at least an S(ubject), V(erb) and O(bject)
  - The clause is continuous
  - S and O are not embedded in a participial clause

Being specific about it

- The criteria of Kirk (2012)
  - The clause contains at least an S(ubject), V(erb) and O(bject)
  - The clause is continuous
  - S and O are not embedded in a participial clause
  - The verb assigns accusative, genitive, or dative to an argument that is a patient or theme

Being specific about it

- The criteria of Kirk (2012)
  - The clause contains at least an S(ubject), V(erb) and O(bject)
  - The clause is continuous
  - S and O are not embedded in a participial clause
  - The verb assigns accusative, genitive, or dative to an argument that is a patient or theme
  - The V consists of one word (no periphrastic forms, modal embeddings or light verbs)

Being specific about it

- The criteria of Kirk (2012)
  - The clause contains at least an S(ubject), V(erb) and O(bject)
  - The clause is continuous
  - S and O are not embedded in a participial clause
  - The verb assigns accusative, genitive, or dative to an argument that is a patient or theme
  - The V consists of one word (no periphrastic forms, modal embeddings or light verbs)
  - S and O are determiner phrases (this includes nominalizations) or quantifier phrases, and not clausal

Being specific about it

- The criteria of Kirk (2012)
  - The clause contains at least an S(ubject), V(erb) and O(bject)
  - The clause is continuous
  - S and O are not embedded in a participial clause
  - The verb assigns accusative, genitive, or dative to an argument that is a patient or theme
  - The V consists of one word (no periphrastic forms, modal embeddings or light verbs)
  - S and O are determiner phrases (this includes nominalizations) or quantifier phrases, and not clausal
  - S and O are continuous strings

Being specific about it

- The criteria of Kirk (2012)
  - The clause contains at least an S(ubject), V(erb) and O(bject)
  - The clause is continuous
  - S and O are not embedded in a participial clause
  - The verb assigns accusative, genitive, or dative to an argument that is a patient or theme
  - The V consists of one word (no periphrastic forms, modal embeddings or light verbs)
  - S and O are determiner phrases (this includes nominalizations) or quantifier phrases, and not clausal
  - S and O are continuous strings
- Admirably explicit, but what are the chances of getting things right at first try?

Being specific about it

- The criteria of Kirk (2012)
  - The clause contains at least an S(ubject), V(erb) and O(bject)
  - The clause is continuous
  - S and O are not embedded in a participial clause
  - The verb assigns accusative, genitive, or dative to an argument that is a patient or theme
  - The V consists of one word (no periphrastic forms, modal embeddings or light verbs)
  - S and O are determiner phrases (this includes nominalizations) or quantifier phrases, and not clausal
  - S and O are continuous strings
- Admirably explicit, but what are the chances of getting things right at first try?
- Even if everything that should be included is included, things may have been excluded that should not have been

The cure

- We need treebanks and linguistic theory

The cure

- We need treebanks and linguistic theory
  - Treebanks for replicability and iterated search
  - Linguistic theory for annotation schemes, and to know what we are looking for

The cure

- We need treebanks and linguistic theory
  - Treebanks for replicability and iterated search
  - Linguistic theory for annotation schemes, and to know what we are looking for
- "annotations are no substitute for the understanding of a phenomenon. They are an encoding of that understanding." Zaenen (2006)

The cure

- We need treebanks and linguistic theory
  - Treebanks for replicability and iterated search
  - Linguistic theory for annotation schemes, and to know what we are looking for
- "annotations are no substitute for the understanding of a phenomenon. They are an encoding of that understanding." Zaenen (2006)
- There is a danger of encoding precisely what we want to test

Creating a treebank

- There are obvious upsides for a linguist in working within an established linguistic theory
  - Build on earlier work

Creating a treebank

- There are obvious upsides for a linguist in working within an established linguistic theory
  - Build on earlier work
  - Compare and contrast analyses for specific languages

Creating a treebank

- There are obvious upsides for a linguist in working within an established linguistic theory
  - Build on earlier work
  - Compare and contrast analyses for specific languages
  - Compare and contrast with other theories

Creating a treebank

- There are obvious upsides for a linguist in working within an established linguistic theory
  - Build on earlier work
  - Compare and contrast analyses for specific languages
  - Compare and contrast with other theories
- Unfortunately there are also downsides in treebanking with an established linguistic theory
  - Requires linguistically (very) sophisticated annotators

Creating a treebank

- There are obvious upsides for a linguist in working within an established linguistic theory
  - Build on earlier work
  - Compare and contrast analyses for specific languages
  - Compare and contrast with other theories
- Unfortunately there are also downsides in treebanking with an established linguistic theory
  - Requires linguistically (very) sophisticated annotators
  - Hard to use for outsiders

Creating a treebank

- There are obvious upsides for a linguist in working within an established linguistic theory
  - Build on earlier work
  - Compare and contrast analyses for specific languages
  - Compare and contrast with other theories
- Unfortunately there are also downsides in treebanking with an established linguistic theory
  - Requires linguistically (very) sophisticated annotators
  - Hard to use for outsiders
  - Risks circular confirmation of biases in the theory

Creating a treebank

- There are obvious upsides for a linguist in working within an established linguistic theory
  - Build on earlier work
  - Compare and contrast analyses for specific languages
  - Compare and contrast with other theories
- Unfortunately there are also downsides in treebanking with an established linguistic theory
  - Requires linguistically (very) sophisticated annotators
  - Hard to use for outsiders
  - Risks circular confirmation of biases in the theory
- So we need "theory-neutral" treebanks

Creating a treebank

- There are obvious upsides for a linguist in working within an established linguistic theory
  - Build on earlier work
  - Compare and contrast analyses for specific languages
  - Compare and contrast with other theories
- Unfortunately there are also downsides in treebanking with an established linguistic theory
  - Requires linguistically (very) sophisticated annotators
  - Hard to use for outsiders
  - Risks circular confirmation of biases in the theory
- So we need "theory-neutral" treebanks

Creating a treebank

- There are obvious upsides for a linguist in working within an established linguistic theory
  - Build on earlier work
  - Compare and contrast analyses for specific languages
  - Compare and contrast with other theories
- Unfortunately there are also downsides in treebanking with an established linguistic theory
  - Requires linguistically (very) sophisticated annotators
  - Hard to use for outsiders
  - Risks circular confirmation of biases in the theory
- So we need "theory-neutral" treebanks
- But the danger with that is that we may contribute to the (already too wide) gulf between theoretical work and corpus linguistics

Principles of annotation

Two conflicting constraints on annotation

1. Encode enough structure to enable reconstruction of theoretically motivated structures

Principles of annotation

Two conflicting constraints on annotation

1. Encode enough structure to enable reconstruction of theoretically motivated structures
2. Encode no more structure than is common to all frameworks

Principles of annotation

Two conflicting constraints on annotation

1. Encode enough structure to enable reconstruction of theoretically motivated structures
2. Encode no more structure than is common to all frameworks

1. is obvious if the goal is theoretical adequacy

Principles of annotation

Two conflicting constraints on annotation

1. Encode enough structure to enable reconstruction of theoretically motivated structures
2. Encode no more structure than is common to all frameworks

1. is obvious if the goal is theoretical adequacy
2. is desirable to minimize the assumptions that go into the annotation and hence cannot be tested using the corpus

Principles of annotation

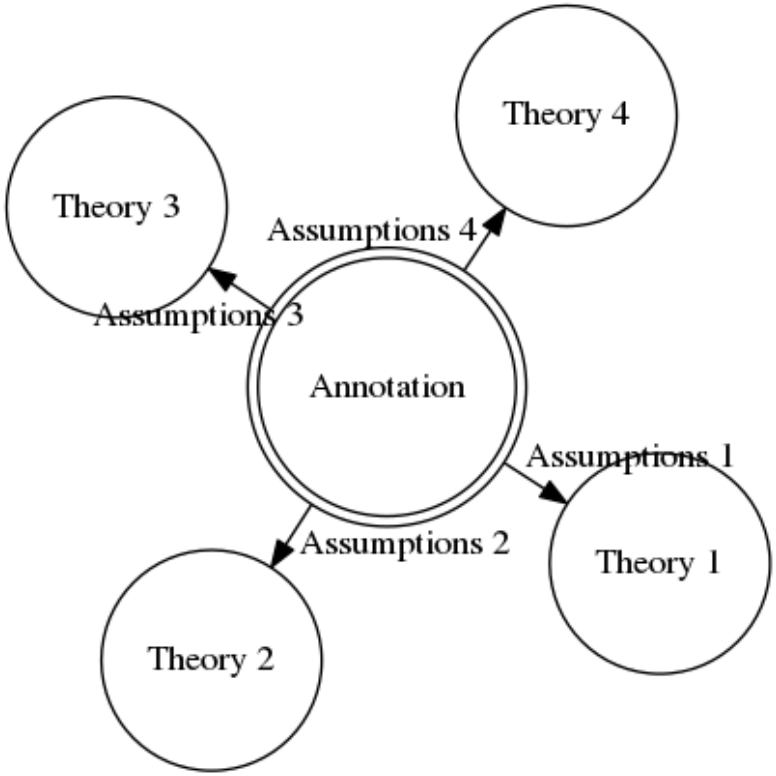
Two conflicting constraints on annotation

1. Encode enough structure to enable reconstruction of theoretically motivated structures
2. Encode no more structure than is common to all frameworks

1. is obvious if the goal is theoretical adequacy
2. is desirable to minimize the assumptions that go into the annotation and hence cannot be tested using the corpus

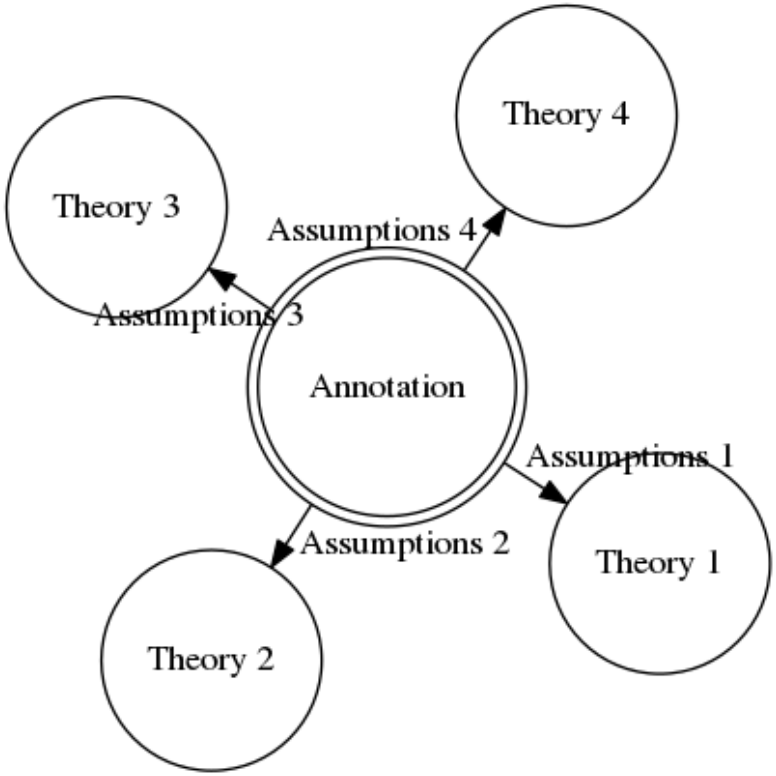
- To find "what is common to all frameworks" it may be necessary to choose concepts that are primary in one theory but derived in others (e.g. grammatical relations in LFG or DG vs. LTAG or CCG)

The ideal situation



- In the context of word order studies, the added assumptions will typically be about phrase structure (rules and constraints on rules)

The ideal situation



- In the context of word order studies, the added assumptions will typically be about phrase structure (rules and constraints on rules)
- Conversion to a theoretically motivated structure can now be seen as testing of the underlying assumptions

A naturally ocurring parallel corpus

- The New Testament in its Greek original and Latin, Gothic, Classical Armenian and OCS translations

A naturally ocurring parallel corpus

- The New Testament in its Greek original and Latin, Gothic, Classical Armenian and OCS translations
- The NT translations are the oldest attestations of Armenian and OCS, and virtually the only attestation of Gothic

A naturally ocurring parallel corpus

- The New Testament in its Greek original and Latin, Gothic, Classical Armenian and OCS translations
- 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

A naturally ocurring parallel corpus

- The New Testament in its Greek original and Latin, Gothic, Classical Armenian and OCS translations
- 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
- So these are important texts, and they are parallel texts

Later extensions

- Classical Greek and Latin:
  - Herodotus
  - Gallic War, Letters to Atticus, De officiis

Later extensions

- Classical Greek and Latin:
  - Herodotus
  - Gallic War, Letters to Atticus, De officiis
- Post-classical Greek and Latin
  - Sphrantzes' Chronicles
  - Peregrinatio Aetheriae
  - Palladius' De Agricultura

Other corpora in the same syntactic annotation scheme

- Poetic Edda (Greinir skáldskapar)

Other corpora in the same syntactic annotation scheme

- Poetic Edda (Greinir skáldskapar)
- Old Norwegian (Medieval Nordic Text Archive)

Other corpora in the same syntactic annotation scheme

- Poetic Edda (Greinir skáldskapar)
- Old Norwegian (Medieval Nordic Text Archive)
- Medieval English and Romance (ISWOC, Oslo)

Other corpora in the same syntactic annotation scheme

- Poetic Edda (Greinir skáldskapar)
- Old Norwegian (Medieval Nordic Text Archive)
- Medieval English and Romance (ISWOC, Oslo)
- Old Slavic texts (Tromsø)

Many-layered annotation

- Morphology

Many-layered annotation

- Morphology
- Syntax

Many-layered annotation

- Morphology
- Syntax
- Semantics and other customised annotation (e.g. animacy)

Many-layered annotation

- Morphology
- Syntax
- Semantics and other customised annotation (e.g. animacy)
- Givenness

Many-layered annotation

- Morphology
- Syntax
- Semantics and other customised annotation (e.g. animacy)
- Givenness
- Experimental discourse structure annotation

Syntactic annotation scheme

- Dependency grammar
  - practicable, easy to teach
  - no assumptions about word order/constituency

Syntactic annotation scheme

- Dependency grammar
  - practicable, easy to teach
  - no assumptions about word order/constituency
- But some aspects were felt too limiting already at the outset
  - We violate the unique head principle with secondary edges in e.g. control structures

Syntactic annotation scheme

- Dependency grammar
  - practicable, easy to teach
  - no assumptions about word order/constituency
- But some aspects were felt too limiting already at the outset
  - We violate the unique head principle with secondary edges in e.g. control structures
  - We use empty nodes in e.g. ellipsis structures
- Experience now vindicates these choices

Syntactic annotation scheme

- Dependency grammar
  - practicable, easy to teach
  - no assumptions about word order/constituency
- But some aspects were felt too limiting already at the outset
  - We violate the unique head principle with secondary edges in e.g. control structures
  - We use empty nodes in e.g. ellipsis structures
- Experience now vindicates these choices
- For standardization a UD version is available

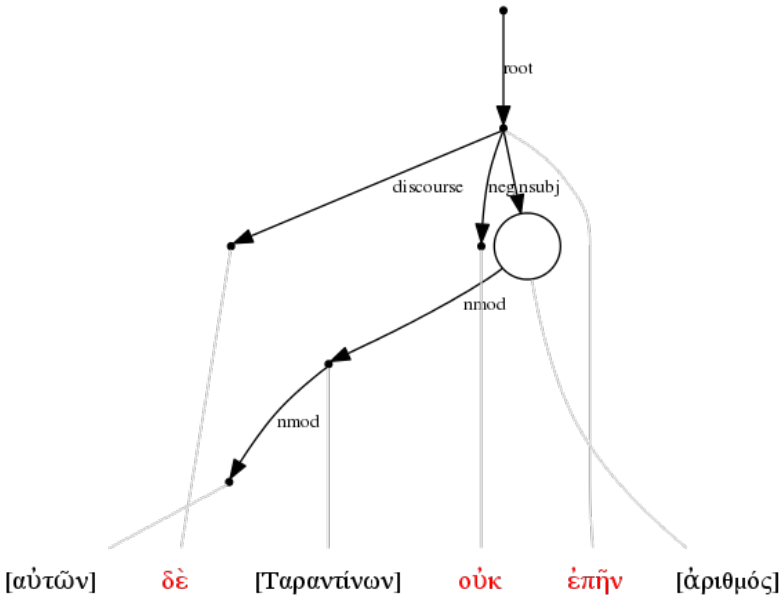
Two questions

1. What can the raw treebank data tell us about the theoretical power needed to capture the data?

Two questions

1. What can the raw treebank data tell us about the theoretical power needed to capture the data?
2. Do treebanks give us what we need to construct theoretically motivated representations of syntactic discontinuities?

Gap degree: the concept



Gap degree (trees) across treebanks in the UD corpora

	0	1	2	3		0	1	2	3
Ancient_Greek	0.368	0.548	0.079	0.005	Hungarian	0.748	0.232	0.016	0.003
Ancient_Greek-PROIEL	0.605	0.351	0.038	0.004	Indonesian	0.981	0.019	0.000	0.000
Arabic	0.918	0.081	0.001	0.000	Irish	0.872	0.126	0.002	0.000
Basque	0.663	0.287	0.045	0.005	Italian	0.961	0.036	0.003	0.000
Bulgarian	0.972	0.028	0.000	0.000	Japanese-KTC	1.000	0.000	0.000	0.000
Croatian	0.925	0.073	0.002	0.000	Latin	0.538	0.416	0.044	0.002
Czech	0.874	0.122	0.004	0.000	Latin-ITT	0.628	0.353	0.018	0.001
Danish	0.772	0.224	0.004	0.000	Latin-PROIEL	0.699	0.274	0.024	0.002
Dutch	0.691	0.286	0.022	0.000	Norwegian	0.923	0.076	0.001	0.000
English	0.950	0.046	0.003	0.000	Old_Church_Slavonic	0.784	0.203	0.012	0.000
Estonian	0.994	0.006	0.000	0.000	Persian	0.949	0.051	0.000	0.000
Finnish	0.923	0.075	0.002	0.000	Polish	0.997	0.003	0.000	0.000
Finnish-FTB	0.932	0.066	0.002	0.000	Portuguese	0.816	0.140	0.038	0.005
French	0.876	0.111	0.011	0.002	Romanian	0.886	0.106	0.006	0.002
German	0.879	0.112	0.007	0.001	Slovenian	0.864	0.128	0.008	0.000
Gothic	0.761	0.224	0.012	0.002	Spanish	0.939	0.060	0.000	0.000
Greek	0.721	0.252	0.026	0.001	Swedish	0.972	0.028	0.000	0.000
Hebrew	1.000	0.000	0.000	0.000	Tamil	0.978	0.022	0.000	0.000
Hindi	0.864	0.133	0.003	0.000					

Gap degree (trees) without punctuation

	0	1	2	3		0	1	2	3
Ancient_Greek	0.372	0.547	0.076	0.004	Hungarian	0.815	0.172	0.010	0.002
Ancient_Greek-PROIEL	0.605	0.351	0.038	0.004	Indonesian	0.994	0.006	0.000	0.000
Arabic	0.945	0.055	0.000	0.000	Irish	0.883	0.117	0.000	0.000
Basque	0.771	0.215	0.013	0.000	Italian	0.988	0.011	0.000	0.000
Bulgarian	0.972	0.028	0.000	0.000	Japanese-KTC	1.000	0.000	0.000	0.000
Croatian	0.936	0.063	0.001	0.000	Latin	0.551	0.409	0.038	0.002
Czech	0.876	0.121	0.003	0.000	Latin-ITT	0.631	0.350	0.018	0.001
Danish	0.750	0.250	0.000	0.000	Latin-PROIEL	0.699	0.274	0.024	0.002
Dutch	0.831	0.156	0.012	0.000	Norwegian	0.926	0.074	0.000	0.000
English	1.000	0.000	0.000	0.000	Old_Church_Slavonic	0.784	0.203	0.012	0.000
Estonian	0.992	0.008	0.000	0.000	Persian	0.955	0.045	0.000	0.000
Finnish	0.924	0.074	0.001	0.000	Polish	0.997	0.003	0.000	0.000
Finnish-FTB	0.933	0.065	0.002	0.000	Portuguese	0.838	0.149	0.013	0.000
French	0.960	0.039	0.001	0.000	Romanian	0.892	0.100	0.006	0.002
German	0.926	0.074	0.001	0.000	Slovenian	0.869	0.122	0.008	0.001
Gothic	0.761	0.224	0.012	0.002	Spanish	0.964	0.036	0.000	0.000
Greek	0.782	0.212	0.007	0.000	Swedish	0.973	0.027	0.000	0.000
Hebrew	1.000	0.000	0.000	0.000	Tamil	0.987	0.013	0.000	0.000
Hindi	0.864	0.132	0.003	0.000					

Gap degree (trees) without punctuation

	0	1	2	3		0	1	2	3
Ancient_Greek	0.372	0.547	0.076	0.004	Hungarian	0.815	0.172	0.010	0.002
Ancient_Greek-PROIEL	0.605	0.351	0.038	0.004	Indonesian	0.994	0.006	0.000	0.000
Arabic	0.945	0.055	0.000	0.000	Irish	0.883	0.117	0.000	0.000
Basque	0.771	0.215	0.013	0.000	Italian	0.988	0.011	0.000	0.000
Bulgarian	0.972	0.028	0.000	0.000	Japanese-KTC	1.000	0.000	0.000	0.000
Croatian	0.936	0.063	0.001	0.000	Latin	0.551	0.409	0.038	0.002
Czech	0.876	0.121	0.003	0.000	Latin-ITT	0.631	0.350	0.018	0.001
Danish	0.750	0.250	0.000	0.000	Latin-PROIEL	0.699	0.274	0.024	0.002
Dutch	0.831	0.156	0.012	0.000	Norwegian	0.926	0.074	0.000	0.000
English	1.000	0.000	0.000	0.000	Old_Church_Slavonic	0.784	0.203	0.012	0.000
Estonian	0.992	0.008	0.000	0.000	Persian	0.955	0.045	0.000	0.000
Finnish	0.924	0.074	0.001	0.000	Polish	0.997	0.003	0.000	0.000
Finnish-FTB	0.933	0.065	0.002	0.000	Portuguese	0.838	0.149	0.013	0.000
French	0.960	0.039	0.001	0.000	Romanian	0.892	0.100	0.006	0.002
German	0.926	0.074	0.001	0.000	Slovenian	0.869	0.122	0.008	0.001
Gothic	0.761	0.224	0.012	0.002	Spanish	0.964	0.036	0.000	0.000
Greek	0.782	0.212	0.007	0.000	Swedish	0.973	0.027	0.000	0.000
Hebrew	1.000	0.000	0.000	0.000	Tamil	0.987	0.013	0.000	0.000
Hindi	0.864	0.132	0.003	0.000					

Immediate lesssons

- Careful with your commas!

Immediate lesssons

- Careful with your commas!
- Results in line with reports on Arabic (Kuhlmann, 2010), Czech (Kuhlmann & Nivre, 2006), Danish (Kuhlmann & Nivre, 2006), Slovene (Kuhlmann, 2010), Swedish (Gómez-rodríguez et al., 2009): gap degree 1 gives very little loss (<0.5%)

Immediate lesssons

- Careful with your commas!
- Results in line with reports on Arabic (Kuhlmann, 2010), Czech (Kuhlmann & Nivre, 2006), Danish (Kuhlmann & Nivre, 2006), Slovene (Kuhlmann, 2010), Swedish (Gómez-rodríguez et al., 2009): gap degree 1 gives very little loss (<0.5%)
- The same holds for Croatian, English, Estonian, Finnish, French, German, Hindi, Indonesian, Irish, Italian, Japanese, Norwegian, Persian, Polish, Spanish and Tamil

Immediate lesssons

- Careful with your commas!
- Results in line with reports on Arabic (Kuhlmann, 2010), Czech (Kuhlmann & Nivre, 2006), Danish (Kuhlmann & Nivre, 2006), Slovene (Kuhlmann, 2010), Swedish (Gómez-rodríguez et al., 2009): gap degree 1 gives very little loss (<0.5%)
- The same holds for Croatian, English, Estonian, Finnish, French, German, Hindi, Indonesian, Irish, Italian, Japanese, Norwegian, Persian, Polish, Spanish and Tamil
- Ancient Greek (Gothic, OCS), Basque, Dutch, Greek, Hungarian, Latin, Portuguese, Romanian and Slovenian: more than 0.5% loss

Immediate lesssons

- Careful with your commas!
- Results in line with reports on Arabic (Kuhlmann, 2010), Czech (Kuhlmann & Nivre, 2006), Danish (Kuhlmann & Nivre, 2006), Slovene (Kuhlmann, 2010), Swedish (Gómez-rodríguez et al., 2009): gap degree 1 gives very little loss (<0.5%)
- The same holds for Croatian, English, Estonian, Finnish, French, German, Hindi, Indonesian, Irish, Italian, Japanese, Norwegian, Persian, Polish, Spanish and Tamil
- Ancient Greek (Gothic, OCS), Basque, Dutch, Greek, Hungarian, Latin, Portuguese, Romanian and Slovenian: more than 0.5% loss
- Ancient Greek stands out (Mambrini & Passarotti, 2013)

Theoretical implications

- 'Standard' (linguistically well-developed) mildly context-sensitive grammar formalisms such as TAG and CCG do not generate structures with gap degree >1

Theoretical implications

- 'Standard' (linguistically well-developed) mildly context-sensitive grammar formalisms such as TAG and CCG do not generate structures with gap degree >1
- There are extensions (e.g. multiple-component tree-adjoining grammars), but we lose the benefits of working with well-developed theories for cross-linguistic comparison

Theoretical implications

- 'Standard' (linguistically well-developed) mildly context-sensitive grammar formalisms such as TAG and CCG do not generate structures with gap degree >1
- There are extensions (e.g. multiple-component tree-adjoining grammars), but we lose the benefits of working with well-developed theories for cross-linguistic comparison
- So it is reasonable to "step up" to unification-based formalisms

Theoretical implications

- 'Standard' (linguistically well-developed) mildly context-sensitive grammar formalisms such as TAG and CCG do not generate structures with gap degree >1
- There are extensions (e.g. multiple-component tree-adjoining grammars), but we lose the benefits of working with well-developed theories for cross-linguistic comparison
- So it is reasonable to "step up" to unification-based formalisms
- Lexical-functional grammar (LFG) is particularly well-developed when it comes to studies of discontinuous syntax

Lexical-functional grammar

- LFG analyses sentences in terms of a surface-oriented c(onstituent)-structure, a more abstract f(eature)-structure and the mapping between them

Lexical-functional grammar

- LFG analyses sentences in terms of a surface-oriented c(onstituent)-structure, a more abstract f(eature)-structure and the mapping between them
- Bröker et al. (1994): Dependency grammar is an LFG that only knows f-structure

Lexical-functional grammar

- LFG analyses sentences in terms of a surface-oriented c(onstituent)-structure, a more abstract f(eature)-structure and the mapping between them
- Bröker et al. (1994): Dependency grammar is an LFG that only knows f-structure
- A DG treebank is a perfect match if we want to study c-structures given particular f-structures!

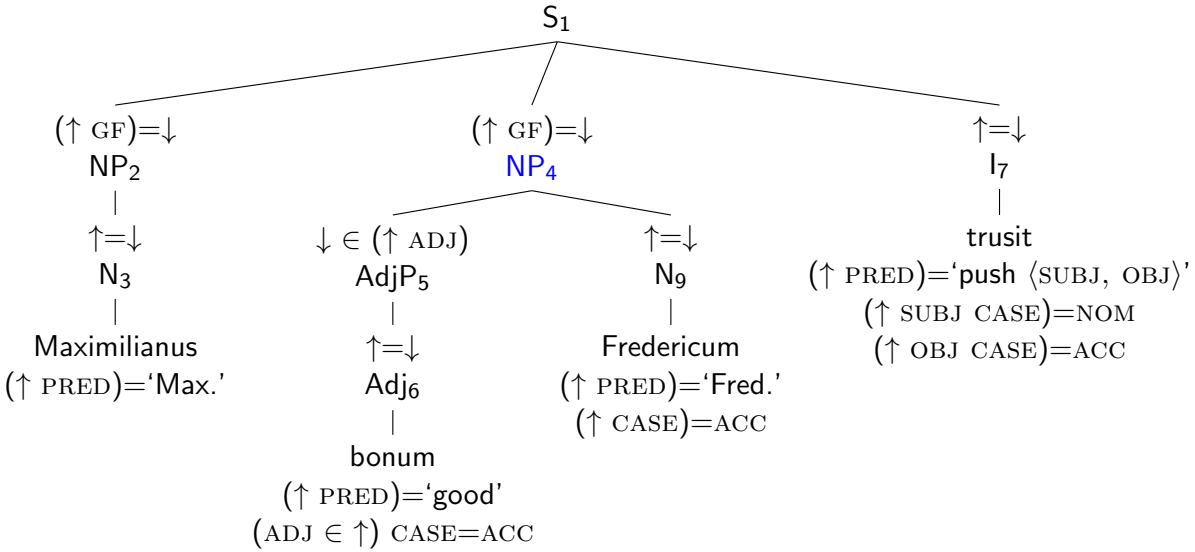
Lexical-functional grammar

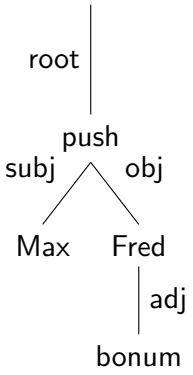
- LFG analyses sentences in terms of a surface-oriented c(onstituent)-structure, a more abstract f(eature)-structure and the mapping between them
- Bröker et al. (1994): Dependency grammar is an LFG that only knows f-structure
- A DG treebank is a perfect match if we want to study c-structures given particular f-structures!
- In practice, DG analyses are in between c- and f-structures:
  - Typically surface-oriented: the tokens are the nodes

Lexical-functional grammar

- LFG analyses sentences in terms of a surface-oriented c(onstituent)-structure, a more abstract f(eature)-structure and the mapping between them
- Bröker et al. (1994): Dependency grammar is an LFG that only knows f-structure
- A DG treebank is a perfect match if we want to study c-structures given particular f-structures!
- In practice, DG analyses are in between c- and f-structures:
  - Typically surface-oriented: the tokens are the nodes
  - But encodes abstract grammatical relations

An LFG analysis

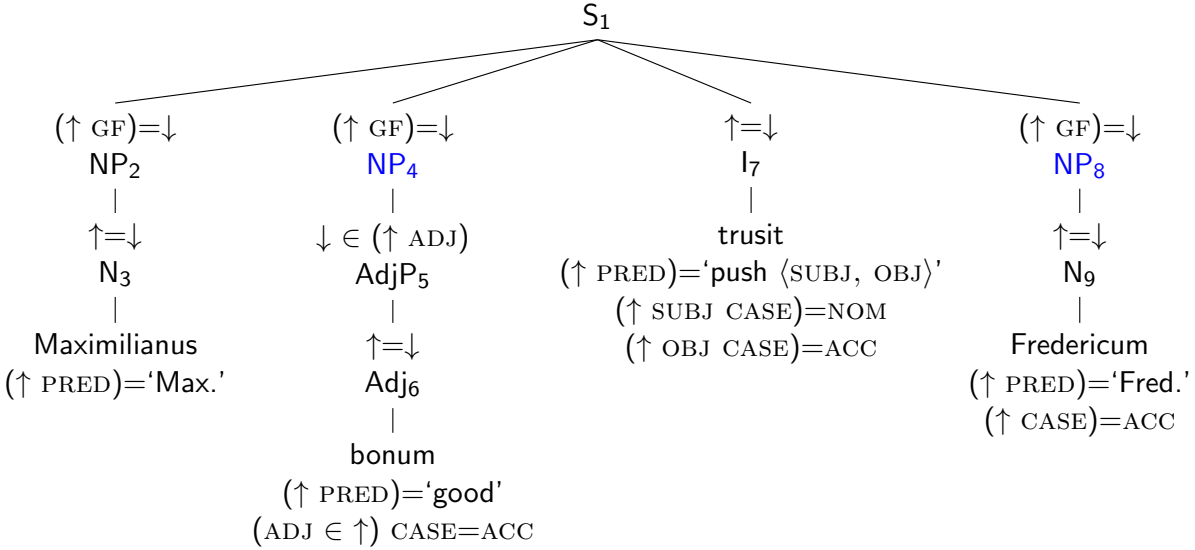


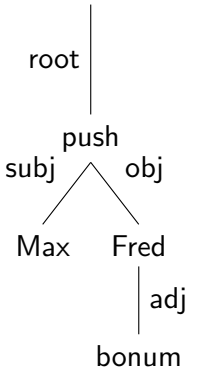


$$\begin{bmatrix} \text{PRED} & \text{'PUSH } \langle \text{SUBJ, OBJ} \rangle \text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'MAX.'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'FRED.'} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'GOOD'} \end{bmatrix} \right\} \end{bmatrix} \end{bmatrix}$$

- Dependency graph ≈ f-structure

An LFG analysis

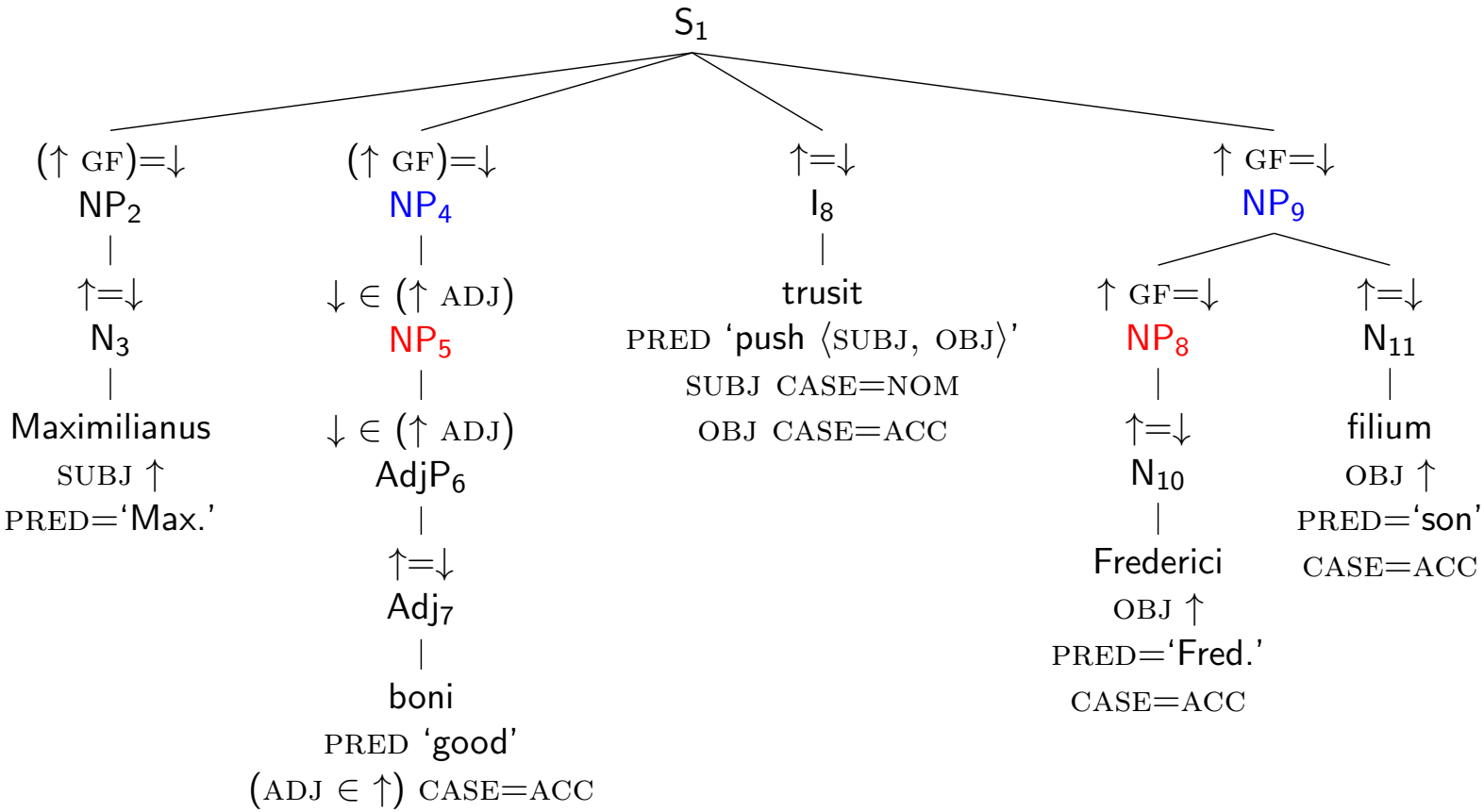




$$\begin{bmatrix} \text{PRED} & \text{'PUSH } \langle \text{SUBJ, OBJ} \rangle \text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'MAX.'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'FRED.'} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'GOOD'} \end{bmatrix} \right\} \end{bmatrix} \end{bmatrix}$$

- Dependency graph ≈ f-structure
- Discontinuities correspond to reentrancies (one f-structure corresponding to multiple phrase structure nodes)

LFG analysis: gap depth



- Deeper gaps require more rentrancies

Depth vs. degree (edges)

		Depth						
Universal dependencies	Degree	0	1	2	3	4	5	6+
	0	1416015	0	0	0	0	0	0
	1	0	58667	4833	592	128	30	43
	2	0	3475	294	49	8	2	0
	3	0	266	15	2	0	0	0
	4	0	40	3	0	0	0	0
	5	0	18	3	1	0	0	0
	6	0	4	1	0	0	0	0
	7	0	2	0	0	0	0	0

Depth vs. degree (edges)

		Depth						
Universal dependencies	Degree	0	1	2	3	4	5	6+
	0	1416015	0	0	0	0	0	0
	1	0	58667	4833	592	128	30	43
	2	0	3475	294	49	8	2	0
	3	0	266	15	2	0	0	0
	4	0	40	3	0	0	0	0
	5	0	18	3	1	0	0	0
	6	0	4	1	0	0	0	0
	7	0	2	0	0	0	0	0

UD-Ancient Greek		0	1	2	3	4	5	6
	0	65707	0	0	0	0	0	0
	1	0	16182	1548	187	21	3	1
	2	0	1259	116	16	3	0	0
	3	0	73	2	0	0	0	0
	4	0	2	0	0	0	0	0

UD-Ancient Greek-PROIEL		0	1	2	3	4	5	6+
	0	75129	0	0	0	0	0	0
	1	0	8900	596	65	13	2	8
	2	0	742	57	5	2	0	0
	3	0	88	5	0	0	0	0
	4	0	18	1	0	0	0	0
	5	0	4	0	0	0	0	0
	6	0	1	0	0	0	0	0
	7	0	1	0	0	0	0	0

Aiming too high

- For practical parsing purposes we could limit ourselves to gap degree 1 + gap degree 2 with depth 1 only

Aiming too high

- For practical parsing purposes we could limit ourselves to gap degree 1 + gap degree 2 with depth 1 only
- LFG as a theory as a theory could derive any gap degree and depth, but

Aiming too high

- For practical parsing purposes we could limit ourselves to gap degree 1 + gap degree 2 with depth 1 only
- LFG as a theory as a theory could derive any gap degree and depth, but
  - reflects the low complexity in a low number of reentrancies in the LFG analyses
  - offers a body of theoretical and cross-linguistic work to lean on

Aiming too high

- For practical parsing purposes we could limit ourselves to gap degree 1 + gap degree 2 with depth 1 only
- LFG as a theory as a theory could derive any gap degree and depth, but
  - reflects the low complexity in a low number of reentrancies in the LFG analyses
  - offers a body of theoretical and cross-linguistic work to lean on
- To connect we need to derive c-structures from the dependencies (Haug, 2012)

Order domains (Adapted from Bröker 1998)

Definition

The order domain $\mathcal{D}_w$ of a word w is the largest subset of $\mathcal{W}$ such that

1. $w \in \mathcal{D}_w$

Order domains (Adapted from Bröker 1998)

Definition

The order domain $\mathcal{D}_w$ of a word w is the largest subset of $\mathcal{W}$ such that

1. $w \in \mathcal{D}_w$
2. all words in $\mathcal{D}_w$ are dominated by w

Order domains (Adapted from Bröker 1998)

Definition

The order domain $\mathcal{D}_w$ of a word w is the largest subset of $\mathcal{W}$ such that

1. $w \in \mathcal{D}_w$
2. all words in $\mathcal{D}_w$ are dominated by w
3. $\mathcal{D}_w$ is continuous, i.e. for any two words in $\mathcal{D}_w$, all words in between are also contained in $\mathcal{D}_w$

Order domains (Adapted from Bröker 1998)

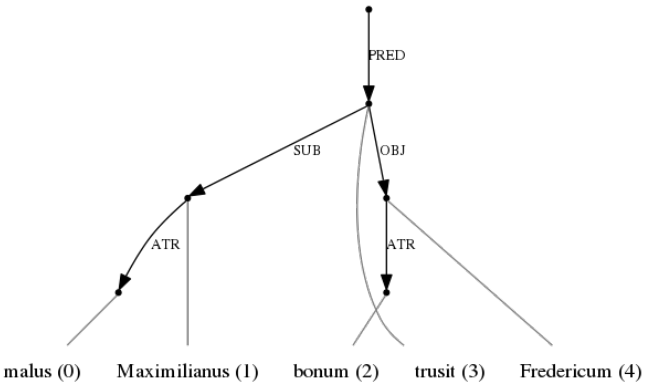
Definition

The order domain $\mathcal{D}_w$ of a word w is the largest subset of $\mathcal{W}$ such that

1. $w \in \mathcal{D}_w$
2. all words in $\mathcal{D}_w$ are dominated by w
3. $\mathcal{D}_w$ is continuous, i.e. for any two words in $\mathcal{D}_w$, all words in between are also contained in $\mathcal{D}_w$

- Intuitively, the order domain corresponds to all of the node's dependents that are not 'displaced'

Example



- malus → {malus}
- Maximilianus → {malus,Max.}
- bonum → {bonum}
- trusit → {malus,Max.,bonum,trusit,Fred.}
- Fredericum → {Fred.}

Order domain structures

Definition

The order domain structure $\mathcal{O}$ of a sentence S with the words $\mathcal{W}$ is the set of order domains of all words $w \in \mathcal{W}$.

Order domain structures

Definition

The order domain structure $\mathcal{O}$ of a sentence S with the words $\mathcal{W}$ is the set of order domains of all words $w \in \mathcal{W}$.

Order domain structures

Definition

The order domain structure $\mathcal{O}$ of a sentence S with the words $\mathcal{W}$ is the set of order domains of all words $w \in \mathcal{W}$.

- Subset inclusion correponds to phrase structure dominance

Order domain structures

Definition

The order domain structure $\mathcal{O}$ of a sentence S with the words $\mathcal{W}$ is the set of order domains of all words $w \in \mathcal{W}$.

- Subset inclusion correponds to phrase structure dominance
- Each order domain is continuous, so we have a total precedence relation

Order domain structures

Definition

The order domain structure $\mathcal{O}$ of a sentence S with the words $\mathcal{W}$ is the set of order domains of all words $w \in \mathcal{W}$.

- Subset inclusion correponds to phrase structure dominance
- Each order domain is continuous, so we have a total precedence relation
- $\mathcal{O}$ is an ordered tree

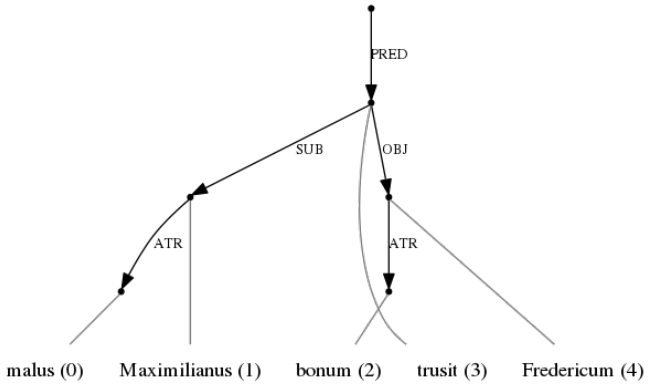
Order domain structures

Definition

The order domain structure $\mathcal{O}$ of a sentence S with the words $\mathcal{W}$ is the set of order domains of all words $w \in \mathcal{W}$.

- Subset inclusion correponds to phrase structure dominance
- Each order domain is continuous, so we have a total precedence relation
- $\mathcal{O}$ is an ordered tree
- A useful intermediate structure between PS and DS

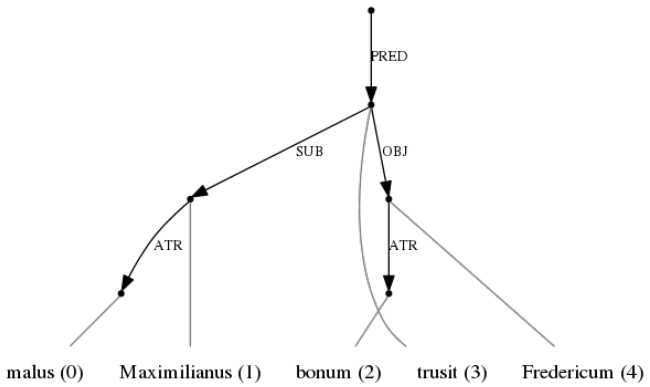
Example

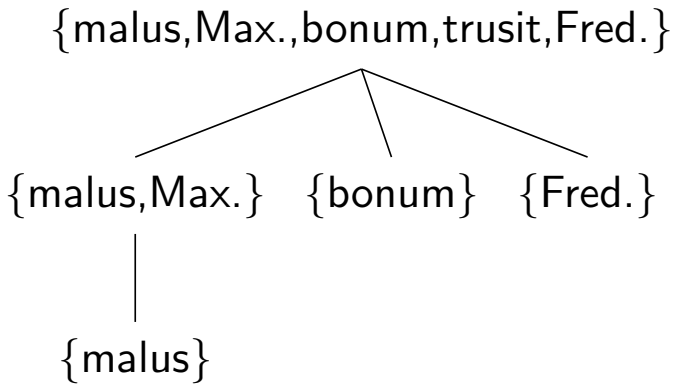


{malus,Max.,bonum,trusit,Fred.},
{malus,Max.} {malus}, {bonum}, {Fred.}

- We order the order domains by subset inclusion and precedence

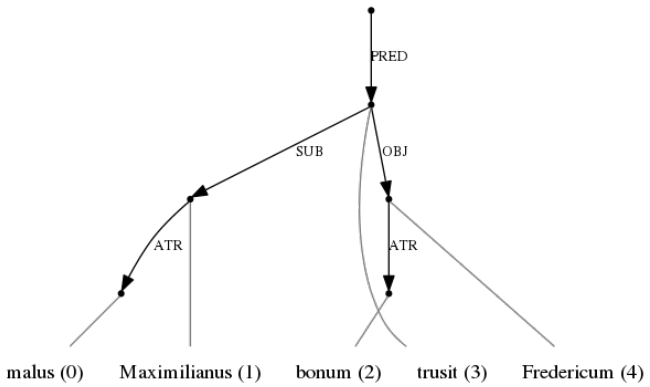
Example

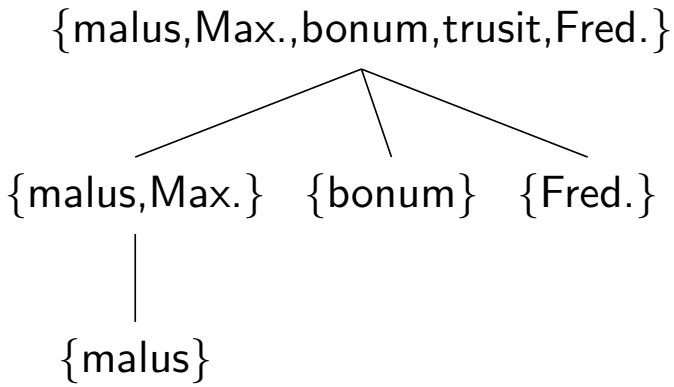




- We order the order domains by subset inclusion and precedence
- Problem: no way to retrieve the dependency of *bonum* on *Fredericum*

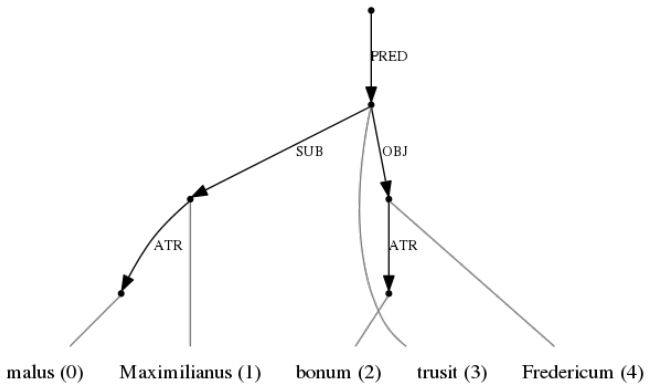
Example

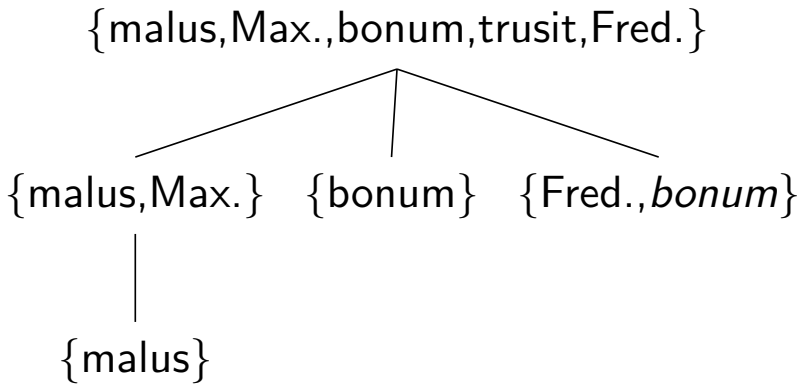




- We order the order domains by subset inclusion and precedence
- Problem: no way to retrieve the dependency of *bonum* on *Fredericum*
- Solution: add a trace keeping track of the discontinuity

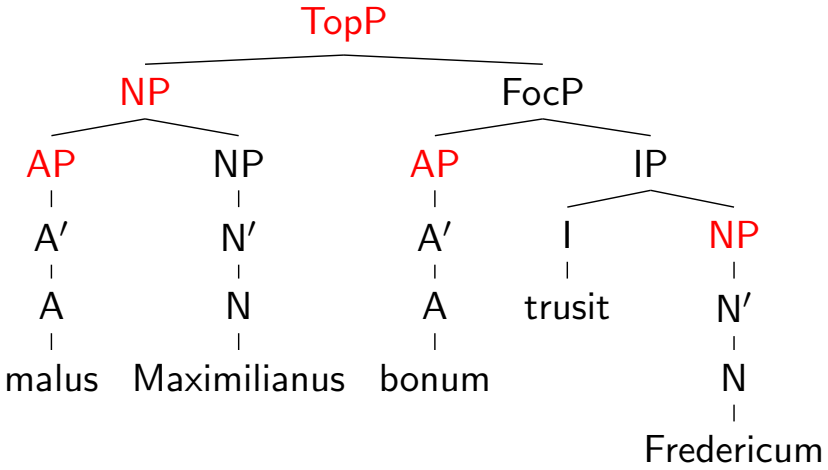
Example

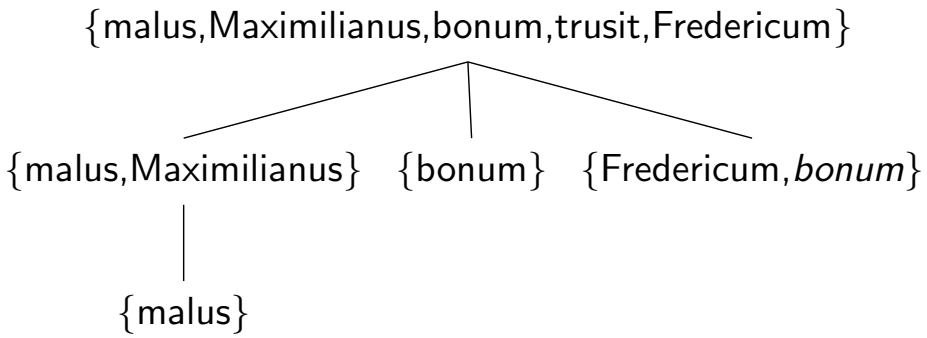




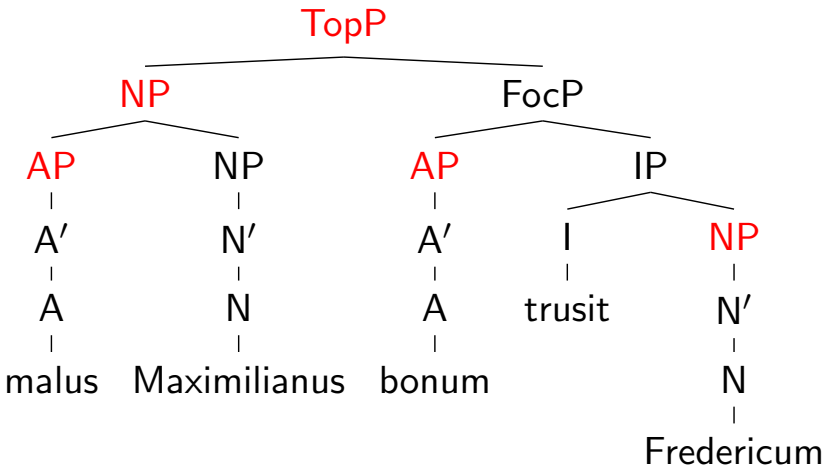
- We order the order domains by subset inclusion and precedence
- Problem: no way to retrieve the dependency of *bonum* on *Fredericum*
- Solution: add a trace keeping track of the discontinuity
- We get a structure that is implicitly present in the dependency graph. but is isomorphic to the expected phrase structure

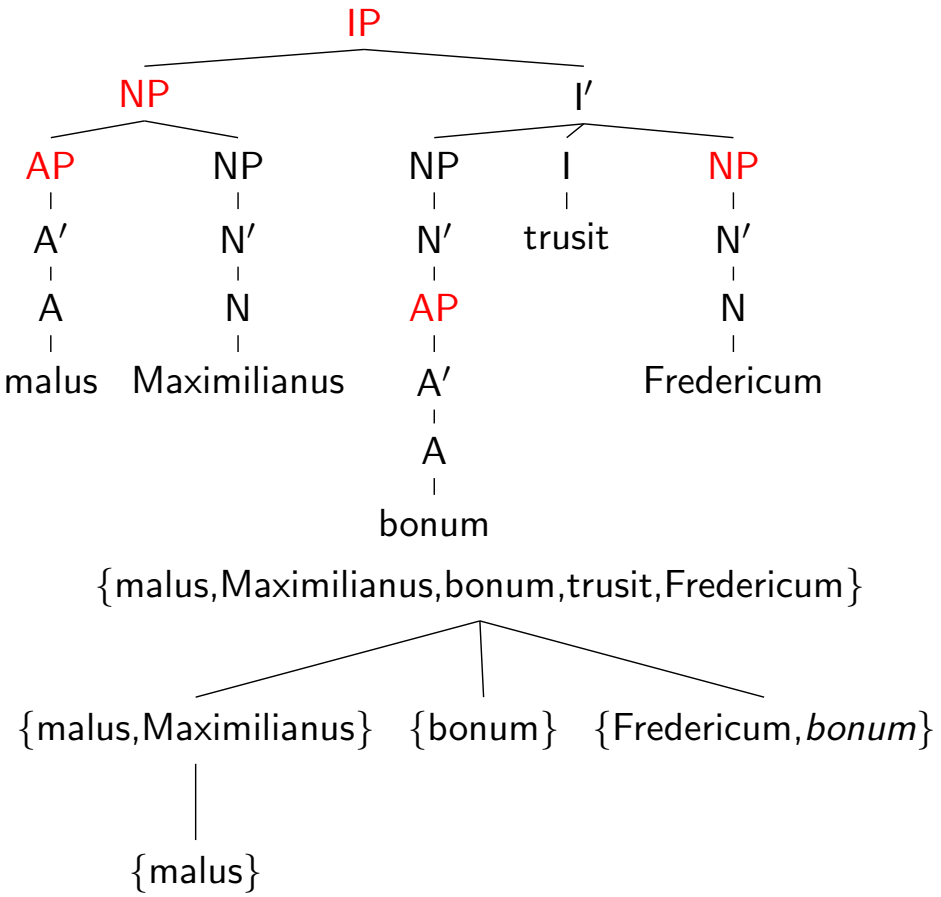
Some alternatives



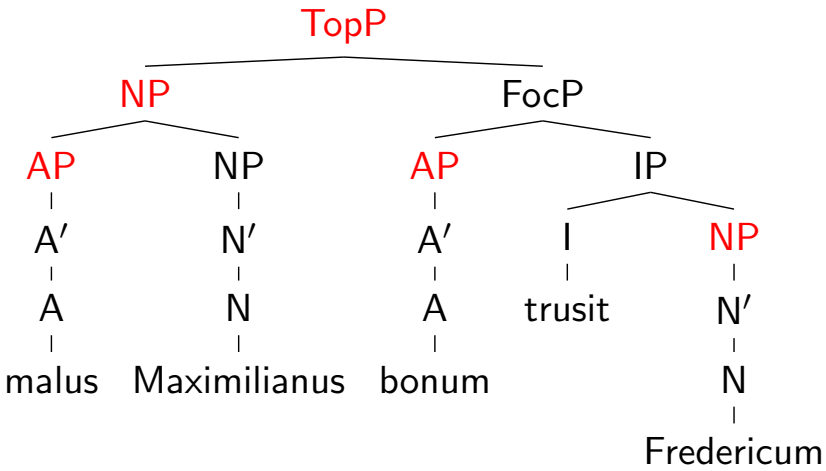


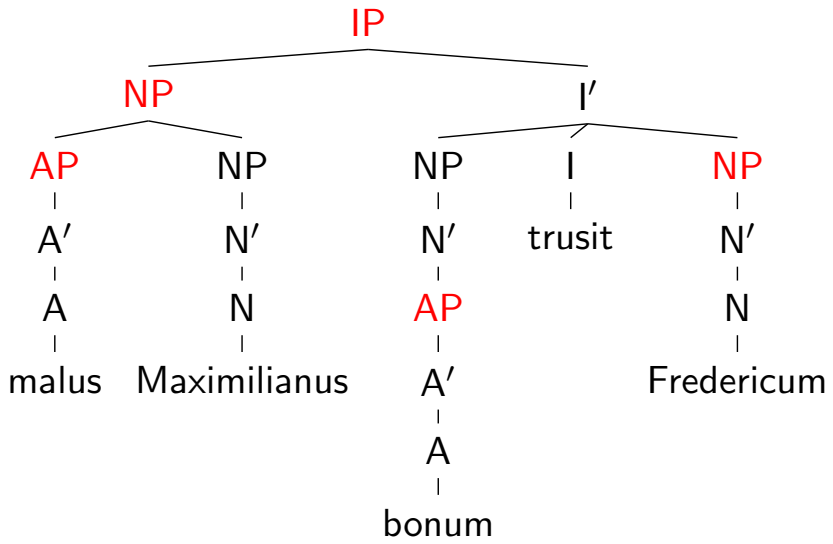
Some alternatives

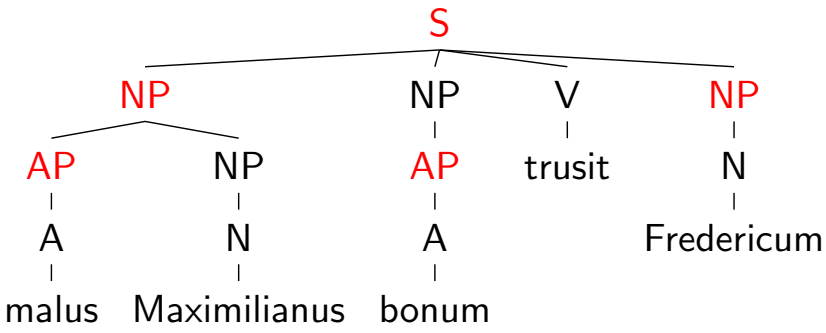


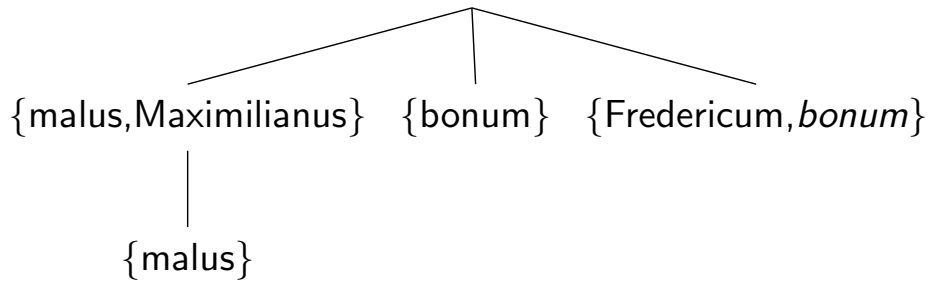


Some alternatives









Hypothesis testing

- The configurations of the maximal projections in these trees are all isomorphic to the order domain structure

Hypothesis testing

- The configurations of the maximal projections in these trees are all isomorphic to the order domain structure
- So we can view the creation of phrase structures from the dependency structure as an expansion of the order domain structure

Hypothesis testing

- The configurations of the maximal projections in these trees are all isomorphic to the order domain structure
- So we can view the creation of phrase structures from the dependency structure as an expansion of the order domain structure
- The task then is to determine the internal structure of each word's projection

Hypothesis testing

- The configurations of the maximal projections in these trees are all isomorphic to the order domain structure
- So we can view the creation of phrase structures from the dependency structure as an expansion of the order domain structure
- The task then is to determine the internal structure of each word's projection
- This must be done by injecting linguistic theory

Hypothesis testing

- The configurations of the maximal projections in these trees are all isomorphic to the order domain structure
- So we can view the creation of phrase structures from the dependency structure as an expansion of the order domain structure
- The task then is to determine the internal structure of each word's projection
- This must be done by injecting linguistic theory
- We can write constrained expansion rules and test them against the data

Why does this work?

- In Greek and Latin syntax, we know more about grammatical relations than about phrase structure

Why does this work?

- In Greek and Latin syntax, we know more about grammatical relations than about phrase structure
- This may be true for discontinuous syntax in general

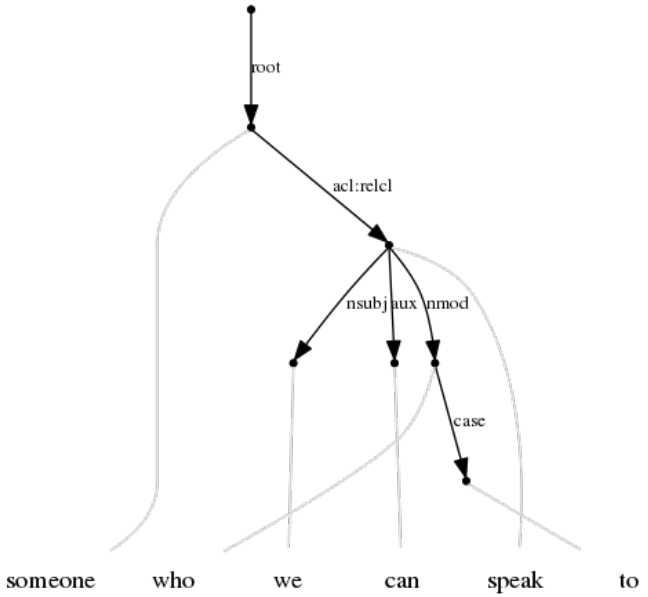
Why does this work?

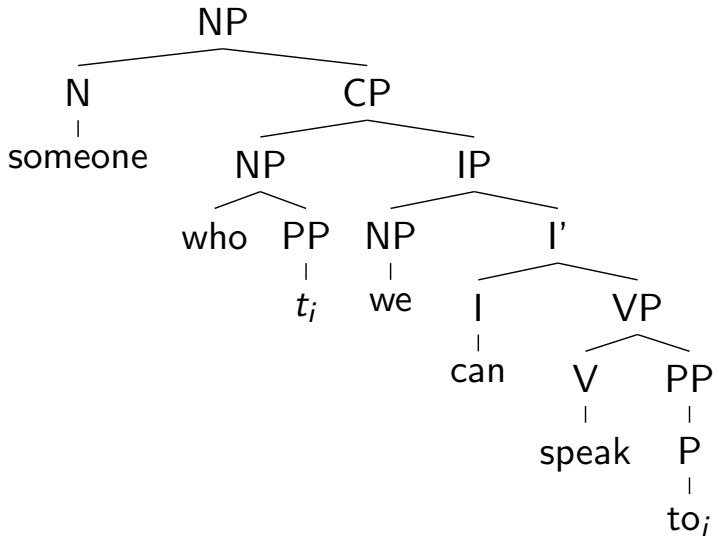
- In Greek and Latin syntax, we know more about grammatical relations than about phrase structure
- This may be true for discontinuous syntax in general
- DG here achieves the ideal situation: the treebank encodes our linguistic understanding, but does not make presuppositions about uncertain things

Why does this work?

- In Greek and Latin syntax, we know more about grammatical relations than about phrase structure
- This may be true for discontinuous syntax in general
- DG here achieves the ideal situation: the treebank encodes our linguistic understanding, but does not make presuppositions about uncertain things
- Function words are typically challenging, as they are often c-structure heads, but typically taken as dependents in DG analyses

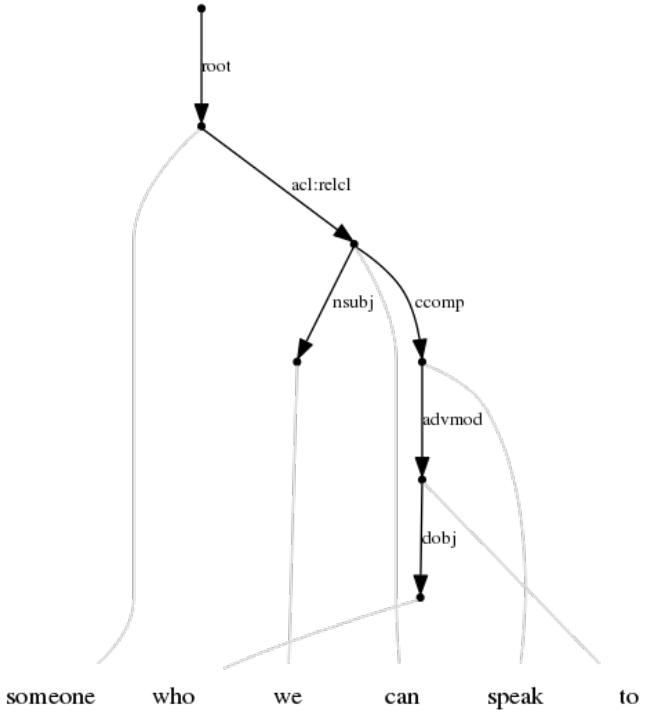
Constituency/function mismatch

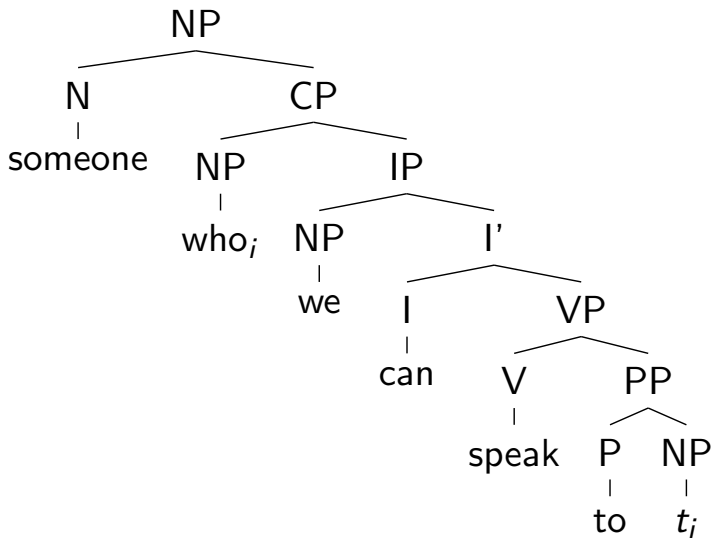




- UD makes adpositions dependents, giving us a final displaced dependent

Constituency/function mismatch





- The constituency evidence rather points to the clause-initial element begin displaced

Participle clauses

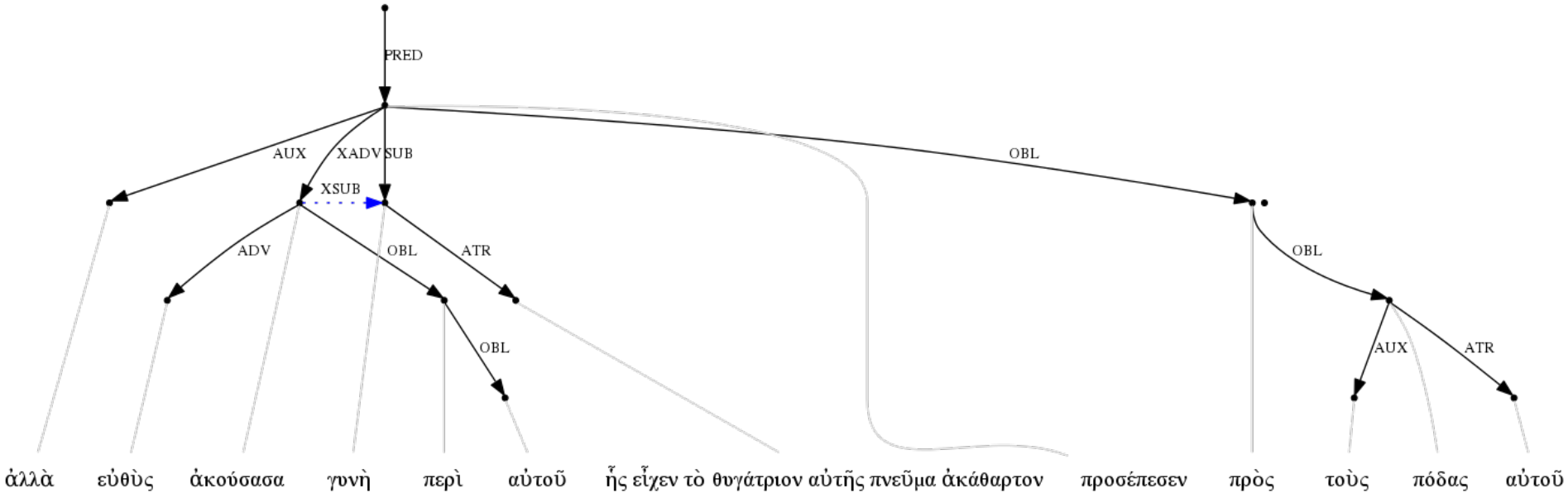
- Let us now look closer at one of Kirk's exclusion cases:
  - S and O are not embedded in a participial clause

Participle clauses

- Let us now look closer at one of Kirk's exclusion cases:
  - S and O are not embedded in a participial clause

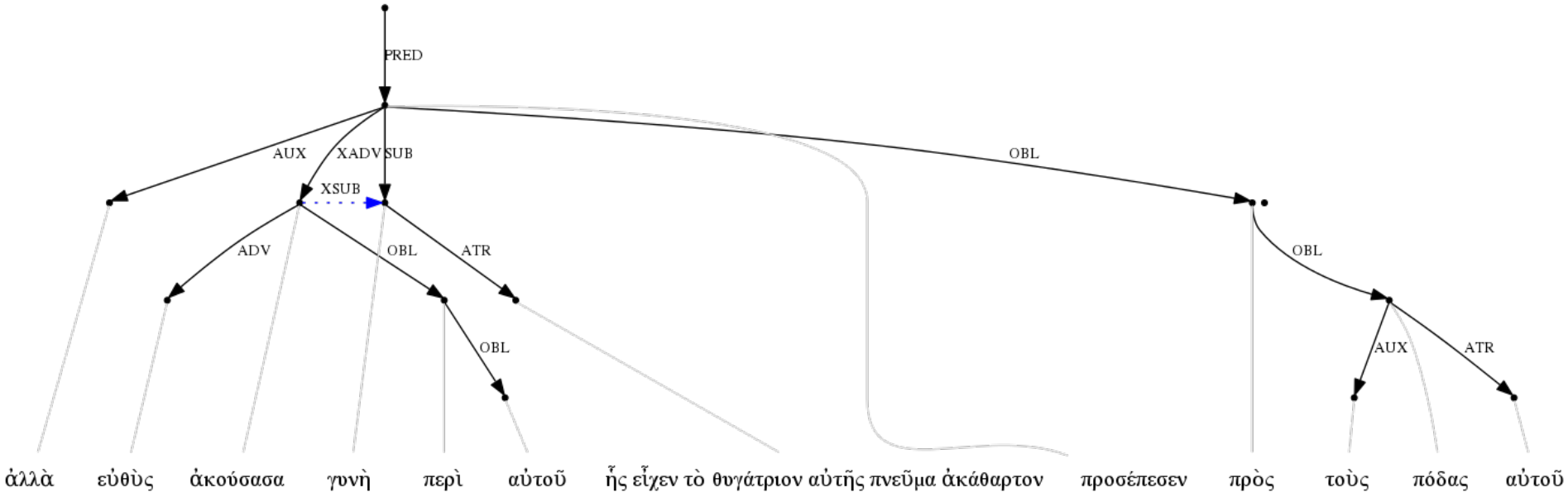
(1) akousasa gunê peri autou
hearing woman about him
hês eikhen to thugatrion autês pneuma akatharton
whose daughter had an evil spirit
prosepesen pros tous podas autou
fell down before his feet (Mark 7.25)

The analysis



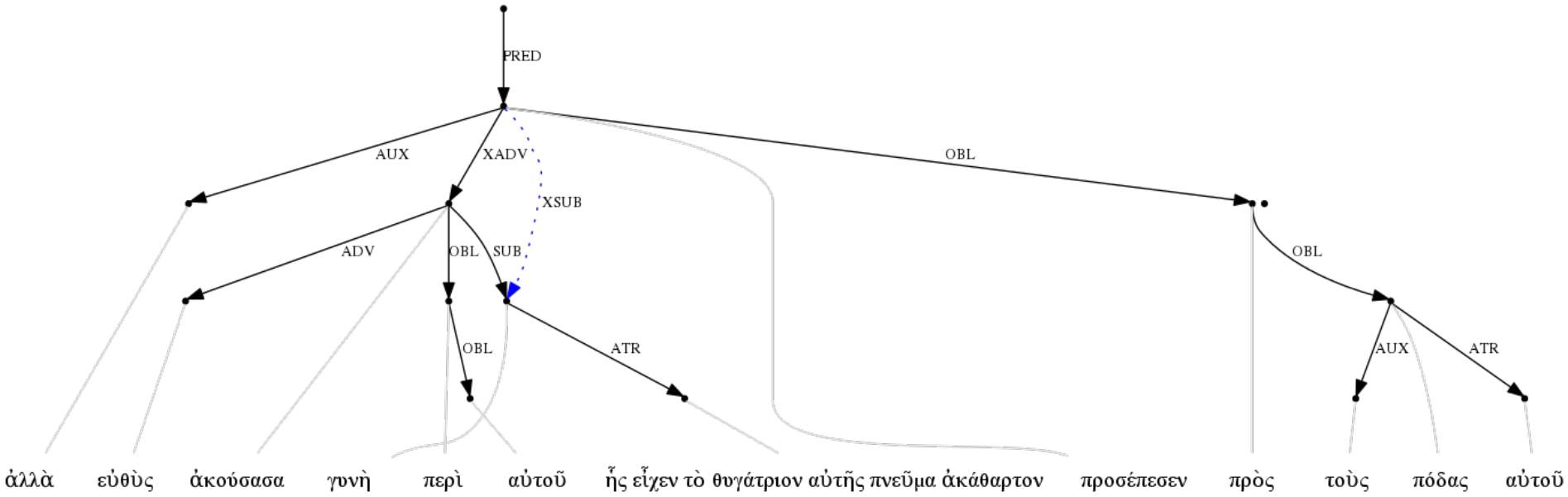
- On this analysis, it is legitimate to analyze the clause as S - V - OBL

The analysis



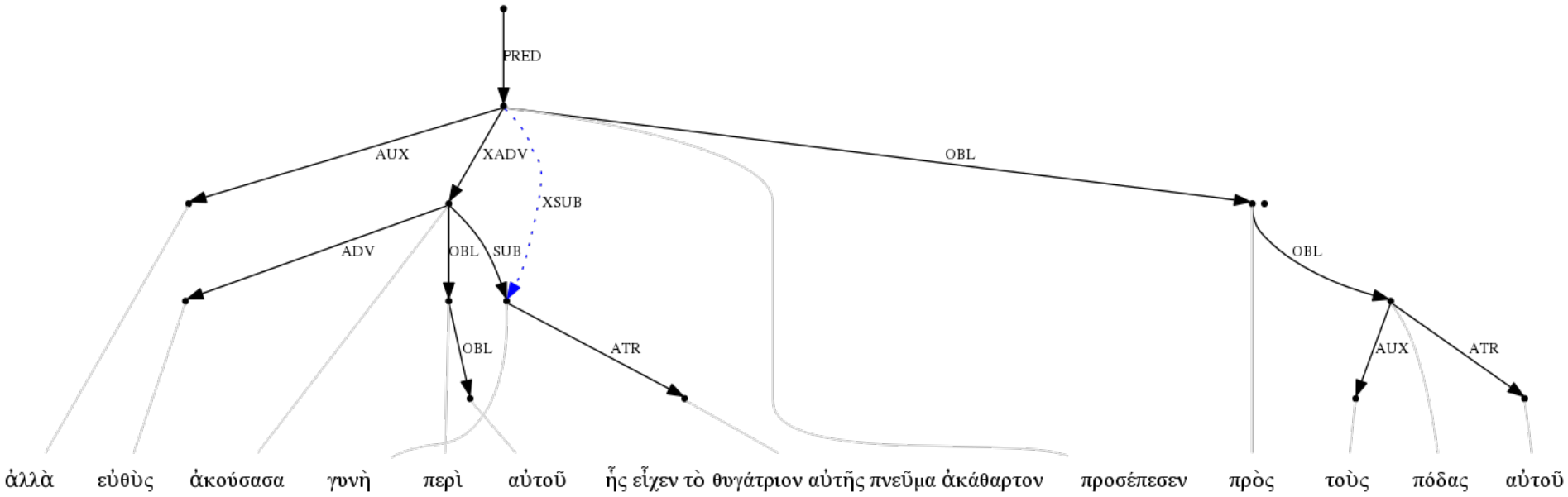
- On this analysis, it is legitimate to analyze the clause as S - V - OBL
- But we get two discontinuities which moreover are illnested

The alternative



- We get one discontinuity (the relative clause) and no illnesting

The alternative



- We get one discontinuity (the relative clause) and no illnesting
- On this analysis it is not legitimate to count the matrix clause as S - V - OBL because there is no overt S

How to choose?

- The question is better framed in terms of constituents than dependencies, i.e. as questions about properties of the participle's projection

How to choose?

- The question is better framed in terms of constituents than dependencies, i.e. as questions about properties of the participle's projection
- We can connect with known typologies of discontinuities
  - unbounded dependencies
  - extraposition
  - scrambling

How to choose?

- The question is better framed in terms of constituents than dependencies, i.e. as questions about properties of the participle's projection
- We can connect with known typologies of discontinuities
  - unbounded dependencies
  - extraposition
  - scrambling
- The presence/absence of a subject can be related to the type of the participle's projection (clausal or verb phrase)

How free is the word order?

- The question, then, is how free the word order in participle clauses is

How free is the word order?

- The question, then, is how free the word order in participle clauses is
- We know that unbounded dependencies can break up clausal categories

How free is the word order?

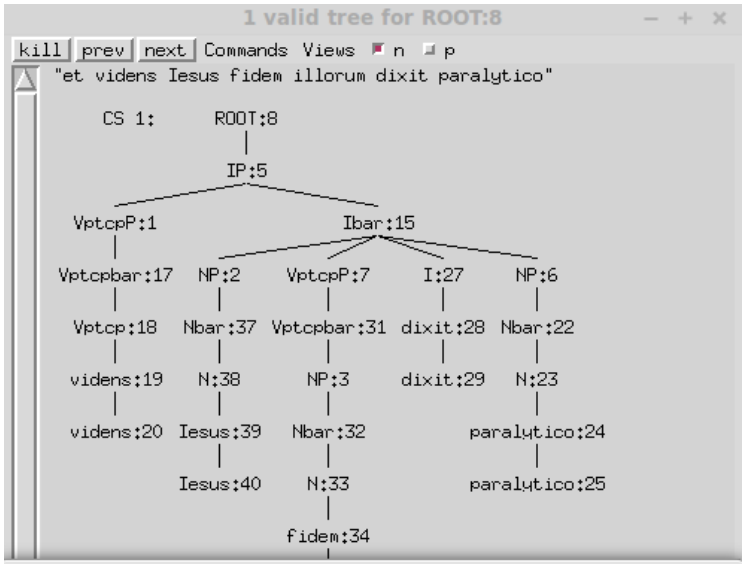
- The question, then, is how free the word order in participle clauses is
- We know that unbounded dependencies can break up clausal categories
- Can participle clauses be broken up by scrambling too?

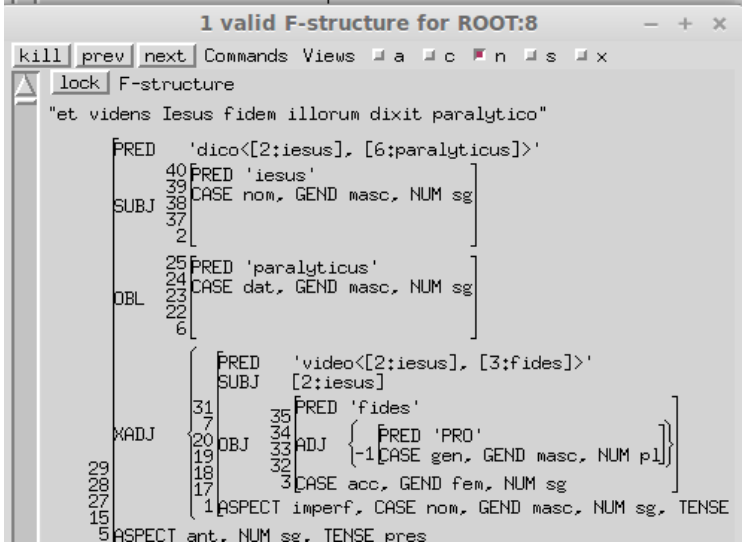
How free is the word order?

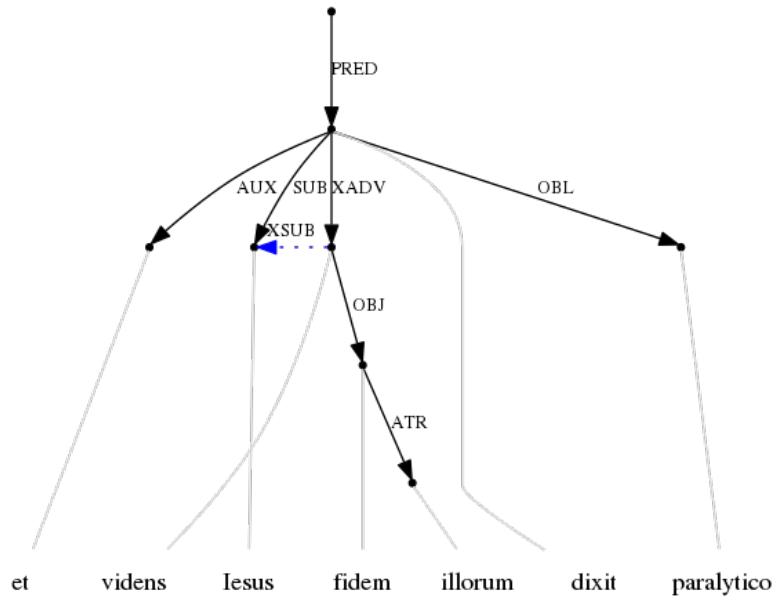
- The question, then, is how free the word order in participle clauses is
- We know that unbounded dependencies can break up clausal categories
- Can participle clauses be broken up by scrambling too?
- We therefore need to count both analyses and test their predictions
  - The internal subject hypothesis predicts only the subject can interrupt the participle clause

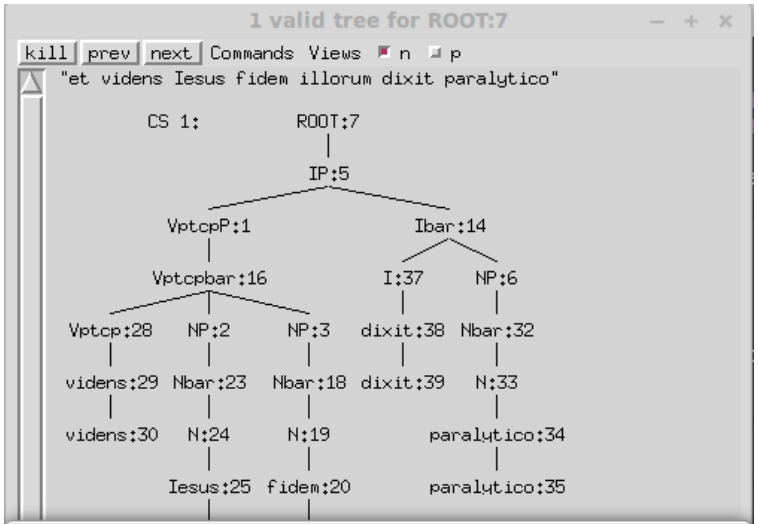
How free is the word order?

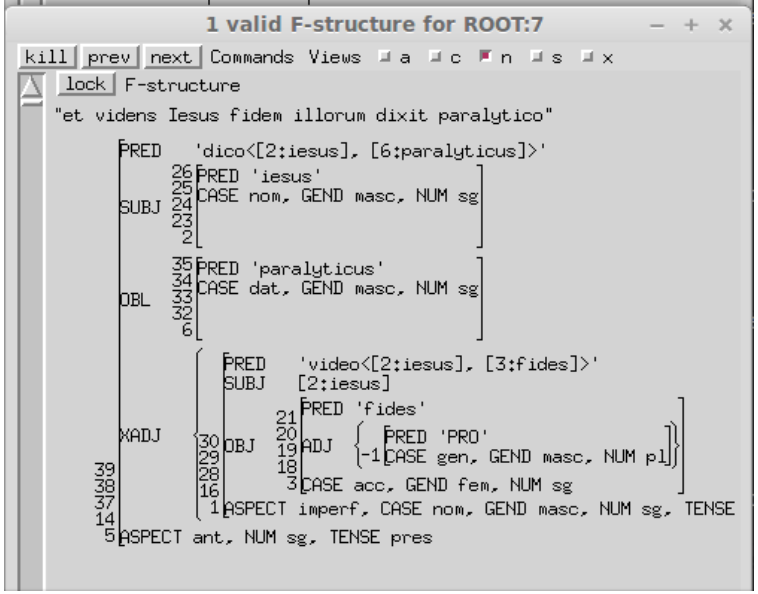
- The question, then, is how free the word order in participle clauses is
- We know that unbounded dependencies can break up clausal categories
- Can participle clauses be broken up by scrambling too?
- We therefore need to count both analyses and test their predictions
  - The internal subject hypothesis predicts only the subject can interrupt the participle clause
  - The discontinuous clause hypothesis predicts other things to intervene as well
- We generate c-structures based on both the SUB and XSUB relation

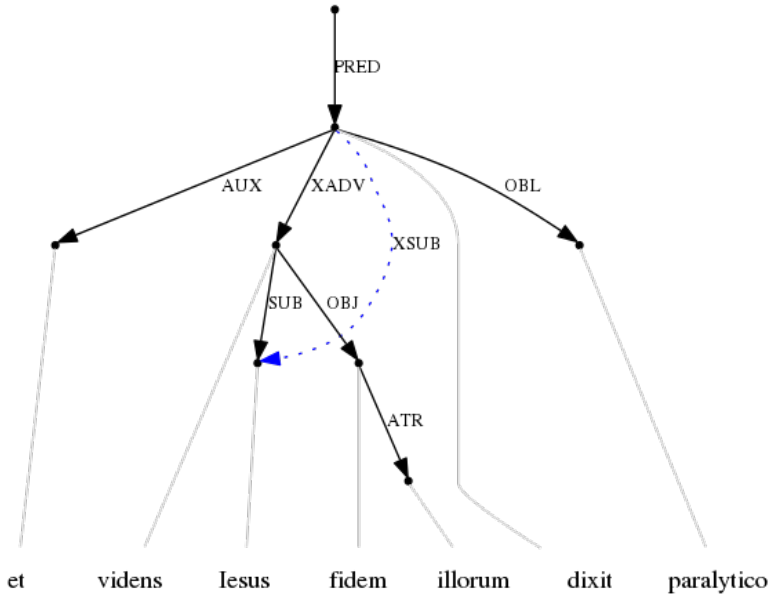

1 valid tree for ROOT:8
kill prev next Commands Views n p
"et videns Iesus fidem illorum dixit paralytico"
CS 1: ROOT:8
IP:5
VptcpP:1
Ibar:15
Vptcpbar:17
NP:2
VptcpP:7
I:27
NP:6
Vptcp:18
Nbar:37
Vptcpbar:31
dixit:28
Nbar:22
videns:19
N:38
NP:3
dixit:29
N:23
videns:20
Iesus:39
Nbar:32
paralytico:24
Iesus:40
N:33
paralytico:25
fidem:34


1 valid F-structure for ROOT:8
kill prev next Commands Views a c n s x
lock F-structure
"et videns Iesus fidem illorum dixit paralytico"
PRED 'dico<[2:iesus], [6:paralyticus]>'
SUBJ
PRED 'iesus'
CASE nom, GEND masc, NUM sg
OBL
PRED 'paralyticus'
CASE dat, GEND masc, NUM sg
XADJ
PRED 'video<[2:iesus], [3:fides]>'
SUBJ [2:iesus]
OBJ
PRED 'fides'
ADJ
PRED 'PRO'
CASE gen, GEND masc, NUM pl
CASE acc, GEND fem, NUM sg
ASPECT imperf, CASE nom, GEND masc, NUM sg, TENSE
ASPECT ant, NUM sg, TENSE pres


PRED
AUX
SUB
XADV
OBL
XSUB
OBJ
ATR
et
videns
Iesus
fidem
illorum
dixit
paralytico


1 valid tree for ROOT:7
kill prev next Commands Views n p
"et videns Iesus fidem illorum dixit paralytico"
CS 1: ROOT:7
IP:5
VptcpP:1
Ibar:14
Vptcpbar:16
I:37
NP:6
Vptcp:28
NP:2
NP:3
dixit:38
Nbar:32
videns:29
Nbar:23
Nbar:18
dixit:39
N:33
videns:30
N:24
N:19
paralytico:34
Iesus:25
fidem:20
paralytico:35


1 valid F-structure for ROOT:7
kill prev next Commands Views a c n s x
lock F-structure
"et videns Iesus fidem illorum dixit paralytico"
PRED 'dico<[2:iesus], [6:paralyticus]>'
SUBJ
PRED 'iesus'
CASE nom, GEND masc, NUM sg
OBL
PRED 'paralyticus'
CASE dat, GEND masc, NUM sg
XADJ
PRED 'video<[2:iesus], [3:fides]>'
SUBJ [2:iesus]
OBJ
PRED 'fides'
ADJ
PRED 'PRO'
CASE gen, GEND masc, NUM pl
CASE acc, GEND fem, NUM sg
ASPECT imperf, CASE nom, GEND masc, NUM sg, TENSE
ASPECT ant, NUM sg, TENSE pres


PRED
AUX
XADV
OBL
XSUB
SUB
OBJ
ATR
et
videns
Iesus
fidem
illorum
dixit
paralytico

The data

	Caesar			Gospels			Herodotus		
	op	no-op	cont	op	no-op	cont	op	no-op	cont
CP	3	0	707	7	0	1381	1	0	209

The data

	Caesar			Gospels			Herodotus		
	op	no-op	cont	op	no-op	cont	op	no-op	cont
CP	3	0	707	7	0	1381	1	0	209
IP	75	0	2656	33	0	9861	11	0	1185

The data

	Caesar			Gospels			Herodotus		
	op	no-op	cont	op	no-op	cont	op	no-op	cont
CP	3	0	707	7	0	1381	1	0	209
IP	75	0	2656	33	0	9861	11	0	1185
Infinitive	30	85	817	72	36	698	45	64	342

The data

	Caesar			Gospels			Herodotus		
	op	no-op	cont	op	no-op	cont	op	no-op	cont
CP	3	0	707	7	0	1381	1	0	209
IP	75	0	2656	33	0	9861	11	0	1185
Infinitive	30	85	817	72	36	698	45	64	342
Ptcp. (ext. subj.)	12	11	221	2	48	1258	5	16	515

The data

	Caesar			Gospels			Herodotus		
	op	no-op	cont	op	no-op	cont	op	no-op	cont
CP	3	0	707	7	0	1381	1	0	209
IP	75	0	2656	33	0	9861	11	0	1185
Infinitive	30	85	817	72	36	698	45	64	342
Ptcp. (ext. subj.)	12	11	221	2	48	1258	5	16	515
Ptpc. (int. subj.)	0	1	243	0	0	1308	3	2	531

The data

	Caesar			Gospels			Herodotus		
	op	no-op	cont	op	no-op	cont	op	no-op	cont
CP	3	0	707	7	0	1381	1	0	209
IP	75	0	2656	33	0	9861	11	0	1185
Infinitive	30	85	817	72	36	698	45	64	342
Ptcp. (ext. subj.)	12	11	221	2	48	1258	5	16	515
Ptpc. (int. subj.)	0	1	243	0	0	1308	3	2	531

- The data makes it very likely that the subject should be counted as internal to the participle phrase, so Kirk was right to exclude them from the main clause

Why does this work?

- Crucially, the answer is not directly encoded in the representation

Why does this work?

- Crucially, the answer is not directly encoded in the representation
- There is a core of "facts" (incl. multiple subjecthood) encoded as theory-neutral as possible

Why does this work?

- Crucially, the answer is not directly encoded in the representation
- There is a core of "facts" (incl. multiple subjecthood) encoded as theory-neutral as possible
- The research relies on enriching the annotation with theoretical hypotheses and either
  - the most restrictive analysis fails

Why does this work?

- Crucially, the answer is not directly encoded in the representation
- There is a core of "facts" (incl. multiple subjecthood) encoded as theory-neutral as possible
- The research relies on enriching the annotation with theoretical hypotheses and either
  - the most restrictive analysis fails
  - the most restrictive analysis holds in all cases

Why does this work?

- Crucially, the answer is not directly encoded in the representation
- There is a core of "facts" (incl. multiple subjecthood) encoded as theory-neutral as possible
- The research relies on enriching the annotation with theoretical hypotheses and either
  - the most restrictive analysis fails
  - the most restrictive analysis holds in all cases
- With further statistical analysis the second outcome may stand in for negative evidence for historical language states

Why does this work?

- Crucially, the answer is not directly encoded in the representation
- There is a core of "facts" (incl. multiple subjecthood) encoded as theory-neutral as possible
- The research relies on enriching the annotation with theoretical hypotheses and either
  - the most restrictive analysis fails
  - the most restrictive analysis holds in all cases
- With further statistical analysis the second outcome may stand in for negative evidence for historical language states
- But why should scrambling across participle clauses be disallowed?

Pullum's conjecture (Pullum, 1982, p. 8)

"no constituent of a recursive category (one that can immediately dominate itself) can scramble out of that category."

Pullum's conjecture (Pullum, 1982, p. 8)

"no constituent of a recursive category (one that can immediately dominate itself) can scramble out of that category."

- The generative power (and asymptotic complexity) of LFG comes from reentrancies: distinct nodes pieced together in the functional structure

Pullum's conjecture (Pullum, 1982, p. 8)

"no constituent of a recursive category (one that can immediately dominate itself) can scramble out of that category."

- The generative power (and asymptotic complexity) of LFG comes from reentrancies: distinct nodes pieced together in the functional structure
- An LFG grammar with reentrancies across recursive categories gives rise to unbounded reentrancies and an exponential parsing problem

Pullum's conjecture (Pullum, 1982, p. 8)

"no constituent of a recursive category (one that can immediately dominate itself) can scramble out of that category."

- The generative power (and asymptotic complexity) of LFG comes from reentrancies: distinct nodes pieced together in the functional structure
- An LFG grammar with reentrancies across recursive categories gives rise to unbounded reentrancies and an exponential parsing problem
- If on the other hand an LFG grammar puts a bound on reentrancies, parsing becomes tractable (Seki et al., 1993)

Pullum's conjecture (Pullum, 1982, p. 8)

"no constituent of a recursive category (one that can immediately dominate itself) can scramble out of that category."

- The generative power (and asymptotic complexity) of LFG comes from reentrancies: distinct nodes pieced together in the functional structure
- An LFG grammar with reentrancies across recursive categories gives rise to unbounded reentrancies and an exponential parsing problem
- If on the other hand an LFG grammar puts a bound on reentrancies, parsing becomes tractable (Seki et al., 1993)
- So one could imagine that scrambling out of a participle clause is disallowed for complexity reasons

The obvious problem: infinitives

- It is well-known that German has long-distance scrambling across (coherent) infinitive constructions

(2) dass [die Witwen]$_j$ [der Opfer]$_i$ [dem Pfarrer]$_k$ der Rat
that the widows the victims the priest the council
gedenken zu lassen versprochen hat.
to commemorate let promised have
'. . . that the council has promised the priest to let the widows commemorate the victims' (Becker et al., 1991)'

The obvious problem: infinitives

- It is well-known that German has long-distance scrambling across (coherent) infinitive constructions

(2) dass [die Witwen]$_j$ [der Opfer]$_i$ [dem Pfarrer]$_k$ der Rat
that the widows the victims the priest the council
gedenken zu lassen versprochen hat.
to commemorate let promised have
'...that the council has promised the priest to let the widows commemorate the victims' (Becker et al., 1991)'

- Certain types of clause union analyses would merge the argument lists

The obvious problem: infinitives

- It is well-known that German has long-distance scrambling across (coherent) infinitive constructions

(2) dass [die Witwen]$_j$ [der Opfer]$_i$ [dem Pfarrer]$_k$ der Rat
that the widows the victims the priest the council
gedenken zu lassen versprochen hat.
to commemorate let promised have
'. . . that the council has promised the priest to let the widows commemorate the victims' (Becker et al., 1991)'

- Certain types of clause union analyses would merge the argument lists
- Alternatively, Joshi et al. (2002) suggest that bounds on scrambling depth could be determined by the grammar

A challenge for linguistic theories: Illnestedness

- As we saw, the correct analysis of participle clauses in Ancient Greek (and Latin) removes one source of illnestedness in the language

A challenge for linguistic theories: Illnestedness

- As we saw, the correct analysis of participle clauses in Ancient Greek (and Latin) removes one source of illnestedness in the language
- In fact, there are strikingly few illnested constructions in AG and Latin, given the overall nonprojectivity in the languages

A challenge for linguistic theories: Illnestedness

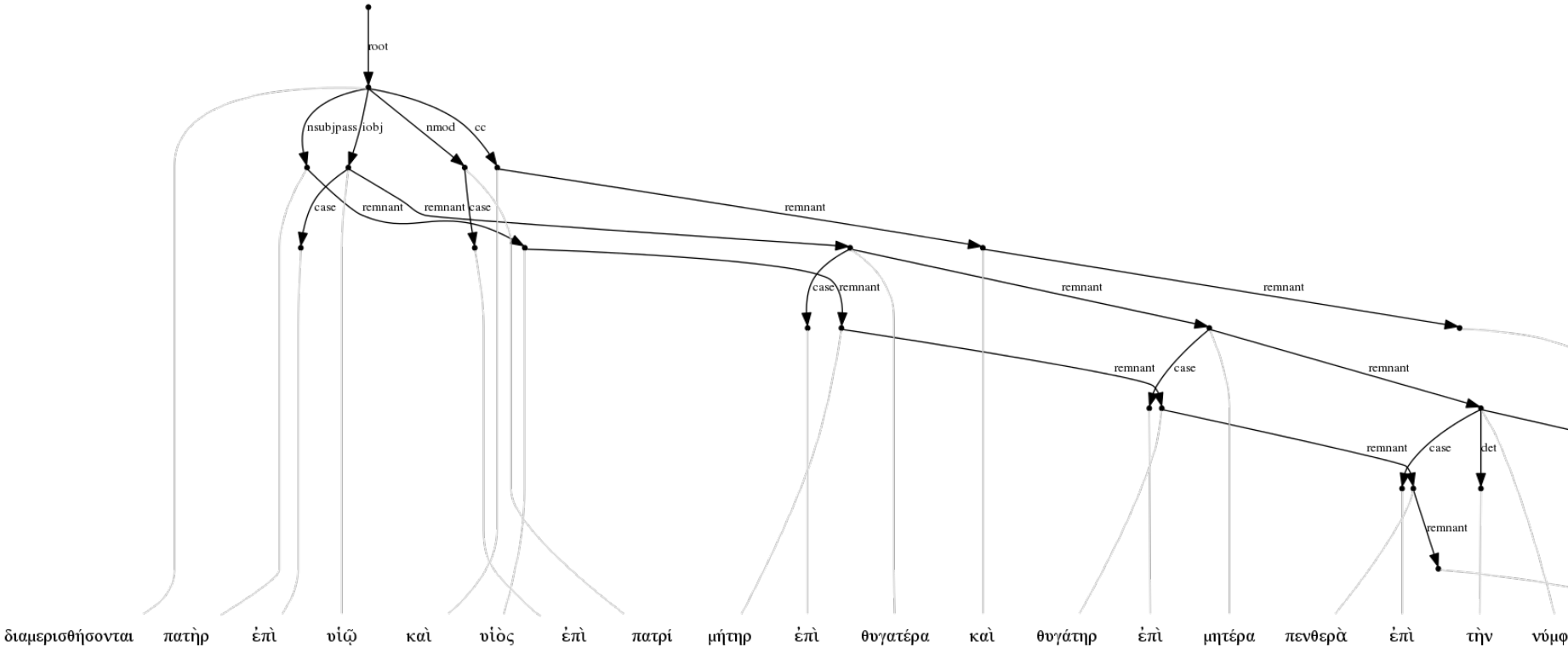
- As we saw, the correct analysis of participle clauses in Ancient Greek (and Latin) removes one source of illnestedness in the language
- In fact, there are strikingly few illnested constructions in AG and Latin, given the overall nonprojectivity in the languages
- They also share some characteristics among themselves and with corresponding examples in German, where Mambrini & Passarotti (2013) report illnestedness at 1.06%

Illnestedness in UD

	illnested	wellnested
Ancient_Greek	0.019	0.981
Ancient_Greek-PROIEL	0.023	0.977
Arabic	0.001	0.999
Basque	0.016	0.984
Bulgarian	0.000	1.000
Croatian	0.007	0.993
Czech	0.001	0.999
Danish	0.009	0.991
Dutch	0.001	0.999
English	0.002	0.998
Estonian	0.000	1.000
Finnish	0.006	0.994
Finnish-FTB	0.006	0.994
French	0.001	0.999
German	0.001	0.999
Gothic	0.019	0.981
Greek	0.005	0.995
Hebrew	0.000	1.000
Hindi	0.002	0.998

	illnested	wellnested
Hungarian	0.022	0.978
Indonesian	0.001	0.999
Irish	0.000	1.000
Italian	0.004	0.996
Japanese-KTC	0.000	1.000
Latin	0.042	0.958
Latin-ITT	0.003	0.997
Latin-PROIEL	0.015	0.985
Norwegian	0.001	0.999
Old_Church_Slavonic	0.018	0.982
Persian	0.000	1.000
Polish	0.000	1.000
Portuguese	0.001	0.999
Romanian	0.019	0.981
Slovenian	0.003	0.997
Spanish	0.000	1.000
Swedish	0.000	1.000
Tamil	0.000	1.000

Deep illnestedness in Greek

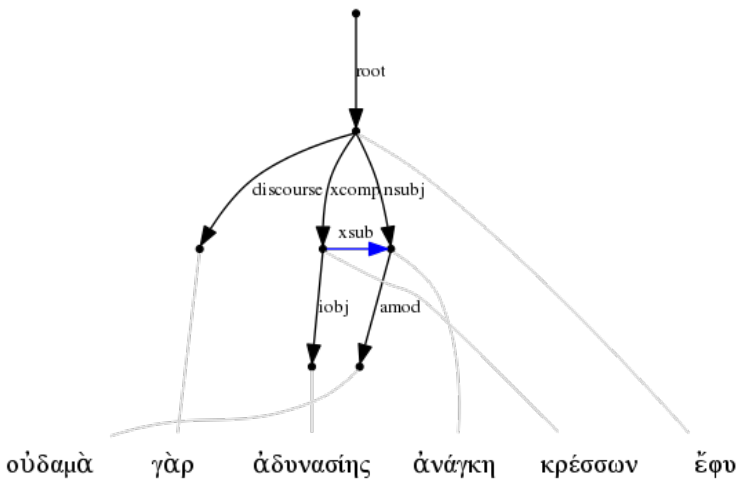


"Real" illnesting (without remnants and punctuation)

	illnested	wellnested
Ancient_Greek	0.014	0.986
Ancient_Greek-PROIEL	0.003	0.997
Arabic	0.000	1.000
Basque	0.001	0.999
Bulgarian	0.000	1.000
Croatian	0.000	1.000
Czech	0.001	0.999
Danish	0.001	0.999
Dutch	0.001	0.999
English	0.000	1.000
Estonian	0.000	1.000
Finnish	0.000	1.000
Finnish-FTB	0.000	1.000
French	0.000	1.000
German	0.000	1.000
Gothic	0.002	0.998
Greek	0.000	1.000
Hebrew	0.000	1.000
Hindi	0.001	0.999

	illnested	wellnested
Hungarian	0.000	1.000
Indonesian	0.000	1.000
Irish	0.000	1.000
Italian	0.000	1.000
Japanese-KTC	0.000	1.000
Latin	0.037	0.963
Latin-ITT	0.002	0.998
Latin-PROIEL	0.002	0.998
Norwegian	0.001	0.999
Old_Church_Slavonic	0.002	0.998
Persian	0.000	1.000
Polish	0.000	1.000
Portuguese	0.000	1.000
Romanian	0.000	1.000
Slovenian	0.000	1.000
Spanish	0.000	1.000
Swedish	0.000	1.000
Tamil	0.000	1.000

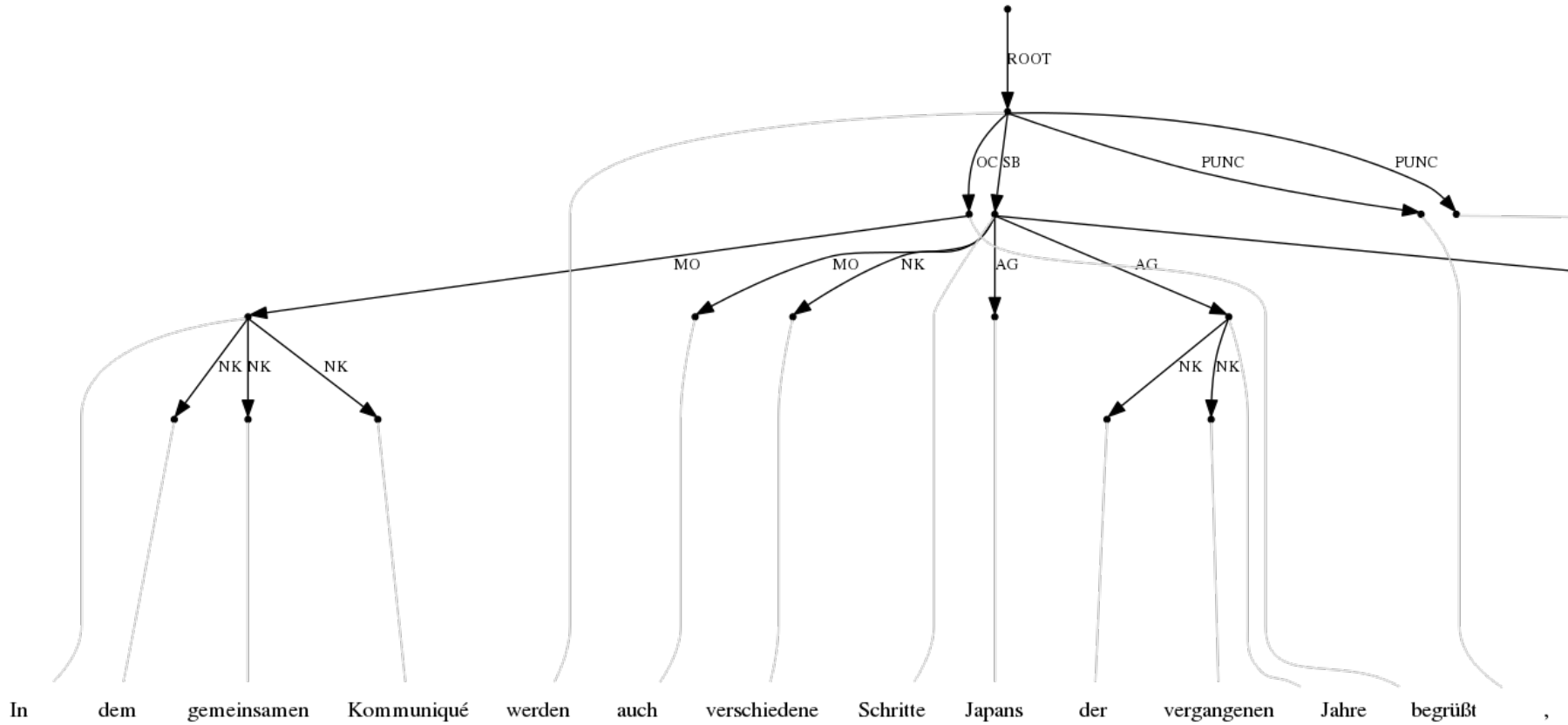
Greek illnestedness



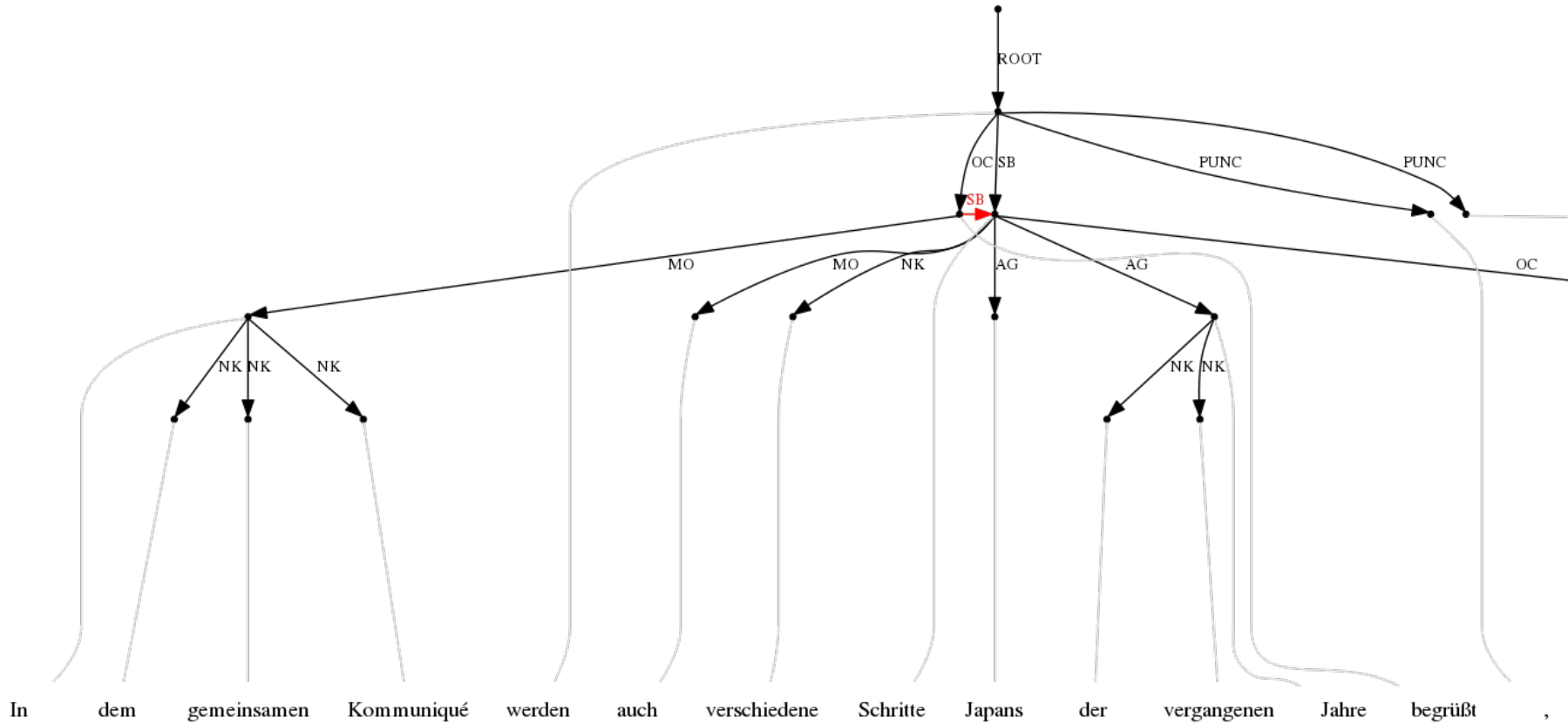
- The subject and the predicative are illnested

(3) oudama gar adunasiēs anagkē kressōn
no.NOM for lack of ability.GEN necessity.NOM stronger.NOM
efu
grow
'For no necessity can grow stronger than lack of ability.'

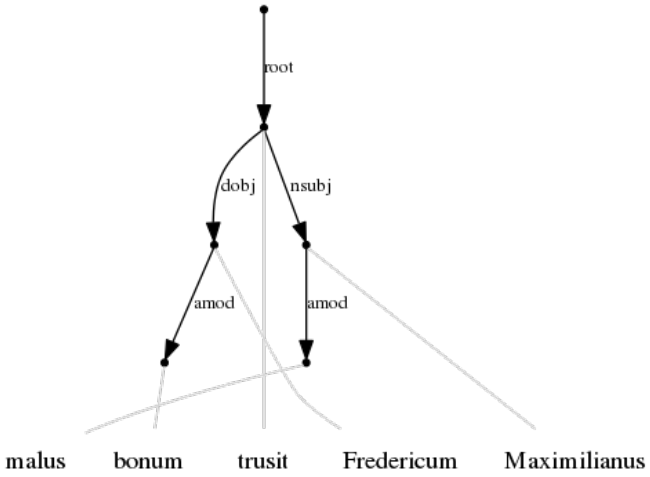
Illnestedness in Tiger

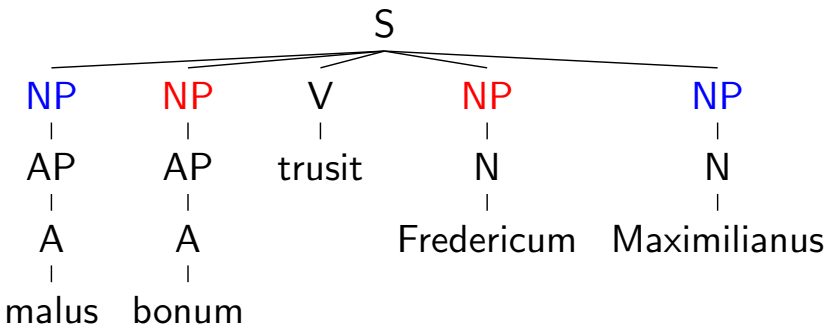


Illnestedness in Tiger

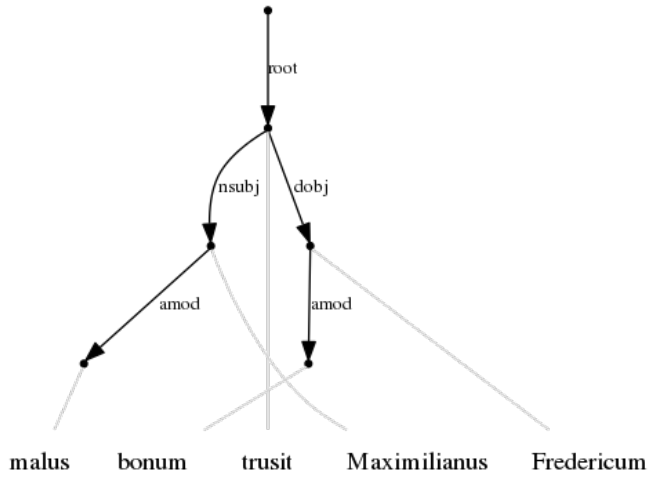


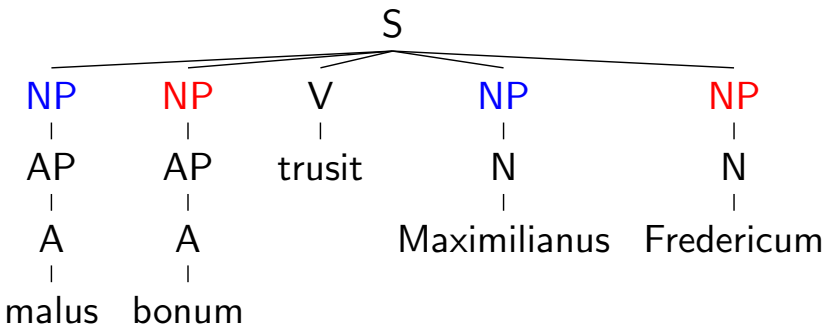
Illnestedness





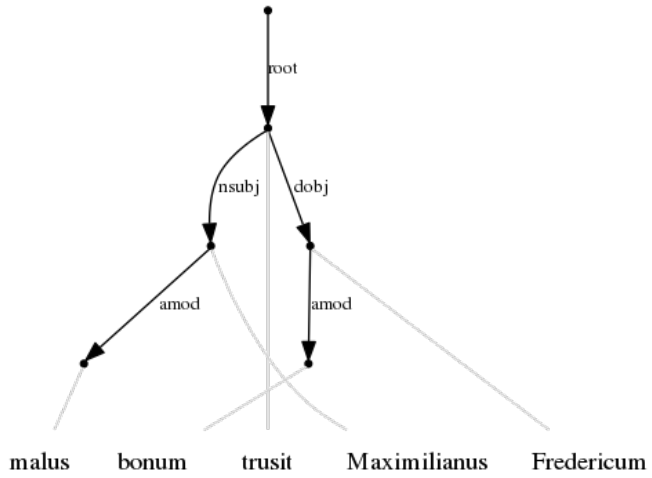
Illnestedness

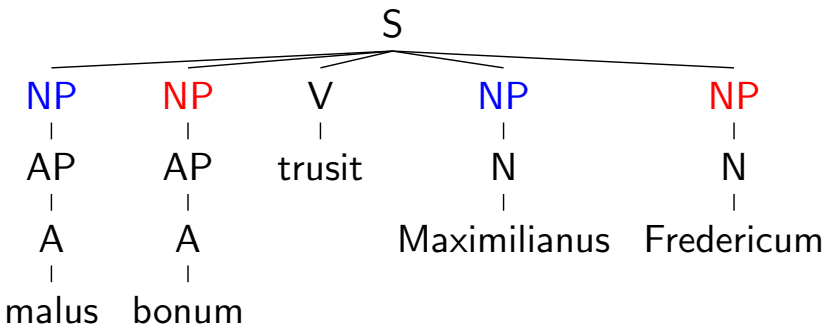




- The illnested structure does not require more reentrancies than the wellnested one

Illnestedness





- The illnested structure does not require more reentrancies than the wellnested one
- And yet illnested structures are practically nonextant in the UD treebanks

Conclusions

- Designing treebanks for theoretical linguistics research requires careful attention to annotation schemes
  - The knowns go in the annotation

Conclusions

- Designing treebanks for theoretical linguistics research requires careful attention to annotation schemes
  - The knowns go in the annotation
  - The known unknowns go in post-annotation experimental enrichments

Conclusions

- Designing treebanks for theoretical linguistics research requires careful attention to annotation schemes
  - The knowns go in the annotation
  - The known unknowns go in post-annotation experimental enrichments
  - Careful with the unknown unknowns!
- Linguistically motivated structures can reflect complexity in a way that corresponds to frequency (e.g. gap degrees and depths)

Conclusions

- Designing treebanks for theoretical linguistics research requires careful attention to annotation schemes
  - The knowns go in the annotation
  - The known unknowns go in post-annotation experimental enrichments
  - Careful with the unknown unknowns!
- Linguistically motivated structures can reflect complexity in a way that corresponds to frequency (e.g. gap degrees and depths)
- Motivated structures can provide evidence for non-standard analyses which avoid exponential blowup (e.g. reentrancies between recursive categories)

Conclusions

- Designing treebanks for theoretical linguistics research requires careful attention to annotation schemes
  - The knowns go in the annotation
  - The known unknowns go in post-annotation experimental enrichments
  - Careful with the unknown unknowns!
- Linguistically motivated structures can reflect complexity in a way that corresponds to frequency (e.g. gap degrees and depths)
- Motivated structures can provide evidence for non-standard analyses which avoid exponential blowup (e.g. reentrancies between recursive categories)
- But sometimes there is no connection between treebank data and theoretical structures, so there is more to do

Becker, Tilman, Aravind K. Joshi & Owen Rambow. 1991. Long-distance scrambling and tree adjoining grammars. In *Proceedings of the fifth conference on european chapter of the association for computational linguistics* EACL '91, 21–26. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/977180.977185. http://dx.doi.org/10.3115/977180.977185.

Bröker, Norbert, Udo Hahn & Susanne Schacht. 1994. Concurrent lexicalized dependency parsing: The parsetalk model. In *Proceedings of the 15th conference on computational linguistics-volume 1*, 379–385.

Davison, M. E. 1989. New testament greek word order. *Literary and Linguistic Computing* 4. 19–28.

Gómez-rodríguez, Carlos, David Weir & John Carrol. 2009. Parsing midly non-projective dependency structures. In *Proceedings of the 12th conference of the european chapter of the acl*, 291–299.

Haug, Dag Trygve Truslew. 2012. From dependency structures to LFG representations. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG12 conference*, 271–291. CSLI Publications.

Kirk, Allison. 2012. *Word order and information structure in new testament greek*: Universiteit Leiden dissertation.

Kuhlmann, Marco. 2010. Mildly non-projective dependency grammar. *Computational Linguistics* 39. 355–387.

Kuhlmann, Marco & Joakim Nivre. 2006. Mildly non-projective dependency structures. In *Proceedings of the coling/acl on main conference poster sessions* COLING-ACL '06, 507–514. Stroudsburg, PA, USA: Association for Computational Linguistics. `http://dl.acm.org/citation.cfm?id=1273073.1273139`.

Mambrini, Francesco & Marco Passarotti. 2013. Non-projectivity in the ancient greek dependency treebank. In *Proceedings of the second international conference on dependency linguistics*, 177–186. Prague.

Pullum, Geoffrey. 1982. Free word order and phrase structure rules. In James Pustejovsky & Peter Sells (eds.), *Proceedings of the twelfth annual meeting of the North Eastern Linguistic Society*, 209–220. Amherst, MA: Graduate Linguistics Students Association.

Rife, J. Merle. 1933. The mechanics of translation greek. *Journal of Biblical Literature* 52. 244–252.

Seki, Hiroyuki, Ryuichi Nakanishi, Yuichi Kaji, Sachiko Ando & Tadao Kasami. 1993. Parallel multiple context-free grammars, finite-state translation systems, and polynomial-time recognizable subclasses of lexical-functional grammars. In *Proceedings of the 31st annual meeting on association for computational linguistics* ACL '93, 130–139. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/981574.981592. http://dx.doi.org/10.3115/981574.981592.

Zaenen, Annie. 2006. Mark-up barking up the wrong tree. *Comput. Linguist.* 32(4). 577–580. doi:10.1162/coli.2006.32.4.577. http://dx.doi.org/10.1162/coli.2006.32.4.577.