digital Grammars
Language technology to rely on.

Concept Alignment for Compositional Translation

Aarne Ranta
Department of Computer Science and Engineering, Chalmers & University of Gothenburg
and
Digital Grammars AB

Logic and Algorithms in Computational Linguistics (LACompLing2018)

Stockholm 30 August 2018

Earlier version: Gothenburg-Stockholm Workshop on Proof Theory, Model Theory, and Probability in Natural Language, Gothenburg 7 February 2018

Plan

Compositional translation

Concept alignment: the problem

- illustrated by GDPR (General Data Protection Regulation), of the EU

What is not compositional?

Concept alignment: towards a solution using neural UD parsing (*new*)

Compositional translation

Compositional translation in 1961: Curry

SOME LOGICAL ASPECTS OF GRAMMATICAL STRUCTURE[1]

BY

HASKELL B. CURRY[2]

pressions. This gives us two levels of grammar, the study of grammatical structure in itself, and a second level which has much the same relation to the first that morphophonemics does to morphology. In order to have terms for immediate use I shall call these two levels *tectogrammatics* and *phenogrammatics* respectively; no doubt someone will propose better terms later.

> It is to be expected that grammatical structure will vary less from language to language than does the phenogrammatics. Different languages use entirely different devices for indicating grammatical phrase composition. In some languages word order is important, as it often is in English; but in Latin the three words in
>
> *Puer puellam amat*
>
> can be arranged in any of the six possible orders without changing the structure. It is not uncommon in such languages to have functors consisting of detached parts, or of parts widely separated from the arguments.

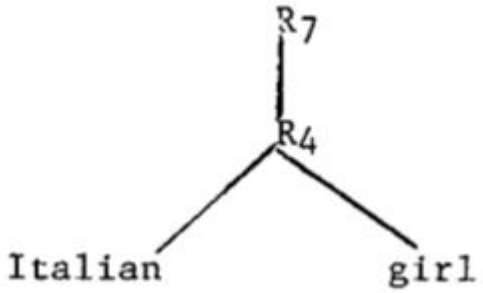
Compositional translation in 1983: Landsbergen

COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

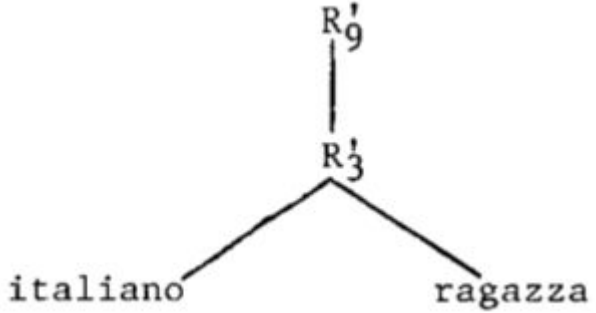
MACHINE TRANSLATION BASED ON LOGICALLY ISOMORPHIC MONTAGUE GRAMMARS

Jan Landsbergen

Philips Research Laboratories
Eindhoven - The Netherlands



"the Italian girl"



"la ragazza italiana"

Compositional translation in 1998-2018: GF

```
abstract Ex = {
  cat
    A ;
    N ;

  fun
    italian : A ;



    girl : N ;

    Mod : A -> N -> N ;




}
```

Compositional translation in 1998-2018: GF

```
abstract Ex = {
  cat
    A ;
    N ;

  fun
    italian : A ;



    girl : N ;

    Mod : A -> N -> N ;




}
```

```
concrete ExEng of Ex = {
  lincat
    A = Str ;
    N = Str ;

  lin
    italian = “Italian” ;



    girl = “girl” ;

    Mod a n = a ++ n ;




}
```

Compositional translation in 1998-2018: GF

```
abstract Ex = {
  cat
    A ;
    N ;

  fun
    italian : A ;



    girl : N ;

    Mod : A -> N -> N ;




}
```

```
concrete ExEng of Ex = {
  lincat
    A = Str ;
    N = Str ;

  lin
    italian = “Italian” ;



    girl = “girl” ;

    Mod a n = a ++ n ;




}
```

```
concrete ExIta of Ex = {
  lincat
    A = Gender => Str ;
    N = {s : Str ;
         g : Gender } ;
  lin
    italian = table {
      Masc => “italiano” ;
      Fem  => “italiana”
      } ;
    girl = {s = “ragazza” ;
            g = Fem} ;
    Mod a n = {
      s = n.s ++ a.s ! n.g ;
      g = n.g
      }
  param Gender = Masc | Fem ;
}
```

Compositional translation, formally

Abstract syntax

- category

C

- function

$F : C_1 \text{ -> } \ldots \text{ -> } C_n \text{ -> } C$

- tree

$F\ t_1 \ldots t_n$

Concrete syntax *L1*

- linearization type

C^o

- linearization function

$F^o : C_1^o \text{ -> } \ldots \text{ -> } C_n^o \text{ -> } C^o$

- linearization

$F^o\ t_1^o \ldots t_n^o$

Concrete syntax *L2*

- linearization type

C^*

- linearization function

$F^* : C_1^* \text{ -> } \ldots \text{ -> } C_n^* \text{ -> } C^*$

- linearization

$F^*\ t_1^* \ldots t_n^*$

Compositional translation, intuitively

Abstract syntax functions are **concepts**, i.e. the components of meaning.

Concrete syntax tells how each concept is expressed in each language.

To translate:

1. Analyse how the source expression is built from concepts.
2. Render the resulting complex meaning by expressing each concept in the target language.

Different kinds of concepts

A concept can be

- “atomic”, i.e. a zero-place function
- “construction”, i.e. a function that takes arguments

A concrete expression can be

- a word
- a **lemgram,** i.e. inflection table + parameters such as gender
- a **multiword,** i.e. several words or a lemgram of several words
- **discontinuous**, i.e. several words or lemgrams separated by words belonging to other concepts
- **a construction**, i.e. a function that combines lemgrams and may add words in between

How to specify a concept

Monolingual lexicon (e.g. WordNet):

sense = lemma + discriminator + part of speech

fly_1_N “two-winged insect”

fly_2_N “opening in trousers”

fly_V “travel through the air”

fly_V2 “cause to fly”

Fine-grained sense distinctions in an ontological hierarchy.

How to specify a concept

Bilingual lexicon:

sense = lemma + lemma + part of speech

fly_fliege_N	“two-winged insect”
fly_latz_N	“opening in trousers”
fly_fliegen_V	“travel through the air”
fly_fliegen_V2	“cause to fly”

Fine enough sense distinctions to express meaning preservation in translation.

Compositionality of semantics?

Language comparison is an excellent source of sense distinctions

The goal of concept alignment is meaning-preserving translation

→ concepts found by alignment are meaning-bearing units

Caveat: they may still be ambiguous.

Concept alignment: the problem

The problem

From **parallel texts**, find out what parts correspond to each other…

… so that an **abstract syntax function** can be built for these parts…

… to enable **compositional translation**

Case study: GDPR

General Data Protection Regulation, Official Journal of the EU

24 official EU languages

~80 pages in each language

60-80k words in each language

2500-3000 unique lemmas in each language

To enter in force 25 May 2018

Our task: identify concepts and how they are expressed in 5 languages (English, German, French, Italian, Spanish); commercial project at Digital Grammars AB

With

Georg Philip Krog	*international law*
Christina Unger	*abstract syntax, German*
Jordi Saludes	*Spanish*
Sara Negri	*Italian*
Daniel von Plato	*Italian*
Grégoire Détrez	*French*
Markus Forsberg	*corpus analysis*
Koen Lindström Claessen	*word alignment*
Thomas Hallgren	*visual effects*
John Camilleri	*all aspects of lexicon and supporting software*

(36) The main establishment of a controller in the Union should be the place of its central administration in the Union, unless the decisions on the purposes and means of the processing of personal data are taken in another establishment of the controller in the Union, in which case that other establishment should be considered to be the main establishment. The main establishment of a controller in the Union should be determined according to objective criteria and should imply the effective and real exercise of management activities determining the main decisions as to the purposes and means of processing through stable arrangements. That criterion should not depend on whether the processing of personal data is carried out at that location. The presence and use of technical means and technologies for processing personal data or processing activities do not, in themselves, constitute a main establishment and are therefore not determining criteria for a main establishment. The main establishment of the processor should be the place of its central administration in the Union or, if it has no central administration in the Union, the place where the main processing activities take place in the Union. In cases involving both the controller and the processor, the competent lead supervisory authority should remain the supervisory authority of the Member State where the controller has its main establishment, but the supervisory authority of the processor should be considered to be a supervisory authority concerned and that supervisory authority should participate in the cooperation procedure provided for by this Regulation. In any case, the supervisory authorities of the Me
authorities
main estab
where the

(36) Die Hauptniederlassung des Verantwortlichen in der Union sollte der Ort seiner Hauptverwaltung in der Union sein, es sei denn, dass Entscheidungen über die Zwecke und Mittel der Verarbeitung personenbezogener Daten in einer anderen Niederlassung des Verantwortlichen in der Union getroffen werden; in diesem Fall sollte die letztgenannte als Hauptniederlassung gelten. Zur Bestimmung der Hauptniederlassung eines Verantwortlichen in der Union sollten objektive Kriterien herangezogen werden; ein Kriterium sollte dabei die effektive und tatsächliche Ausübung von Managementtätigkeiten durch eine feste Einrichtung sein, in deren Rahmen die Grundsatzentscheidungen zur Festlegung der Zwecke und Mittel der Verarbeitung getroffen werden. Dabei sollte nicht ausschlaggebend sein, ob die Verarbeitung der personenbezogenen Daten tatsächlich an diesem Ort ausgeführt wird. Das Vorhandensein und die Verwendung technischer Mittel und Verfahren zur Verarbeitung personenbezogener Daten oder Verarbeitungstätigkeiten begründen an sich noch keine Hauptniederlassung und sind daher kein ausschlaggebender Faktor für das Bestehen einer Hauptniederlassung. Die Hauptniederlassung des Auftragsverarbeiters sollte der Ort sein, an dem der Auftragsverarbeiter seine Hauptverwaltung in der Union hat, oder — wenn er keine Hauptverwaltung in der Union hat — der Ort, an dem die wesentlichen Verarbeitungstätigkeiten in der Union stattfinden. Sind sowohl der Verantwortliche als auch der Auftragsverarbeiter betroffen, so sollte die Aufsichtsbehörde des Mitgliedstaats, in dem der Verantwortliche seine Hauptniederlassung hat, die zuständige federführende Aufsichtsbehörde bleiben, doch sollte die Aufsichtsbehörde des Auftragsverarbeiters als betroffene Aufsichtsbehörde betrachtet werden und diese Aufsichtsbehörde sollte sich an dem in dieser Verordnung vorgesehenen Verfahren der Zusammenarbeit beteiligen. Auf jeden Fall sollten die Aufsichtsbehörden des Mitgliedstaats oder der Mitgliedstaaten, in dem bzw. denen der Auftragsverarbeiter eine oder mehrere Niederlassungen hat, nicht als betroffene Aufsichtsbehörden betrachtet werden, wenn sich der Beschlussentwurf nur auf den Verantwortlichen bezieht. Wird die Verarbeitung durch eine Unternehmensgruppe vorgenommen, so sollte die Hauptniederlassung des herrschenden Unternehmens als Hauptniederlassung der Unternehmensgruppe gelten, es sei denn, die Zwecke und Mittel der Verarbeitung werden von einem anderen Unternehmen festgelegt.

(36) The main establishment of a controller in the Union should be the place of its central administration in the Union, unless the decisions on the purposes and means of the processing of personal data are taken in another establishment of the controller in the Union, in which case that

(36) Lo stabilimento principale di un titolare del trattamento nell'Unione dovrebbe essere il luogo in cui ha sede la sua amministrazione centrale nell'Unione, a meno che le decisioni sulle finalità e i mezzi del trattamento di dati personali siano adottate in un altro stabilimento del titolare del trattamento nell'Unione, nel qual caso tale altro stabilimento dovrebbe essere considerato lo stabilimento principale. Lo stabilimento principale di un titolare del trattamento nell'Unione dovrebbe essere determinato in base a criteri obiettivi e implicare l'effettivo e reale svolgimento di attività di gestione finalizzate alle principali decisioni sulle finalità e sui mezzi del trattamento nel quadro di un'organizzazione stabile. Tale criterio non dovrebbe dipendere dal fatto che i dati personali siano trattati in quella sede. La presenza o l'uso di mezzi tecnici e

(36) El establecimiento principal de un responsable del tratamiento en la Unión debe ser el lugar de su administración central en la Unión, salvo que las decisiones relativas a los fines y medios del tratamiento de los datos personales se tomen en otro establecimiento del responsable en la Unión, en cuyo caso, ese otro establecimiento debe considerarse el establecimiento principal. El establecimiento principal de un responsable en la Unión debe determinarse en función de criterios objetivos y debe implicar el ejercicio efectivo y real de actividades de gestión que determinen las principales decisiones en cuanto a los fines y medios del tratamiento a través de modalidades estables. Dicho criterio no debe depender de si el tratamiento de los datos personales se realiza en dicho lugar. La presencia y utilización de medios técnicos y tecnologías para el tratamiento de datos personales o las actividades de tratamiento no constituyen, en sí mismas, establecimiento principal y no son, por lo tanto, criterios determinantes de un establecimiento principal. El establecimiento principal del encargado del tratamiento debe ser el lugar de su administración central en la Unión o, si careciese de administración central en la Unión, el lugar en el que se llevan a cabo las principales actividades de tratamiento en la Unión. En los casos que impliquen tanto al responsable como al encargado, la autoridad de control principal competente debe seguir siendo la autoridad de control del Estado miembro en el que el responsable tenga su establecimiento principal, pero la autoridad de control del encargado debe considerarse autoridad de control interesada y participar en el procedimiento de cooperación establecido en el presente Reglamento. En cualquier caso, las autoridades de control del Estado miembro o los Estados miembros en los que el encargado tenga uno o varios establecimientos no deben considerarse autoridades de control interesadas cuando el proyecto de decisión afecte únicamente al responsable. Cuando el tratamiento lo realice un grupo empresarial, el establecimiento principal de la empresa que ejerce el control debe considerarse el establecimiento principal del grupo empresarial, excepto cuando los fines y medios del tratamiento los determine otra empresa.

mehrere Niederlassungen hat, nicht als betroffene Aufsichtsbehörden betrachtet werden, wenn sich der Beschlussentwurf nur auf den Verantwortlichen bezieht. Wird die Verarbeitung durch eine Unternehmensgruppe vorgenommen, so sollte die Hauptniederlassung des herrschenden Unternehmens als Hauptniederlassung der Unternehmensgruppe gelten, es sei denn, die Zwecke und Mittel der Verarbeitung werden von einem anderen Unternehmen festgelegt.

The first sentence

Eng: The protection of natural persons in relation to the processing of personal data is a fundamental right.

Ger: Der Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten ist ein Grundrecht.

Fre: La protection des personnes physiques à l'égard du traitement des données à caractère personnel est un droit fondamental.

Ita: La protezione delle persone fisiche con riguardo al trattamento dei dati di carattere personale è un diritto fondamentale .

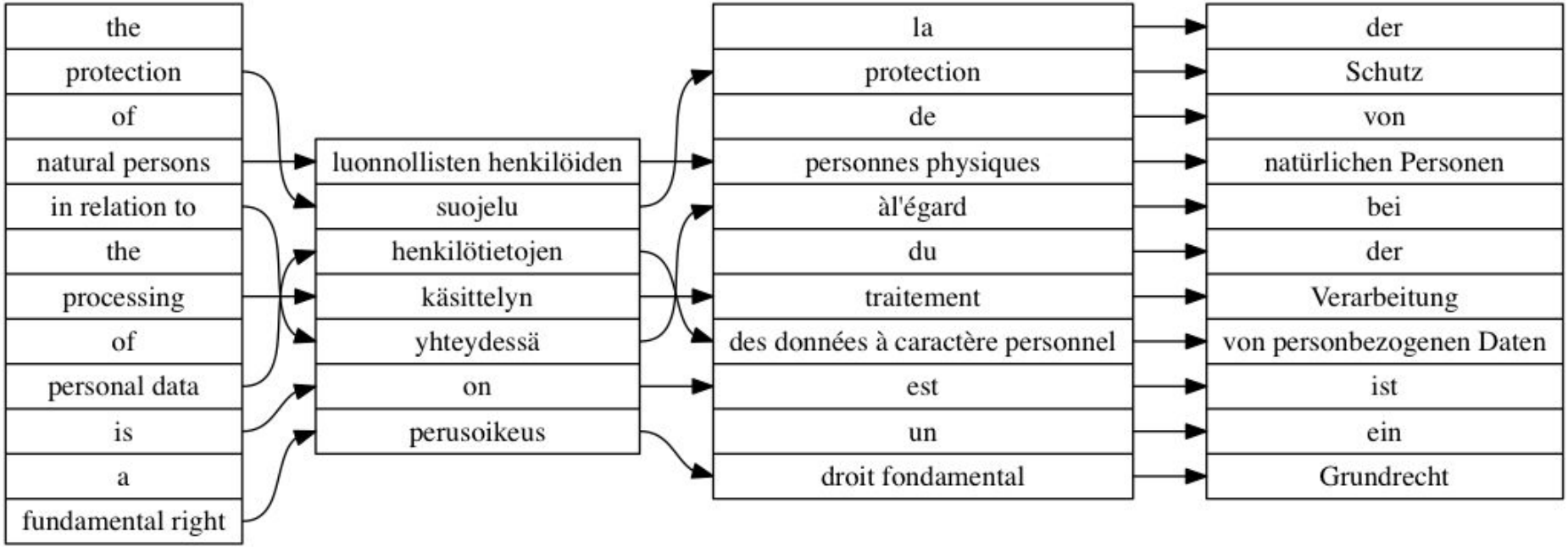
Spa: La protección de las personas físicas en relación con el tratamiento de datos personales es un derecho fundamental.

Fin: Luonnollisten henkilöiden suojelu henkilötietojen käsittelyn yhteydessä on perusoikeus.

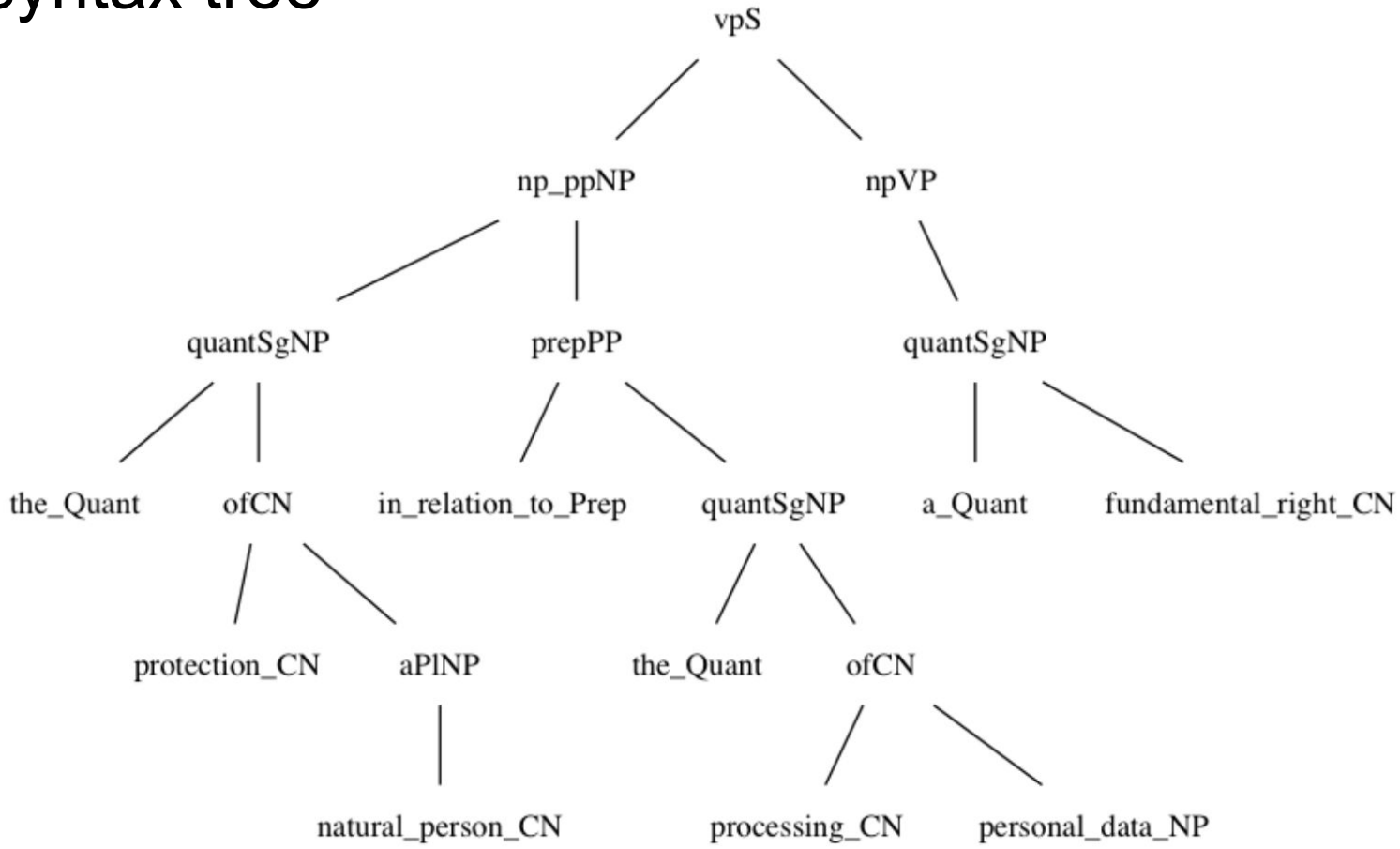
Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

http://eur-lex.europa.eu/legal-content/FI/TXT/HTML/?uri=CELEX:32016R0679&from=EN

Concept alignments in the first sentence



Abstract syntax tree



Easy case: compound to multiword

encourage the establishment of data protection certification mechanisms pursuant to Article 42(1), and approve the criteria of certification pursuant to Article 42(5)

die Einführung von Datenschutzzertifizierungsmechanismen nach Artikel 42 Absatz 1 anregen und Zertifizierungskriterien nach Artikel 42 Absatz 5 billigen

encourage la mise en place de mécanismes de certification en matière de protection des données en application de l'article 42, paragraphe 1, et approuve les critères de certification en application de l'article 42, paragraphe 5

German was a good starting point for identifying these!

Easy case: compound to multiword

data protection certification mechanisms

Datenschutzzertifizierungsmechanismen

mécanismes de certification en matière de protection des données

A trickier example

Tätä asetusta olisi sovellettava myös unionin alueella olevien rekisteröityjen henkilötietojen käsittelyyn, jos sitä suorittava rekisterinpitäjä tai henkilötietojen käsittelijä ei ole sijoittautunut unioniin ja jos käsittely liittyy näiden rekisteröityjen käyttäytymisen seurantaan niiltä osin kuin käyttäytyminen tapahtuu unionissa.

The processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union should also be subject to this Regulation when it is related to the monitoring of the behaviour of such data subjects in so far as their behaviour takes place within the Union.

Die Verarbeitung personenbezogener Daten von betroffenen Personen, die sich in der Union befinden, durch einen nicht in der Union niedergelassenen Verantwortlichen oder Auftragsverarbeiter sollte auch dann dieser Verordnung unterliegen, wenn sie dazu dient, das Verhalten dieser betroffenen Personen zu beobachten, soweit ihr Verhalten in der Union erfolgt.

What is the common structure?

A trickier example

Tätä asetusta olisi sovellettava myös unionin alueella olevien rekisteröityjen henkilötietojen käsittelyyn, jos sitä suorittava rekisterinpitäjä tai henkilötietojen käsittelijä ei ole sijoittautunut unioniin ja jos käsittely liittyy näiden rekisteröityjen käyttäytymisen seurantaan niiltä osin kuin käyttäytyminen tapahtuu unionissa.

The processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union should also be subject to this Regulation when it is related to the monitoring of the behaviour of such data subjects in so far as their behaviour takes place within the Union.

Die Verarbeitung personenbezogener Daten von betroffenen Personen, die sich in der Union befinden, durch einen nicht in der Union niedergelassenen Verantwortlichen oder Auftragsverarbeiter sollte auch dann dieser Verordnung unterliegen, wenn sie dazu dient, das Verhalten dieser betroffenen Personen zu beobachten, soweit ihr Verhalten in der Union erfolgt.

```
fun be_subject_to__unterliegen_NP_NP_Cl : NP -> NP -> Cl
```

A trickier example

Tätä asetusta olisi sovellettava myös unionin alueella olevien rekisteröityjen henkilötietojen käsittelyyn, jos sitä suorittava rekisterinpitäjä tai henkilötietojen käsittelijä ei ole sijoittautunut unioniin ja jos käsittely liittyy näiden rekisteröityjen käyttäytymisen seurantaan niiltä osin kuin käyttäytyminen tapahtuu unionissa.

Fin: `mkCl y (passiveVPSlash (mkVPSlash (mkV3 soveltaa_V part illat) x)`

The processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union should also be subject to this Regulation when it is related to the monitoring of the behaviour of such data subjects in so far as their behaviour takes place within the Union.

Eng: `mkCl x (mkVP (mkAP (mkA2 subject_A to_Prep) y))`

Die Verarbeitung personenbezogener Daten von betroffenen Personen, die sich in der Union befinden, durch einen nicht in der Union niedergelassenen Verantwortlichen oder Auftragsverarbeiter sollte auch dann dieser Verordnung unterliegen, wenn sie dazu dient, das Verhalten dieser betroffenen Personen zu beobachten, soweit ihr Verhalten in der Union erfolgt.

Ger: `mkCl x (mkVP (mkV2 unterliegen_V dative) y)`

Demo page: https://gdprlexicon.com

SIGNATU

9 results for "subject"

English	German	French	Italian	Spanish
be subject to X	X unterliegen	être soumise à X	essere sottoposta a X	ser sujeta a X
be subject to X	X unterstehen	être placée sous les ordres de X	sottostare X	ser sujeta a X
be subject to X	X unterworfen sein	faire l' objet de X	essere soggiogata a X	ser sujeta a X
person subject to law	Rechtsunterworfene	justiciable	persona soggetta alla legge	persona sujeta al ley
subject	betroffen	concerné	soggetto	sujeto
subject	Gegenstand	sujet	soggetto	objeto
subject-matter	Gegenstand	objet	oggetto	objeto
subject to X	vorbehaltlich X	sous reserve de X	soggetto a X	siempre que se den X
subject to X	fallend unter X	soumis à X	soggetto a X	interesados a X

Some statistics about the GDPR lexicon

	abstract
functions	3525
atomic	3272
syntactic	139
construction	114

	English	German	French	Italian	Spanish
word tokens	55186	54903	62198	55296	57383
word types	2625	4153	3206	3520	3498
lemma types	2555	3053	2467	2689	2478
multiword functions	590	227	574	559	594
multiwords in corpus	8-13%	3-10%	10-21%	10-20%	11-21%

Some statistics about the GDPR lexicon

	abstract
functions	3525
atomic	3272
syntactic	139
construction	114

	English	German	French	Italian	Spanish
word tokens	55186	54903	62198	55296	57383
word types	2625	4153	3206	3520	3498
lemma types	2555	3053	2467	2689	2478
multiword functions	590	227	574	559	594
multiwords in corpus	8-13%	3-10%	10-21%	10-20%	11-21%

Is this the largest multilingual corpus ever analysed at this level of precision?

What is not compositional

Aggregation

insurance_company__versicherungsunternehmen_CN : CN

- Fre: *compagnie d'assurance*
- Ita: *compagnia di assicurazioni*

banking_company__finanzunternehmen_CN : CN

- Fre: *banque*
- Ita: *istituto di credito*

GDPR 126.4:

- *insurance and banking companies*
- *Versicherungs- und Finanzunternehmen*
- *les compagnies d'assurance et les banques*
- *compagnie di assicurazione e istituti di credito*

Really?

we can turn any meaning assignment on a recursively enumerable set of expressions into a compositional one, as long as we can replace the syntactic operations with different ones

Zoltán Gendler Szabó, *https://plato.stanford.edu/entries/compositionality/*, citing Janssen, Theo M.V., 1983, *Foundations and Applications of Montague Grammar*, Amsterdam: Mathematisch Centrum.

On Mathematical Proofs of the Vacuity of Compositionality *

Dag Westerståhl
Stockholm University
dag.westerstahl@philosophy.su.se

February 23, 1998

doubtful — part about concatenation and function application), 'Any semantics can be made compositional' amounts to the claim that for any meaning function m from a partial algebra $\mathbf{A}$ to a set M, there is *another* meaning function μ from $\mathbf{A}$ to *another* set M^* such that (a) μ is compositional, and (b) $m(a)$ can be recovered uniformly from $\mu(a)$. Similarly for the other two cases. Now, stated thus explicitly, this claim is *trivially* true (cf. Fact 6.1), and the mathematical machinery invoked in the three examples is quite unnecessary. But then why

Yes

we can turn any meaning assignment on a recursively enumerable set of expressions into a compositional one, as long as we can replace the syntactic operations with different ones

“Replace the syntactic operations with different ones” is precisely what we have been doing when identifying the concepts.

But

Mathematically, compositional linearization in GF is restricted to operations of pairing, selection, and concatenation of strings (mildly context-sensitive).

Linguistically, concept alignment is interesting and useful as long as it identifies “natural concepts”.

Intuitively, it feels that the living language is always one step ahead.

In practice, aspects like aggregation should be treated separately, in order not to clutter the grammar itself.

Maintainability

Compositionality makes it easier to add new concepts, as the old ones don’t need changes.

On the other hand, simple “transfer” passes can make the compositional rules much simpler.

This is essentially what is done in compilers. We use the technique defined in

B. Bringert and A. Ranta. A Pattern for Almost Compositional Functions. *Journal of Functional Programming*, 18(5-6), pp. 567-598, 2008

One way to see this is as the division of labour between data and algorithms. An optimal balance maximizes the simplicity of both of them (data = grammars, algorithms = transfer passes).

Concept alignment: ideas for automation

Two different problems

1. **Concept creation**: given a parallel text, identify the concepts (units of compositional translation).

2. **Concept propagation**: given a concept and a parallel text in a new language, identify the expression for the concept in the new language.

- if this fails, you may have to create a new concept in a higher type

Methods from statistical machine translation

Word alignment by the IBM model:

- find pairs of words that translate to each other
- filter those who fit in the same category
- generalize by using morphological paradigms

Phrase alignment, a generalization of word alignment:

- find pairs of “phrases” i.e. multiword segments (can be different lengths)
- try to parse in a common category
- generalize by using linearization rules

MT techniques in a retrieval system of semantically enriched patents

M González Bermúdez, M Mateva, R Enache, L. Màrquez, B. Popov, A. Ranta, *Machine Translation Summit XIV,* 2013

Problems with standard alignment techniques

Low precision - a lot of junk

- no guarantee of common category
- false positives

A lot of data needed (e.g. because of morphological diversity)

Cannot find discontinuous phrases

Problems with distant word placements (e.g. German verbs in the end)

Syntax-based alignment: from function to arguments

Parse the parallel texts and obtain abstract syntax trees

- in a shared system of categories and functions
- both sides can initially have different words

Align matching subtrees that have a shared context:

$$f(a) \sim f(b) \Rightarrow a \sim b$$

- the result is a concept *a_b_A : A*
- iterate this to obtain bigger shared trees *f(…)*
- “given the function, infer the arguments”

Variation: from arguments to function

Find functions when arguments are given:

$$a \sim b \;\; \& \; f(a) \sim g(b) \quad \Rightarrow \quad f \sim g$$

The result is a concept *f_g_A_B : A -> B*

This generalizes to many-place functions.

Example:

x is subject to y ~ x unterliegt y

Variation: robust parsing

It is enough to have a forest of parse trees of chunks with large enough local subtrees that permit the same category.

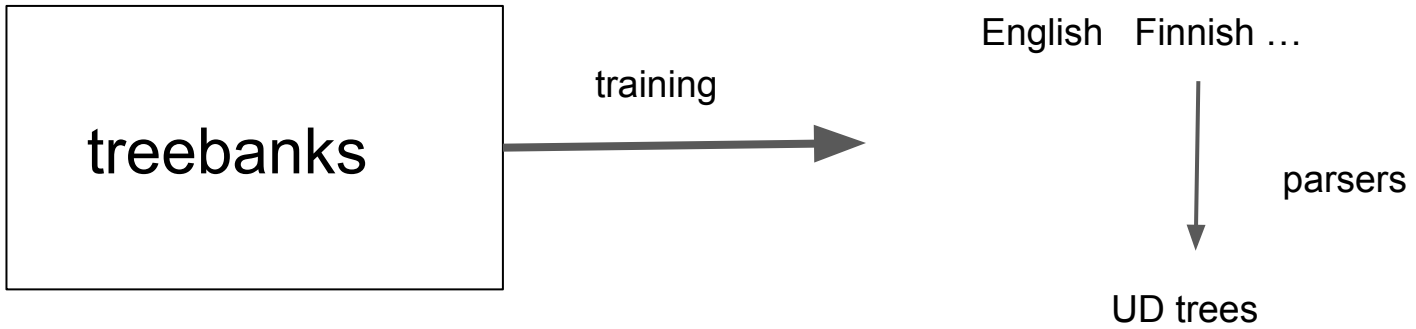
Then we can use robust parsing techniques e.g. UD (machine-learned dependency parsing) converted to GF.

UD = Universal Dependencies

Dependency tree: labelled arcs between words

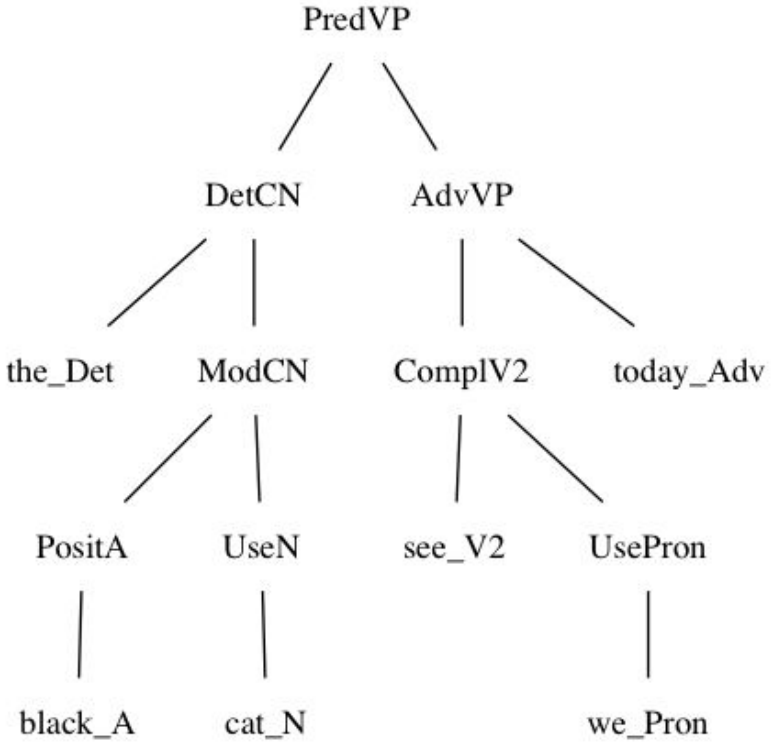
Universal: same labels in different languages

Parsing: machine-learned from treebanks



abstract syntax

```
PredVP  : NP  -> VP  -> Cl

ComplV2 : V2  -> NP  -> VP

AdvVP   : VP  -> Adv -> VP

DetCN   : Det -> CN  -> NP

ModCN   : AP  -> CN  -> CN

UseN    : N    -> CN

UsePron : Pron -> NP

PositA  : A    -> AP
```

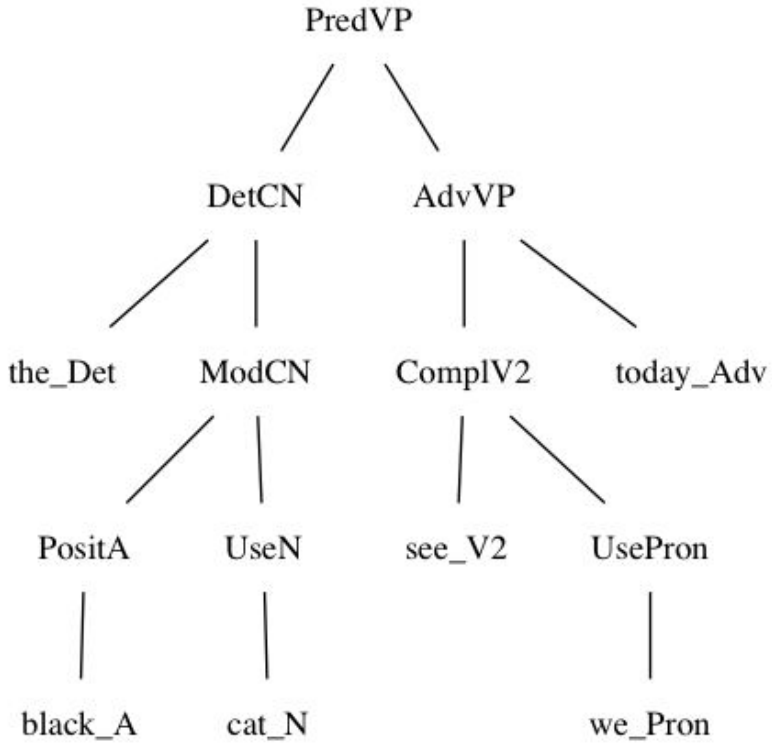


abstract syntax

```
PredVP  : NP  -> VP  -> Cl
ComplV2 : V2  -> NP  -> VP
AdvVP   : VP  -> Adv -> VP
DetCN   : Det -> CN  -> NP
ModCN   : AP  -> CN  -> CN
UseN    : N    -> CN
UsePron : Pron -> NP
PositA  : A    -> AP
```

dependency configuration

```
nsubj  head
head   dobj
head   advmod
det    head
amod   head
head
head
head
```

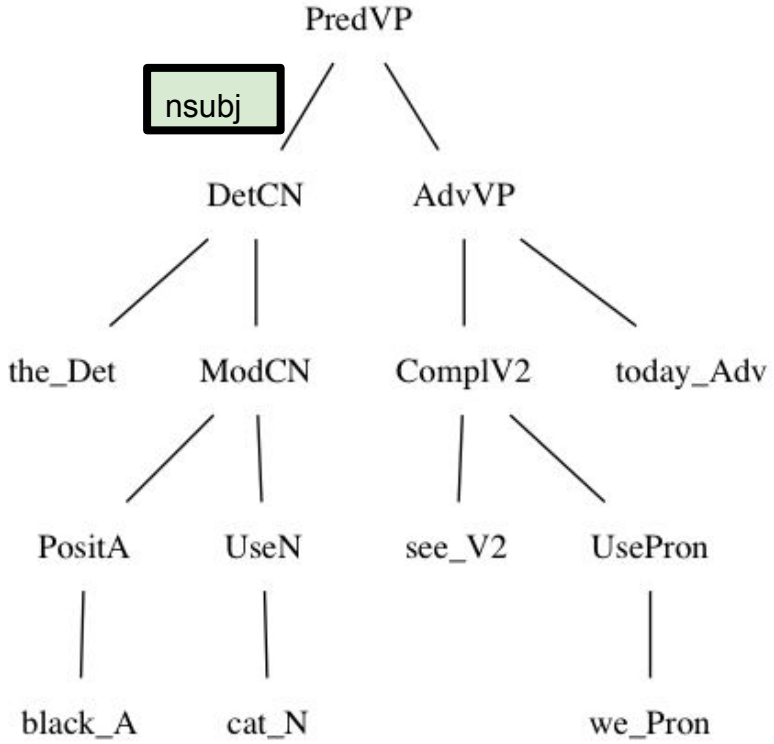


abstract syntax

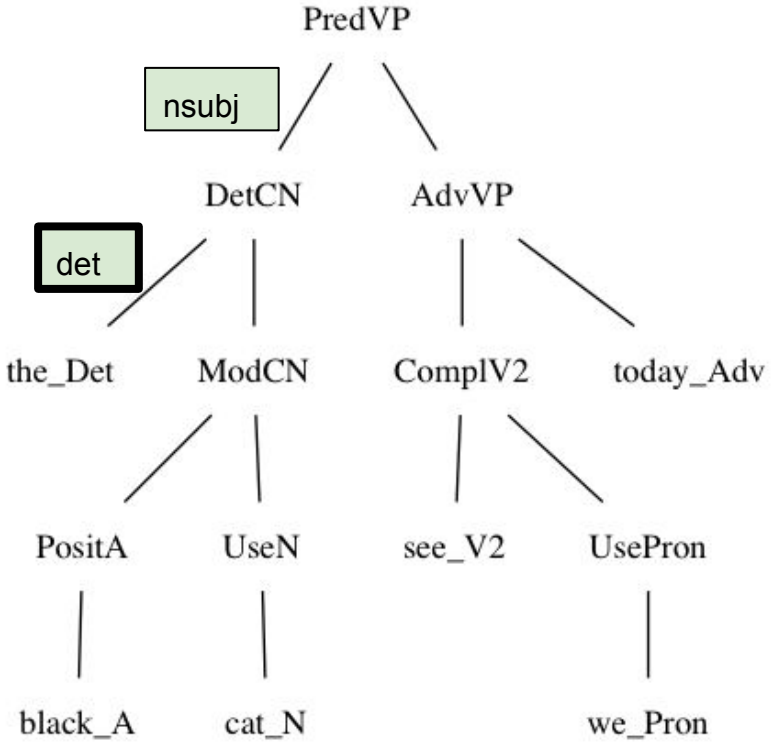
```
PredVP   : NP   -> VP   -> Cl
ComplV2  : V2   -> NP   -> VP
AdvVP    : VP   -> Adv  -> VP
DetCN    : Det  -> CN   -> NP
ModCN    : AP   -> CN   -> CN
UseN     : N    -> CN
UsePron  : Pron -> NP
PositA   : A    -> AP
```

dependency configuration

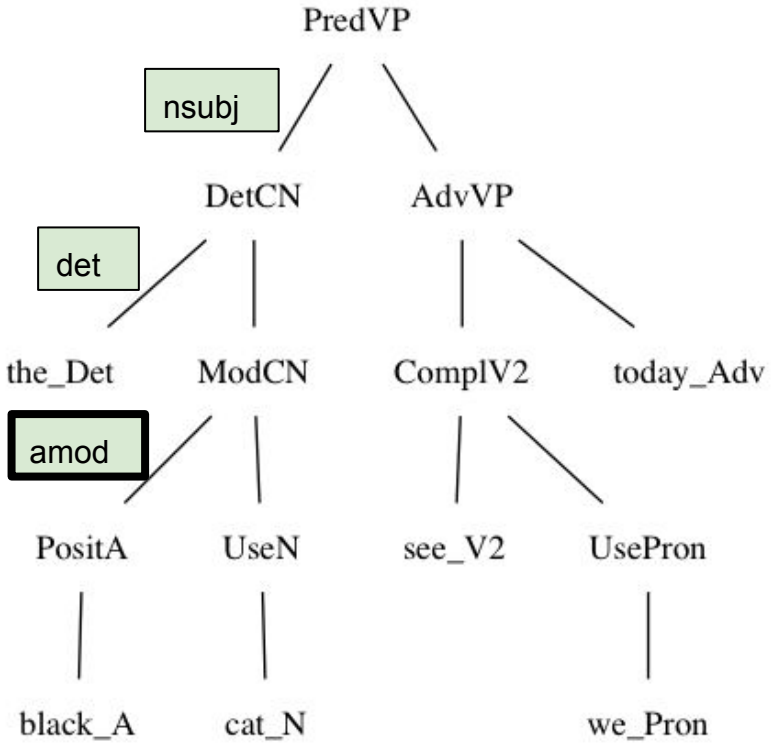
```
nsubj  head
head   dobj
head   advmod
det    head
amod   head
head
head
head
```



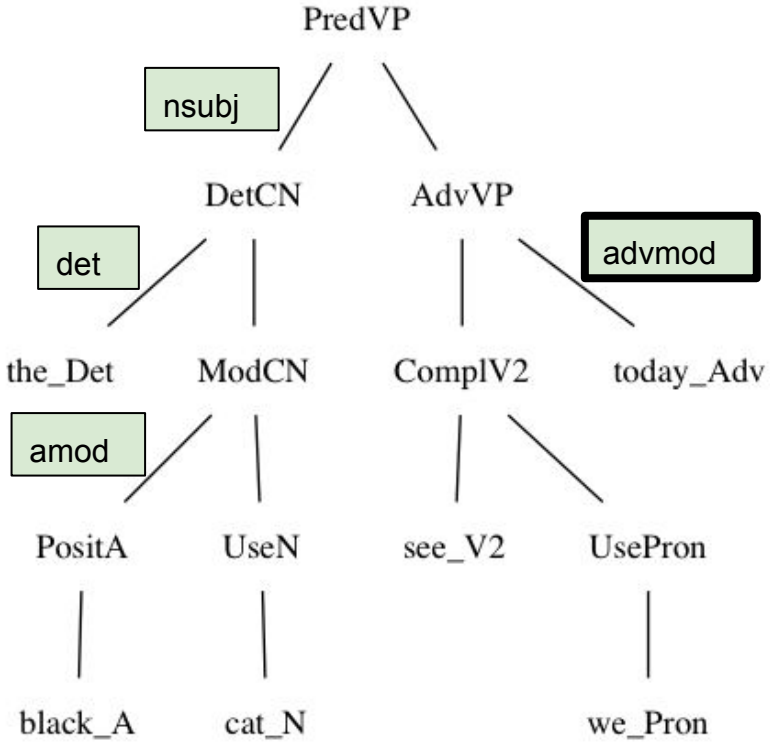
abstract syntax	dependency configuration
`PredVP : NP -> VP -> Cl`	`nsubj head`
`ComplV2 : V2 -> NP -> VP`	`head dobj`
`AdvVP : VP -> Adv -> VP`	`head advmod`
`DetCN : Det -> CN -> NP`	`det head`
`ModCN : AP -> CN -> CN`	`amod head`
`UseN : N -> CN`	`head`
`UsePron : Pron -> NP`	`head`
`PositA : A -> AP`	`head`



abstract syntax	dependency configuration
PredVP : NP -> VP -> Cl	nsubj head
ComplV2 : V2 -> NP -> VP	head dobj
AdvVP : VP -> Adv -> VP	head advmod
DetCN : Det -> CN -> NP	det head
ModCN : AP -> CN -> CN	amod head
UseN : N -> CN	head
UsePron : Pron -> NP	head
PositA : A -> AP	head



abstract syntax	dependency configuration
`PredVP : NP -> VP -> Cl`	`nsubj head`
`ComplV2 : V2 -> NP -> VP`	`head dobj`
`AdvVP : VP -> Adv -> VP`	`head advmod`
`DetCN : Det -> CN -> NP`	`det head`
`ModCN : AP -> CN -> CN`	`amod head`
`UseN : N -> CN`	`head`
`UsePron : Pron -> NP`	`head`
`PositA : A -> AP`	`head`

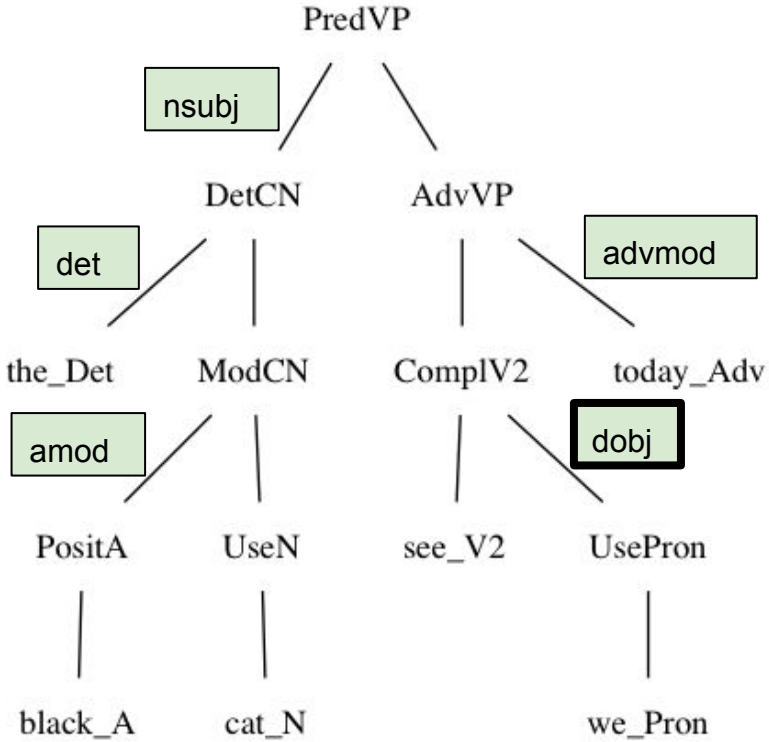


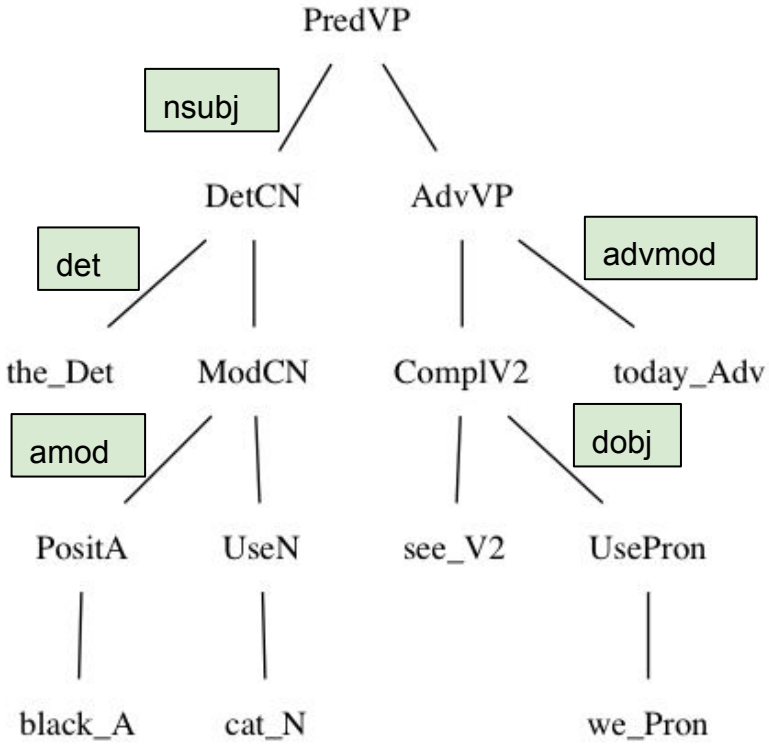
abstract syntax

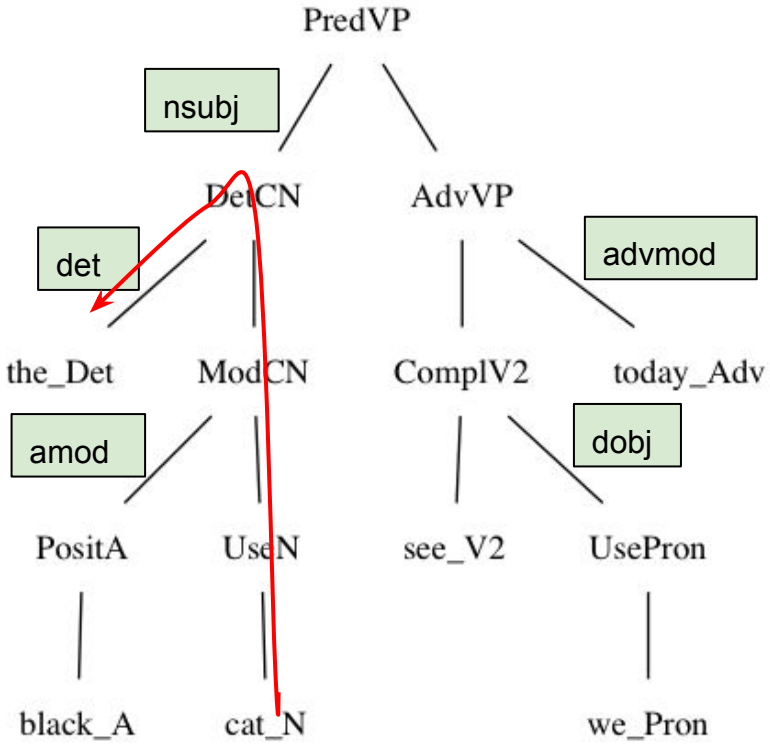
```
PredVP   : NP   -> VP   -> Cl
ComplV2  : V2   -> NP   -> VP
AdvVP    : VP   -> Adv  -> VP
DetCN    : Det  -> CN   -> NP
ModCN    : AP   -> CN   -> CN
UseN     : N    -> CN
UsePron  : Pron -> NP
PositA   : A    -> AP
```

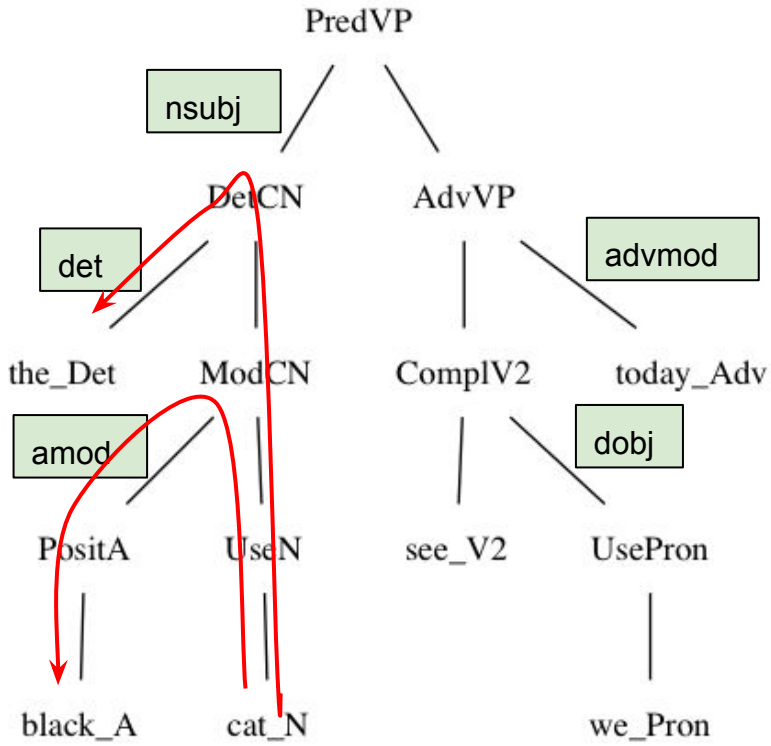
dependency configuration

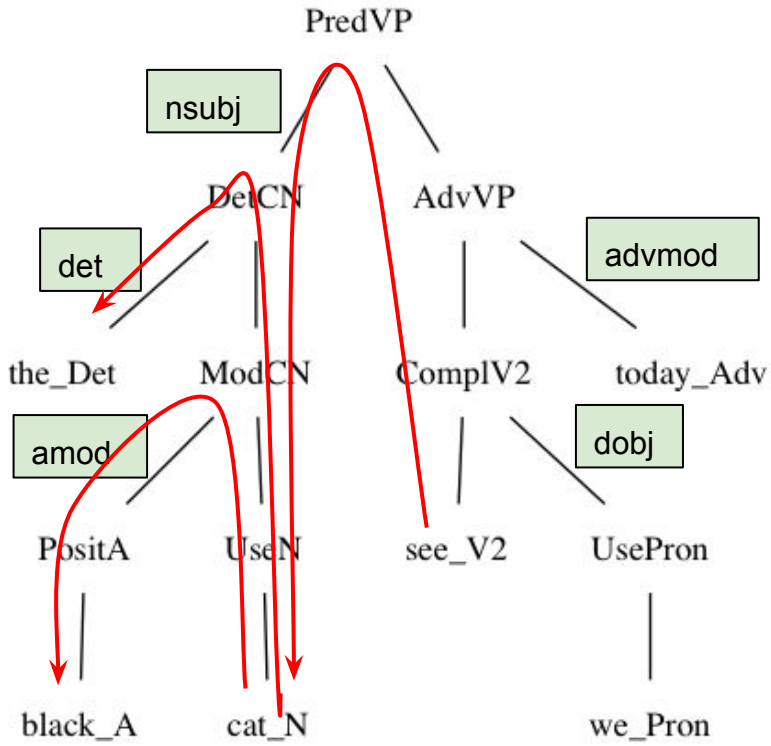
```
nsubj   head
head    dobj
head    advmod
det     head
amod    head
head
head
head
```

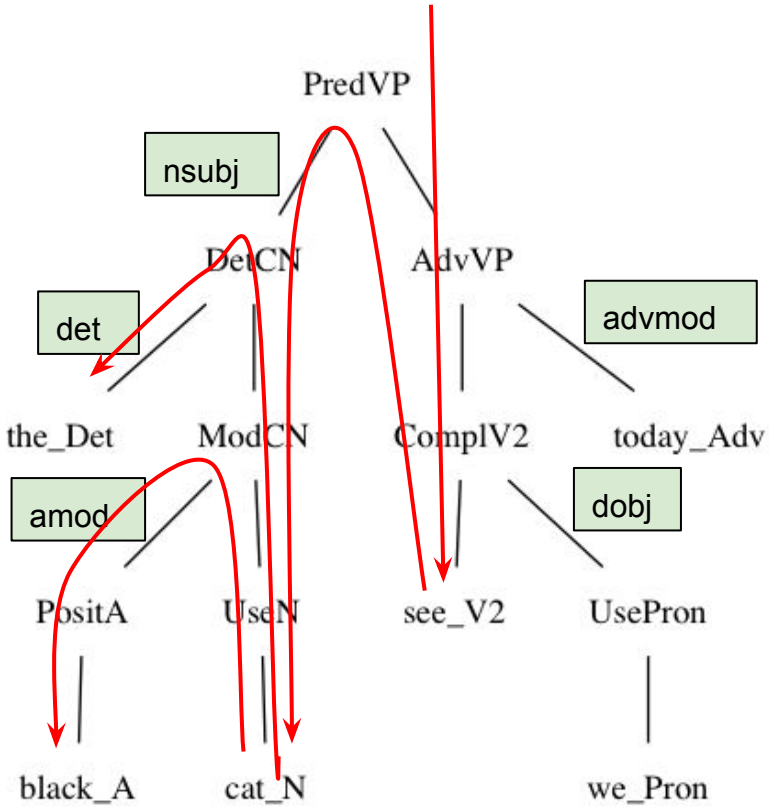


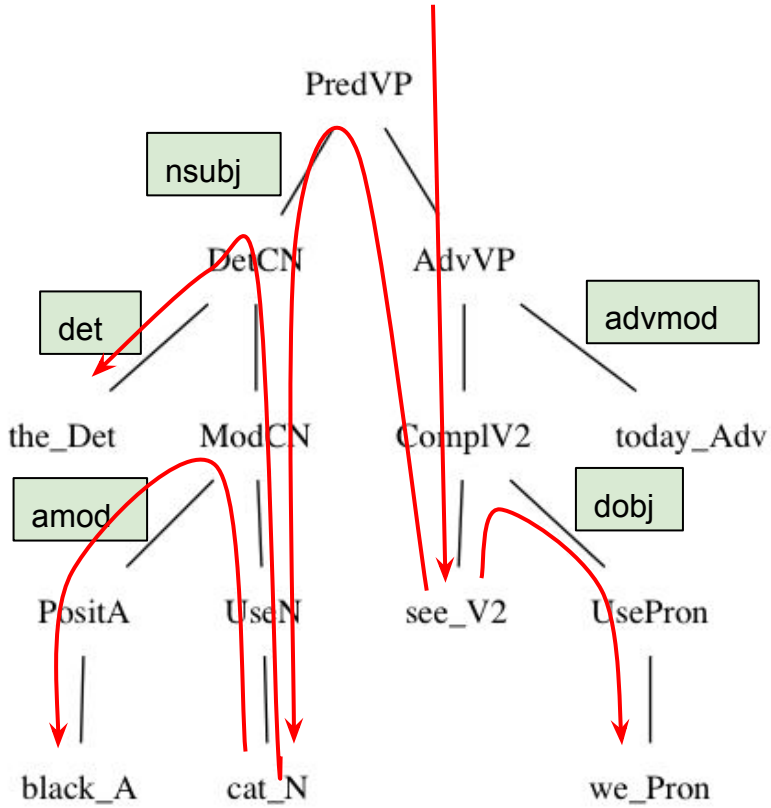

PredVP
nsubj
DetCN
AdvVP
det
advmod
the_Det
ModCN
ComplV2
today_Adv
amod
dobj
PositA
UseN
see_V2
UsePron
black_A
cat_N
we_Pron

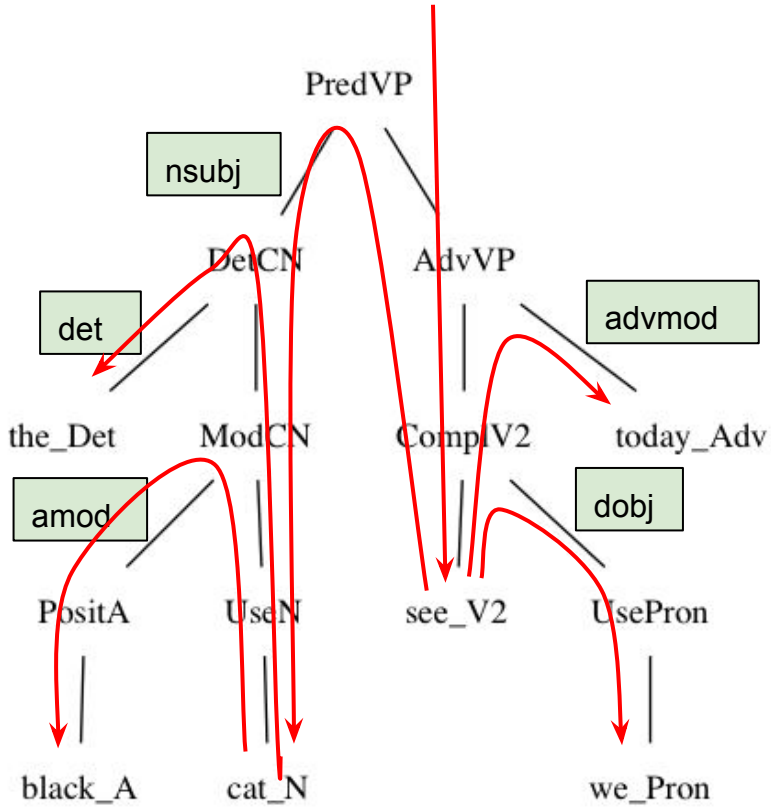

PredVP
nsubj
DetCN
AdvVP
det
advmod
the_Det
ModCN
ComplV2
today_Adv
amod
dobj
PositA
UseN
see_V2
UsePron
black_A
cat_N
we_Pron

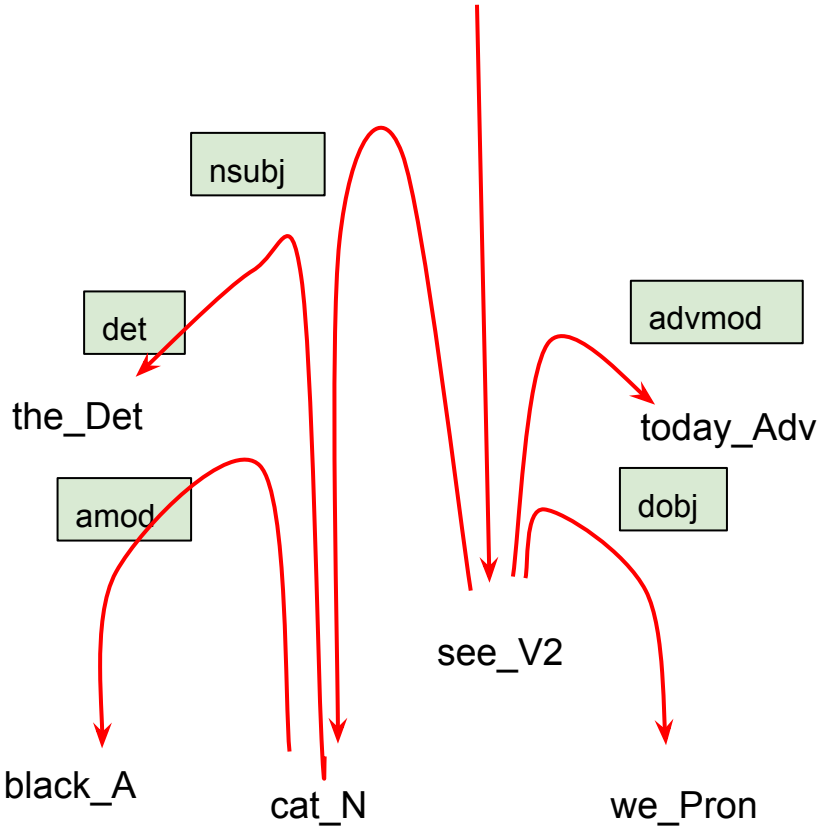

PredVP
nsubj
DetCN
AdvVP
det
advmod
the_Det
ModCN
ComplV2
today_Adv
amod
dobj
PositA
UseN
see_V2
UsePron
black_A
cat_N
we_Pron

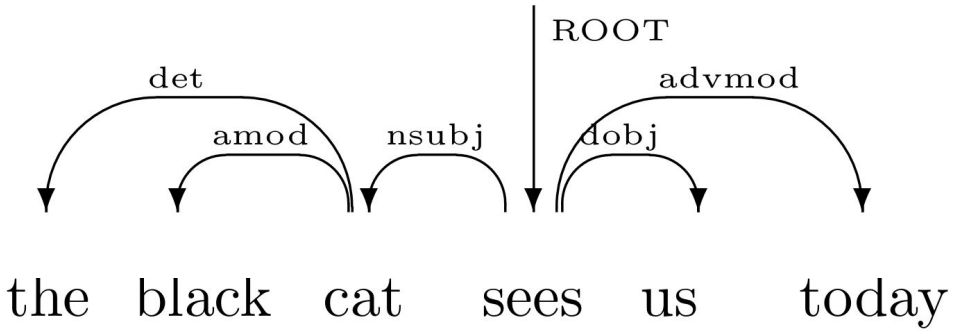

PredVP
nsubj
DetCN
AdvVP
det
advmod
the_Det
ModCN
ComplV2
today_Adv
amod
dobj
PositA
UseN
see_V2
UsePron
black_A
cat_N
we_Pron

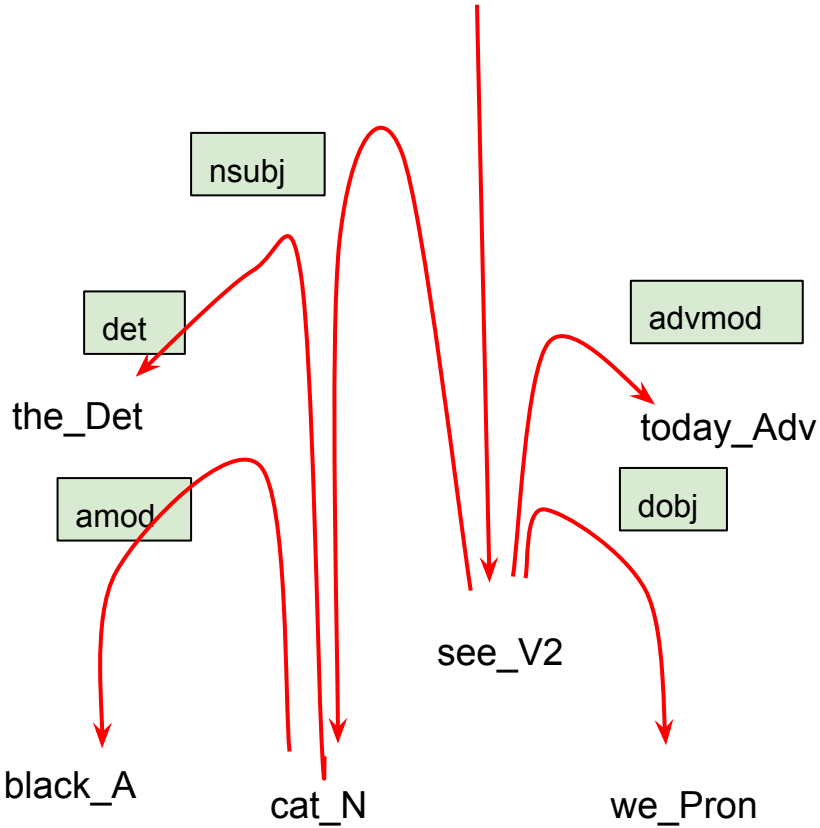

PredVP
nsubj
DetCN
AdvVP
det
advmod
the_Det
ModCN
ComplV2
today_Adv
amod
dobj
PositA
UseN
see_V2
UsePron
black_A
cat_N
we_Pron


PredVP
nsubj
DetCN
AdvVP
det
advmod
the_Det
ModCN
ComplV2
today_Adv
amod
dobj
PositA
UseN
see_V2
UsePron
black_A
cat_N
we_Pron

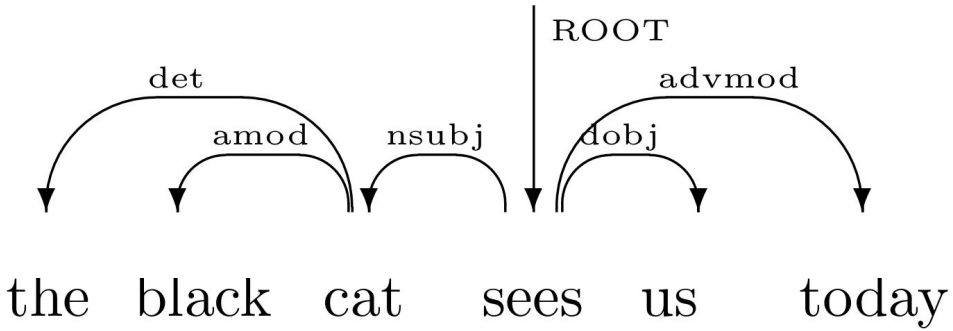

PredVP
nsubj
DetCN
AdvVP
det
advmod
the_Det
ModCN
CompV2
today_Adv
amod
dobj
PositA
UseN
see_V2
UsePron
black_A
cat_N
we_Pron

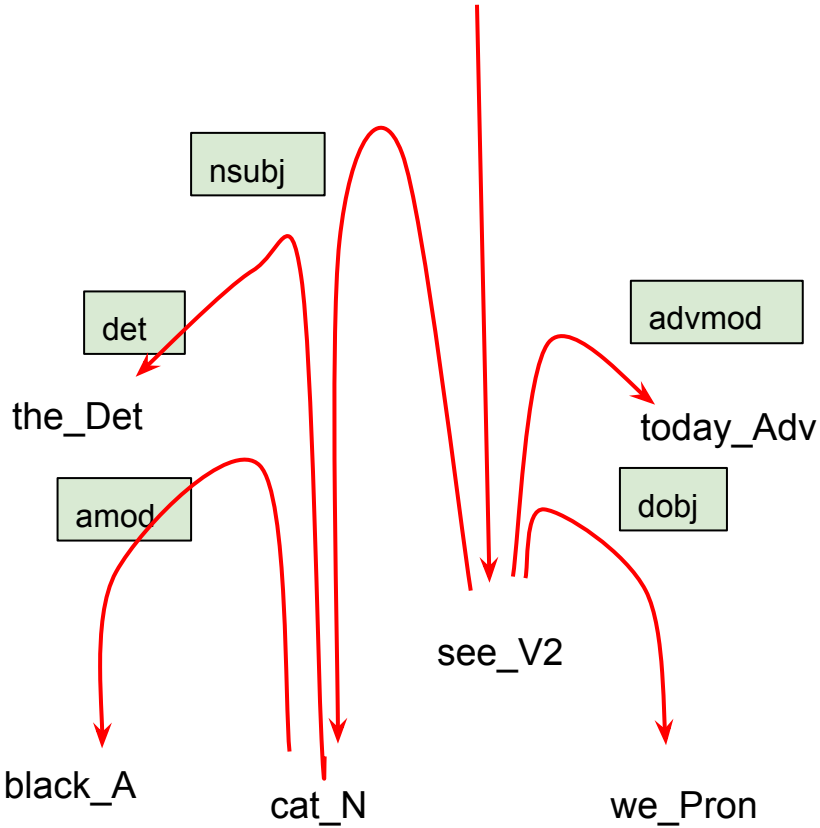

nsubj
det
the_Det
advmod
today_Adv
amod
dobj
see_V2
black_A
cat_N
we_Pron


ROOT
det
amod
nsubj
dobj
advmod
the black cat sees us today


nsubj
det
advmod
the_Det
today_Adv
amod
dobj
see_V2
black_A
cat_N
we_Pron

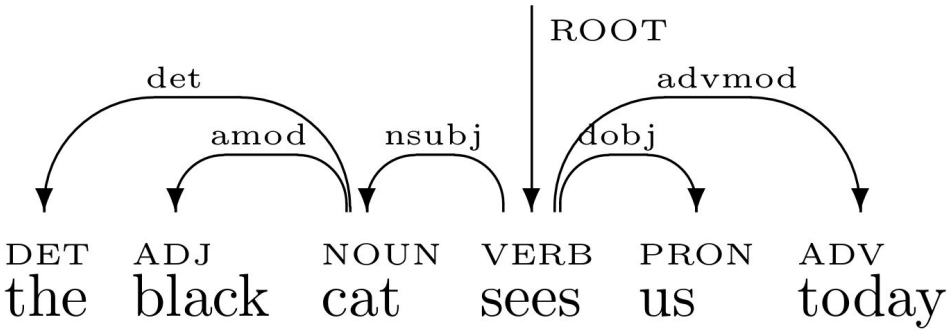
```
abstract syntax category configuration

Det       DET
A         ADJ
N         NOUN
V2        VERB
Pron      PRON
Adv       ADV
```

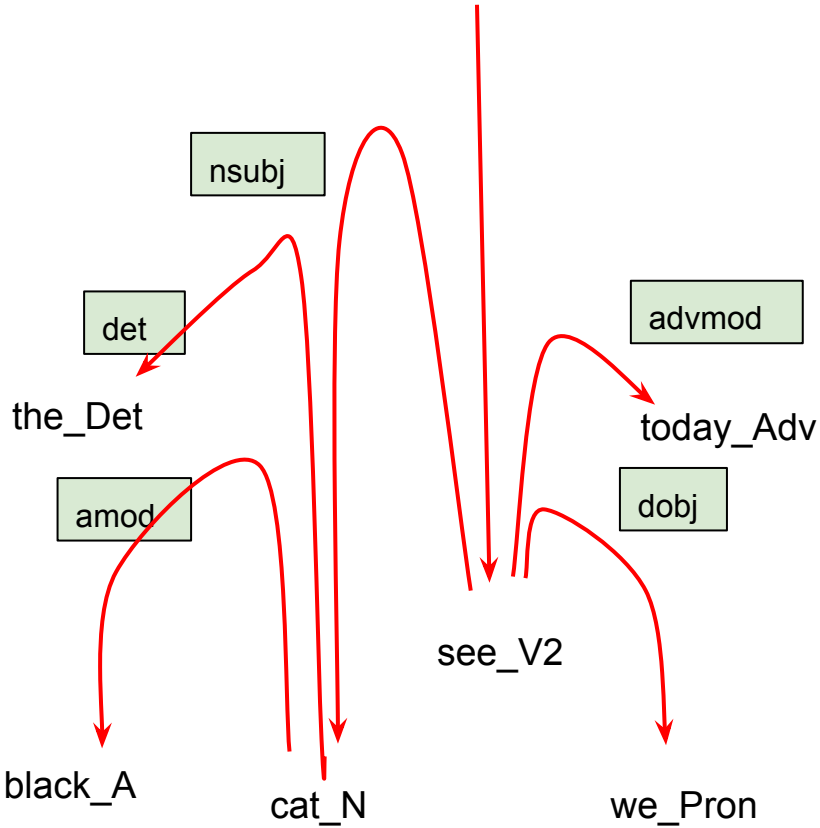




abstract syntax category configuration

Det	DET
A	ADJ
N	NOUN
V2	VERB
Pron	PRON
Adv	ADV

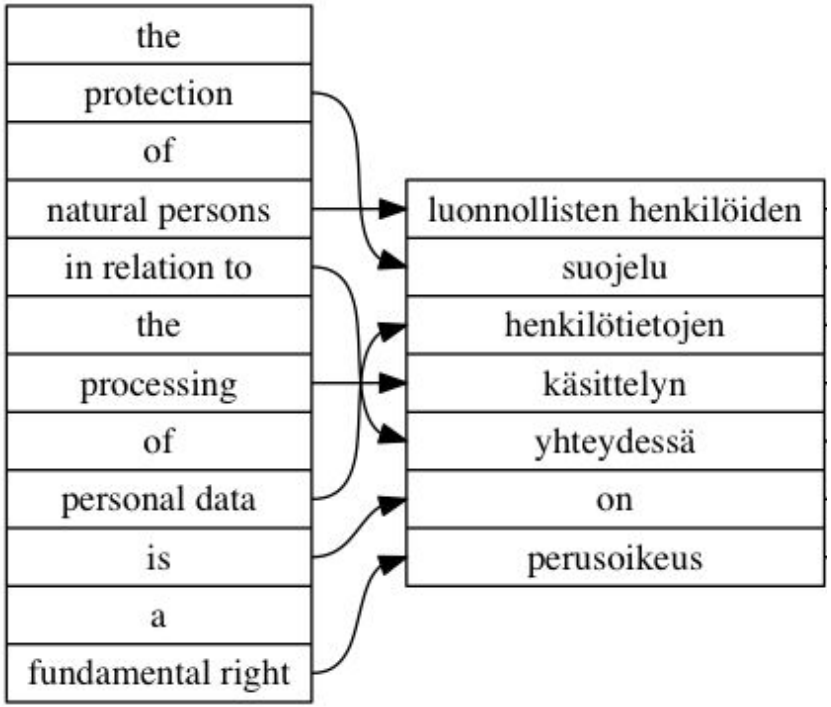


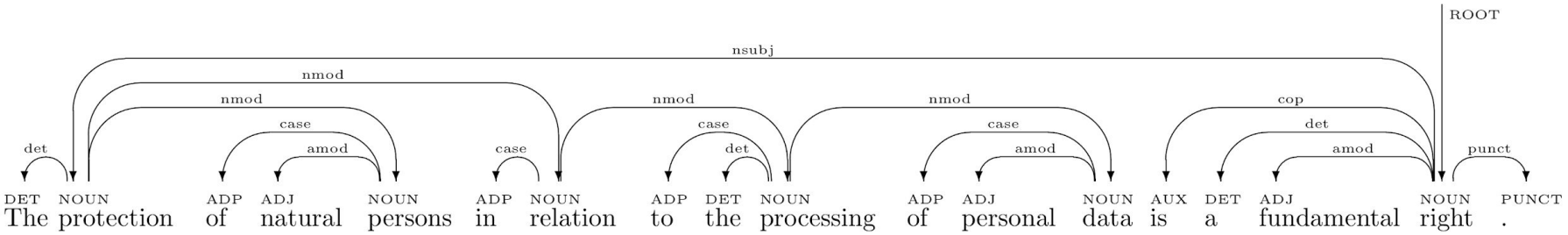


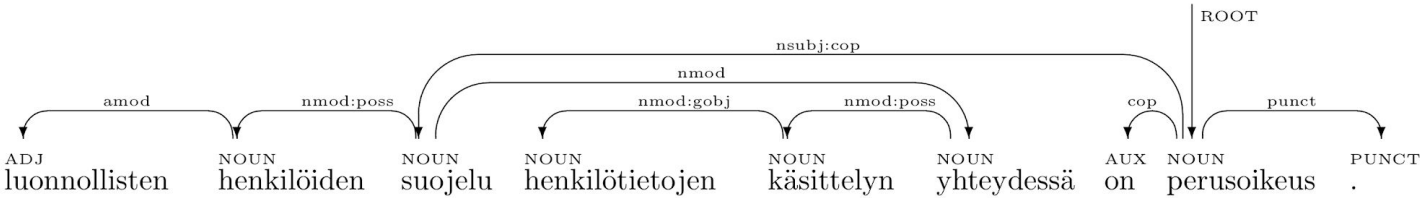
Aligning parallel UD trees

1. Order the trees in tree form
2. Recursively sort subtrees by their label
3. Add PAD nodes so that subtrees with the same label are aligned
4. Collect resulting pairs of aligned subtrees

Example: GDPR English-Finnish




ROOT
nsubj
nmod
nmod
case
amod
det
case
nmod
case
det
nmod
case
amod
cop
det
amod
punct
DET NOUN ADP ADJ NOUN ADP NOUN ADP DET NOUN ADP ADJ NOUN AUX DET ADJ NOUN PUNCT
The protection of natural persons in relation to the processing of personal data is a fundamental right .


ROOT
nsubj:cop
nmod
amod
nmod:poss
nmod:gobj
nmod:poss
cop
punct
ADJ NOUN NOUN NOUN NOUN NOUN AUX NOUN PUNCT
luonnollisten henkilöiden suojelu henkilötietojen käsittelyn yhteydessä on perusoikeus .

```
1	The	the	DET	DT	Definite=Def|PronType=Art	2	det	_	_
2	protection	protection	NOUN	NN	Number=Sing	17	nsubj	_	_
3	of	of	ADP	IN	_	5	case	_	_
4	natural	natural	ADJ	JJ	Degree=Pos	5	amod	_	_
5	persons	person	NOUN	NNS	Number=Plur	2	nmod	_	_
6	in	in	ADP	IN	_	7	case	_	_
7	relation	relation	NOUN	NN	Number=Sing	2	nmod	_	_
8	to	to	ADP	IN	_	10	case	_	_
9	the	the	DET	DT	Definite=Def|PronType=Art	10	det	_	_
10	processing	processing	NOUN	NN	Number=Sing	7	nmod	_	_
11	of	of	ADP	IN	_	13	case	_	_
12	personal	personal	ADJ	JJ	Degree=Pos	13	amod	_	_
13	data	data	NOUN	NN	Number=Sing	10	nmod	_	_
14	is	be	AUX	VBZ	Mood=Ind|Number=Sing|Person=3|Tense=Pres	17	cop	_	_
15	a	a	DET	DT	Definite=Ind|PronType=Art	17	det	_	_
16	fundamental	fundamental	ADJ	JJ	Degree=Pos	17	amod	_	_
17	right	right	NOUN	NN	Number=Sing	0	root	_	_
18	.	.	PUNCT	.	_	17	punct	_	_
```

```
1	luonnollisten	luonnollinen	ADJ	A	Case=Gen|Degree=Pos|Number=Plur	2	amod	_	_
2	henkilöiden	henkilö	NOUN	N	Case=Gen|Number=Plur	3	nmod:poss	_	_
3	suojelu	suojelu	NOUN	N	Case=Nom|Number=Sing	8	nsubj:cop	_	_
4	henkilötietojen	henkilö#tiedot	NOUN	N	Case=Gen|Number=Plur	5	nmod:gobj	_	_
5	käsittelyn	käsittely	NOUN	N	Case=Gen|Number=Sing	6	nmod:poss	_	_
6	yhteydessä	yhteys	NOUN	N	Case=Ine|Number=Sing	3	nmod	_	_
7	on	olla	AUX	V	Mood=Ind|Number=Sing|Person=3|Tense=Pres	8	cop	_	_
8	perusoikeus	perus#oikeus	NOUN	N	Case=Nom|Number=Sing	0	root	_	_
9	.	.	PUNCT	Punct	_	8	punct	_	SpacesAfter
```

1. Order the trees in tree form

```
root right NOUN Number=Sing 17
    nsubj protection NOUN Number=Sing 2
        det the DET Definite=Def|PronType=Art 1
        nmod person NOUN Number=Plur 5
            case of ADP _ 3
            amod natural ADJ Degree=Pos 4
        nmod relation NOUN Number=Sing 7
            case in ADP _ 6
            nmod processing NOUN Number=Sing 10
                case to ADP _ 8
                det the DET Definite=Def|PronType=Art 9
                nmod data NOUN Number=Sing 13
                    case of ADP _ 11
                    amod personal ADJ Degree=Pos 12
    cop be AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 14
    det a DET Definite=Ind|PronType=Art 15
    amod fundamental ADJ Degree=Pos 16
    punct . PUNCT _ 18
```

```
root perus#oikeus NOUN Case=Nom|Number=Sing 8
    nsubj:cop suojelu NOUN Case=Nom|Number=Sing 3
        nmod:poss henkilö NOUN Case=Gen|Number=Plur 2
            amod luonnollinen ADJ Case=Gen|Number=Plur 1
        nmod yhteys NOUN Case=Ine|Number=Sing 6
            nmod:poss käsittely NOUN Case=Gen|Number=Sing 5
                nmod:gobj henkilö#tiedot NOUN Case=Gen 4
    cop olla AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 7
    punct . PUNCT _ 9
```

2&3. Sort by label and add PAD nodes

```
root right NOUN Number=Sing 17
    nsubj protection NOUN Number=Sing 2
        det the DET Definite=Def|PronType=Art 1
        nmod person NOUN Number=Plur 5
            case of ADP _ 3
            amod natural ADJ Degree=Pos 4
        nmod relation NOUN Number=Sing 7
            case in ADP _ 6
            nmod processing NOUN Number=Sing 10
                case to ADP _ 8
                det the DET Definite=Def|PronType=Art 9
                nmod data NOUN Number=Sing 13
                    case of ADP _ 11
                    amod personal ADJ Degree=Pos 12
    cop be AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 14
    det a DET Definite=Ind|PronType=Art 15
    amod fundamental ADJ Degree=Pos 16
    punct . PUNCT _ 18
```

```
root perus#oikeus NOUN Case=Nom|Number=Sing 8
    nsubj:cop suojelu NOUN Case=Nom|Number=Sing 3
        PAD
        nmod:poss henkilö NOUN Case=Gen|Number=Plur 2
            PAD
            amod luonnollinen ADJ Case=Gen|Number=Plur 1
        nmod yhteys NOUN Case=Ine|Number=Sing 6
            PAD
            nmod:poss käsittely NOUN Case=Gen|Number=Sing 5
                PAD
                PAD
                nmod:gobj henkilö#tiedot NOUN Case=Gen 4
                    PAD
                    PAD
    cop olla AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 7
    PAD
    PAD
    punct . PUNCT _ 9
```

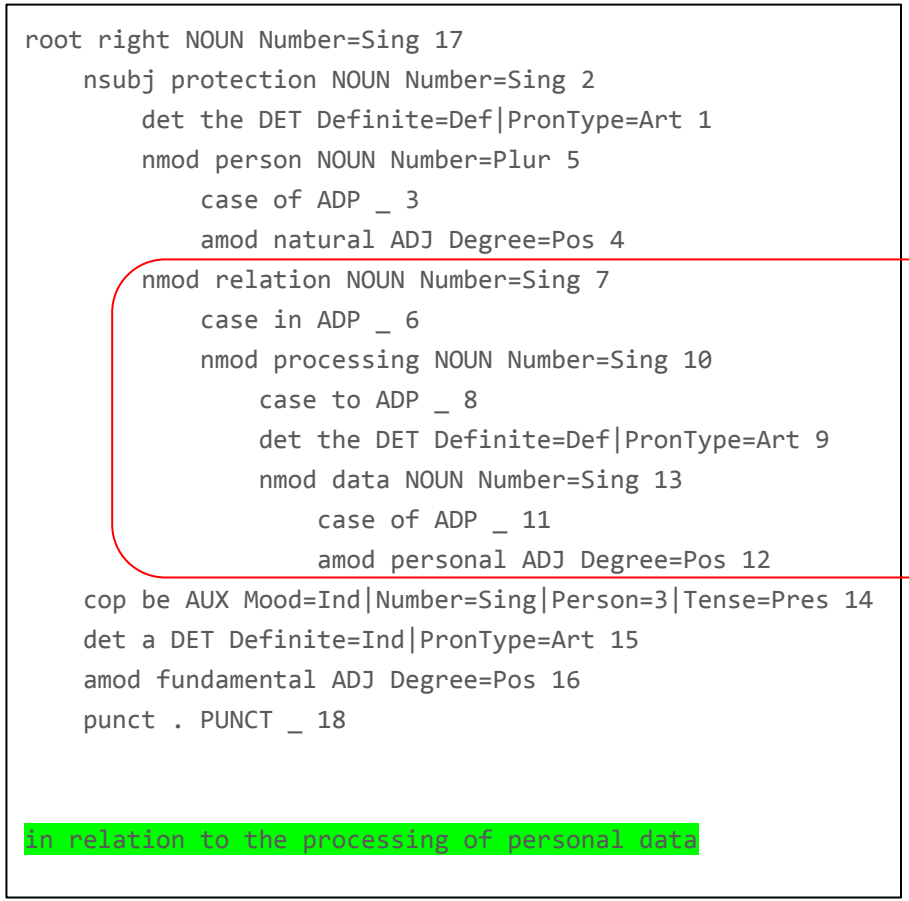
4. Collect resulting pairs of aligned subtrees

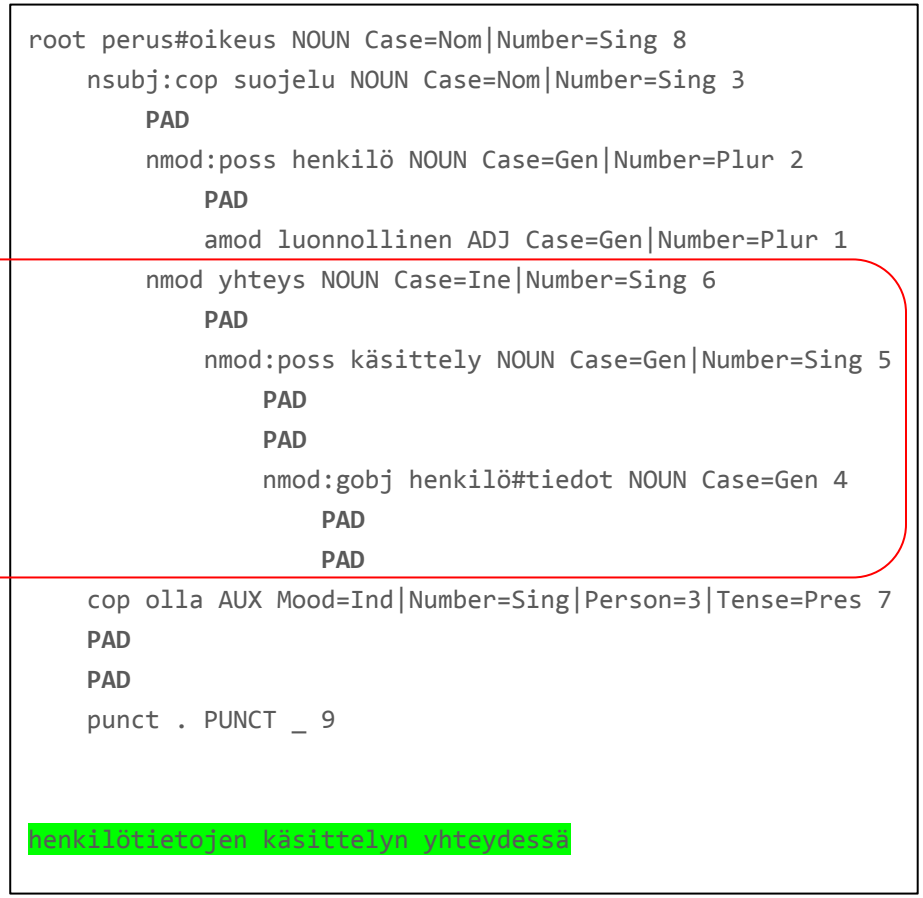
```
root right NOUN Number=Sing 17
    nsubj protection NOUN Number=Sing 2
        det the DET Definite=Def|PronType=Art 1
        nmod person NOUN Number=Plur 5
            case of ADP _ 3
            amod natural ADJ Degree=Pos 4
        nmod relation NOUN Number=Sing 7
            case in ADP _ 6
            nmod processing NOUN Number=Sing 10
                case to ADP _ 8
                det the DET Definite=Def|PronType=Art 9
                nmod data NOUN Number=Sing 13
                    case of ADP _ 11
                    amod personal ADJ Degree=Pos 12
    cop be AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 14
    det a DET Definite=Ind|PronType=Art 15
    amod fundamental ADJ Degree=Pos 16
    punct . PUNCT _ 18

natural
be
.
```

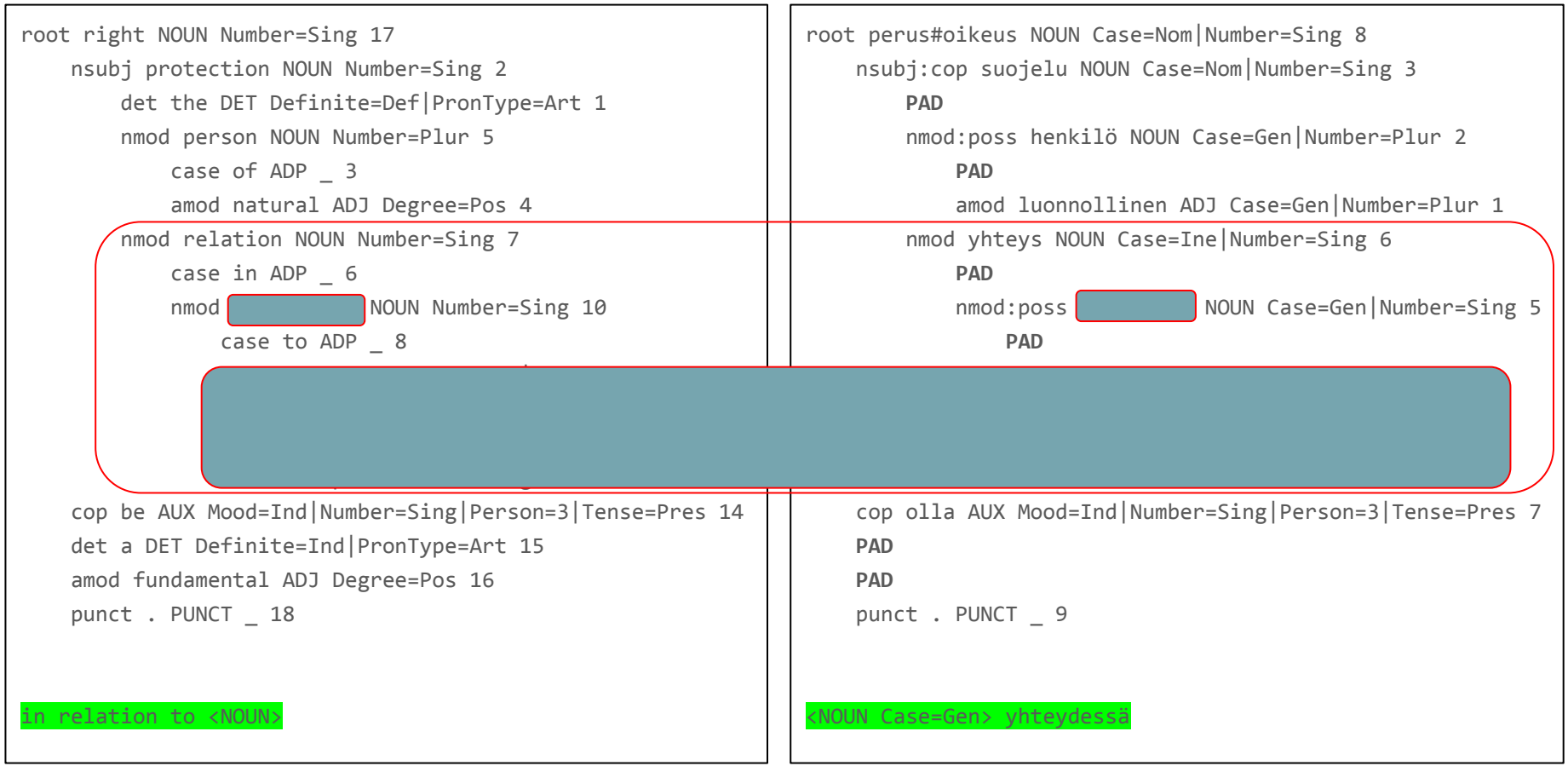
```
root perus#oikeus NOUN Case=Nom|Number=Sing 8
    nsubj:cop suojelu NOUN Case=Nom|Number=Sing 3
        PAD
        nmod:poss henkilö NOUN Case=Gen|Number=Plur 2
            PAD
            amod luonnollinen ADJ Case=Gen|Number=Plur 1
        nmod yhteys NOUN Case=Ine|Number=Sing 6
            PAD
            nmod:poss käsittely NOUN Case=Gen|Number=Sing 5
                PAD
                PAD
                nmod:gobj henkilö#tiedot NOUN Case=Gen 4
                    PAD
                    PAD
    cop olla AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 7
    PAD
    PAD
    punct . PUNCT _ 9

luonnollinen
olla
.
```

```
root right NOUN Number=Sing 17
    nsubj protection NOUN Number=Sing 2
        det the DET Definite=Def|PronType=Art 1
        nmod person NOUN Number=Plur 5
            case of ADP _ 3
            amod natural ADJ Degree=Pos 4
        nmod relation NOUN Number=Sing 7
            case in ADP _ 6
            nmod processing NOUN Number=Sing 10
                case to ADP _ 8
                det the DET Definite=Def|PronType=Art 9
                nmod data NOUN Number=Sing 13
                    case of ADP _ 11
                    amod personal ADJ Degree=Pos 12
    cop be AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 14
    det a DET Definite=Ind|PronType=Art 15
    amod fundamental ADJ Degree=Pos 16
    punct . PUNCT _ 18

natural person
personal data
```

```
root perus#oikeus NOUN Case=Nom|Number=Sing 8
    nsubj:cop suojelu NOUN Case=Nom|Number=Sing 3
        PAD
        nmod:poss henkilö NOUN Case=Gen|Number=Plur 2
            PAD
            amod luonnollinen ADJ Case=Gen|Number=Plur 1
        nmod yhteys NOUN Case=Ine|Number=Sing 6
            PAD
            nmod:poss käsittely NOUN Case=Gen|Number=Sing 5
                PAD
                PAD
                nmod:gobj henkilö#tiedot NOUN Case=Gen 4
                    PAD
                    PAD
    cop olla AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 7
    PAD
    PAD
    punct . PUNCT _ 9

luonnollinen henkilö
henkilötiedot
```


root right NOUN Number=Sing 17
nsubj protection NOUN Number=Sing 2
det the DET Definite=Def|PronType=Art 1
nmod person NOUN Number=Plur 5
case of ADP _ 3
amod natural ADJ Degree=Pos 4
nmod relation NOUN Number=Sing 7
case in ADP _ 6
nmod processing NOUN Number=Sing 10
case to ADP _ 8
det the DET Definite=Def|PronType=Art 9
nmod data NOUN Number=Sing 13
case of ADP _ 11
amod personal ADJ Degree=Pos 12
cop be AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 14
det a DET Definite=Ind|PronType=Art 15
amod fundamental ADJ Degree=Pos 16
punct . PUNCT _ 18
in relation to the processing of personal data


root perus#oikeus NOUN Case=Nom|Number=Sing 8
nsubj:cop suojelu NOUN Case=Nom|Number=Sing 3
PAD
nmod:poss henkilö NOUN Case=Gen|Number=Plur 2
PAD
amod luonnollinen ADJ Case=Gen|Number=Plur 1
nmod yhteys NOUN Case=Ine|Number=Sing 6
PAD
nmod:poss käsittely NOUN Case=Gen|Number=Sing 5
PAD
PAD
nmod:gobj henkilö#tiedot NOUN Case=Gen 4
PAD
PAD
cop olla AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 7
PAD
PAD
punct . PUNCT _ 9
henkilötietojen käsittelyn yhteydessä


root right NOUN Number=Sing 17
nsubj protection NOUN Number=Sing 2
det the DET Definite=Def|PronType=Art 1
nmod person NOUN Number=Plur 5
case of ADP _ 3
amod natural ADJ Degree=Pos 4
nmod relation NOUN Number=Sing 7
case in ADP _ 6
nmod NOUN Number=Sing 10
case to ADP _ 8
cop be AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 14
det a DET Definite=Ind|PronType=Art 15
amod fundamental ADJ Degree=Pos 16
punct . PUNCT _ 18
in relation to <NOUN>
root perus#oikeus NOUN Case=Nom|Number=Sing 8
nsubj:cop suojelu NOUN Case=Nom|Number=Sing 3
PAD
nmod:poss henkilö NOUN Case=Gen|Number=Plur 2
PAD
amod luonnollinen ADJ Case=Gen|Number=Plur 1
nmod yhteys NOUN Case=Ine|Number=Sing 6
PAD
nmod:poss NOUN Case=Gen|Number=Sing 5
PAD
cop olla AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres 7
PAD
PAD
punct . PUNCT _ 9
<NOUN Case=Gen> yhteydessä

After the first implementation, there’s still much to do

Very low precision:

```
(3,("NOUN/NOUN",(["subject"],["Daten"])))
(3,("NOUN/NOUN",(["protection"],["Daten"])))
(3,("NOUN/NOUN",(["processing"],["Verantwortliche"])))
(3,("NOUN/NOUN",(["data"],["Verarbeitung"])))
(3,("NOUN/NOUN",(["controller"],["Verarbeitung"])))
(3,("NOUN/NOUN",(["controller"],["Aufsichtsbehörde"])))
(3,("NOUN/NOUN",(["authority"],["Verantwortliche"])))
(3,("NOUN/NOUN",(["authority"],["Aufsichtsbehörden"])))
(3,("NOUN/NOUN",(["Board"],["Aufsichtsbehörde"])))
```

TODO:

- improve the algorithm
- recover from parse errors
- recover from unnecessary cross-lingual differences in UD parsing

Cross-lingual variations in UD parsing

```
root right NOUN Number=Sing 17                                                  |||  root Grundrecht NOUN _ 12
    amod fundamental ADJ Degree=Pos 16                                          |||      pad   _ 0
    cop be AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 14         |||      cop sein VERB Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 10
    det a DET Definite=Ind|PronType=Art 15                                      |||      det ein DET Definite=Ind|PronType=Art 11
    nsubj protection NOUN Number=Sing 2                                         |||      nsubj Datum NOUN Case=Nom|Gender=Neut|Number=Plur 9
        pad   _ 0                                                               |||          amod personenbezogen ADJ Case=Nom|Degree=Pos|Gender=Masc,Neut|Number=Sing 8
            pad   _ 0                                                           |||              nmod Verarbeitung NOUN Case=Dat|Gender=Fem|Number=Sing 7
                pad   _ 0                                                       |||                  case bei ADP _ 5
                pad   _ 0                                                       |||                  det der DET Case=Dat|Definite=Def|Gender=Fem|Number=Sing|PronType=Art 6
        det the DET Definite=Def|PronType=Art 1                                 |||          pad   _ 0
        nmod person NOUN Number=Plur 5                                          |||          pad   _ 0
            amod natural ADJ Degree=Pos 4                                       |||              pad   _ 0
            case of ADP _ 3                                                     |||              pad   _ 0
            nmod relation NOUN Number=Sing 7                                    |||              pad   _ 0
                case in ADP _ 6                                                 |||                  pad   _ 0
        nmod processing NOUN Number=Sing 10                                     |||          pad   _ 0
            case to ADP _ 8                                                     |||              pad   _ 0
            det the DET Definite=Def|PronType=Art 9                             |||              pad   _ 0
            nmod data NOUN Number=Sing 13                                       |||              pad   _ 0
                amod personal ADJ Degree=Pos 12                                 |||                  pad   _ 0
                case of ADP _ 11                                                |||                  pad   _ 0
        pad   _ 0                                                               |||          nsubj Schutz NOUN Case=Nom|Gender=Masc|Number=Sing 2
            pad   _ 0                                                           |||              det der DET Case=Nom|Definite=Def|Gender=Masc|Number=Sing|PronType=Art 1
            pad   _ 0                                                           |||              nmod Person NOUN _ 4
                pad   _ 0                                                       |||                  amod natürlicher ADJ Degree=Cmp,Pos 3
    punct . PUNCT _ 18                                                          |||      punct . PUNCT _ 13
```

Conclusion

Translation is compositional when performed on appropriate abstract concepts.

These concepts can be realized differently in different languages.

Concept alignment is the process of finding them in a parallel corpus.

Starting with string-based alignment, we move to syntax-based methods.

Concepts need not be unambiguous semantic units.

Compositional translation may be broken by aggregation.

The GDPR project: 3500 concepts, 5 languages - 22 to do.

Concepts can be found without exact parsing, e.g. from dependency trees.