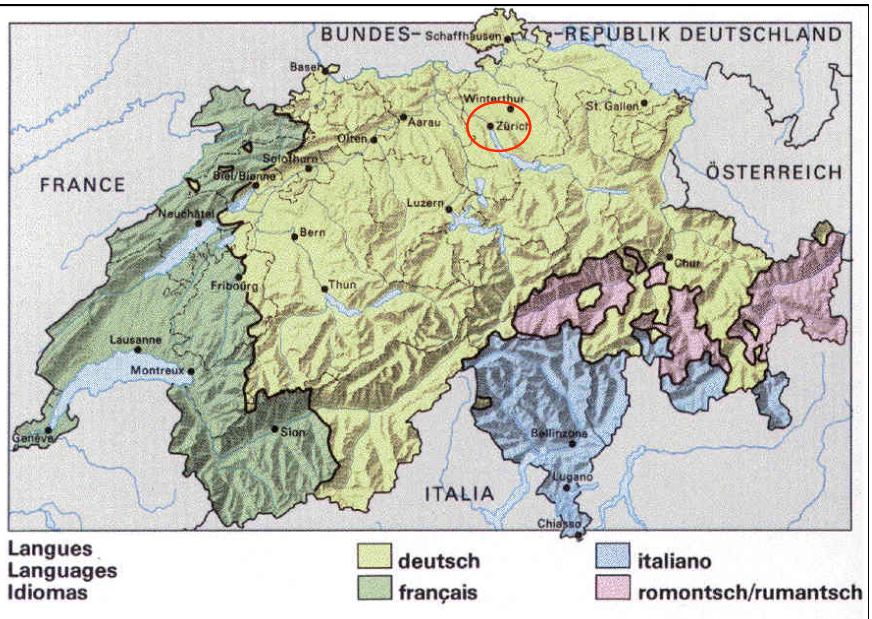
University of Zurich[UZH]

Institute of Computational Linguistics

Multiword Expressions in Parallel Treebanks

Martin Volk
volk@cl.uzh.ch
Institut für Computerlinguistik
Universität Zürich
Parseme COST, Warschau, September 2013



University of Zurich[UZH]

Institute of Computational Linguistics

- Biomedical Text Mining (with Novartis)
- Style Checking in Law Texts
- Sentiment Detection
- Digital Humanities (digitizing heritage texts)
- Machine Translation

University of Zurich[UZH]

Institute of Computational Linguist

Teaching

Bachelor and Master in Computational Linguistics

Multiword / Multitoken Expressions

In this talk:

- Names
  - Person, Location, Organization, Event
- Dates and Numbers
- Support/Light verb constructions
- DE verbs with separated prefixes / EN particle verbs
- ES and FR Multiword prepositions

Excluded:

– Idioms, Collocations, …

University of Zurich[UZH]
Institute of Computational Linguistics

Multiword / Multitoken Expressions

Contiguous … vs.

- Names → *John F. Kennedy*

… Discontiguous

- DE verbs with separated prefixes → *fängt das Spiel an*

Static … vs.

- Dates → *15. June 2005*

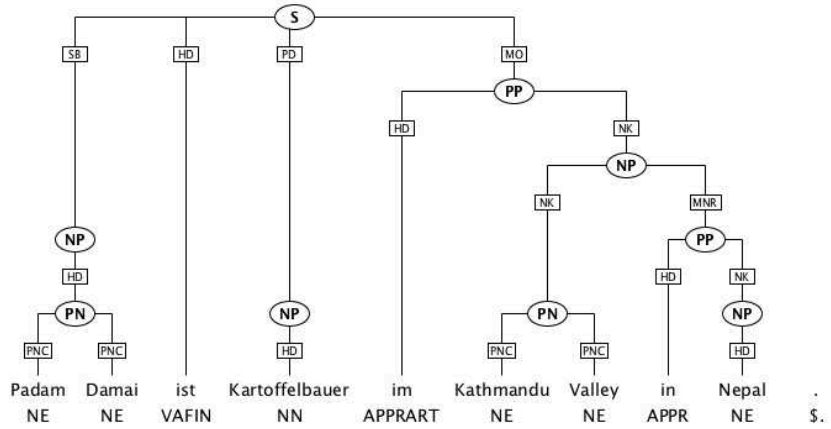
… Dynamic / Inflectable / Modifiable

- Support verb constructions → *zur Verfügung stellen*
- Geo names → *refuge solitaire de Sponda*

University of Zurich[UZH]

Institute of Computational Linguistics

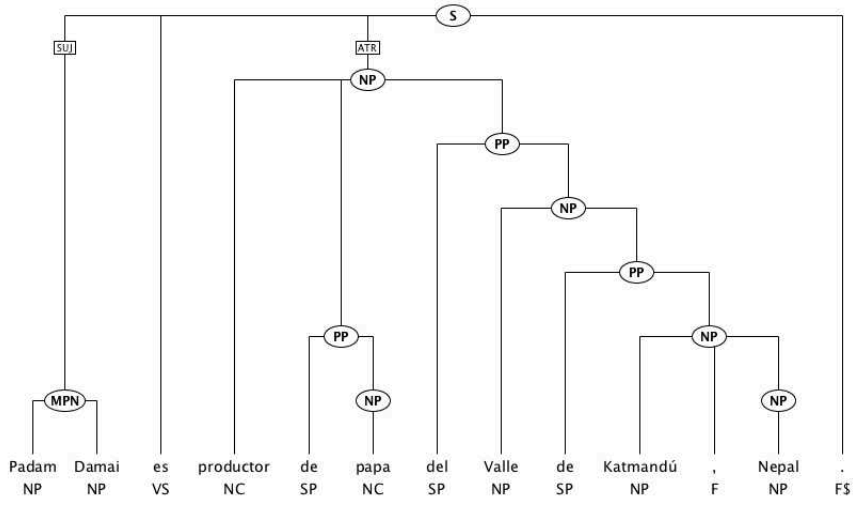
MWE: Proper Names - DE



University of Zurich[UZH]

Institute of Computational Linguistics

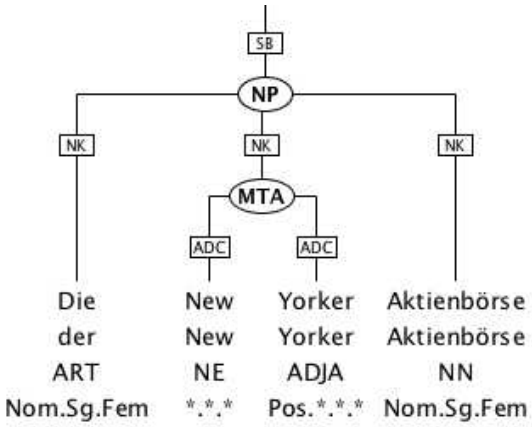
MWE: Proper Names - ES



University of Zurich[UZH]

Institute of Computational Linguistics

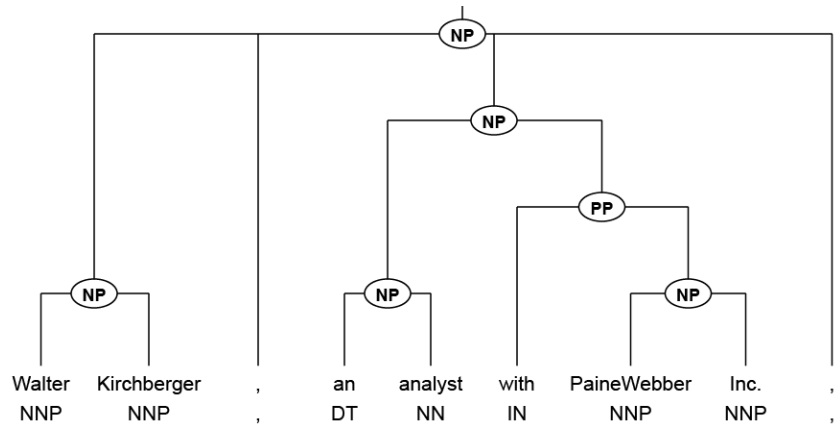
MWE: DE Proper Names as Adjectives



University of Zurich[UZH]

Institute of Computational Linguistics

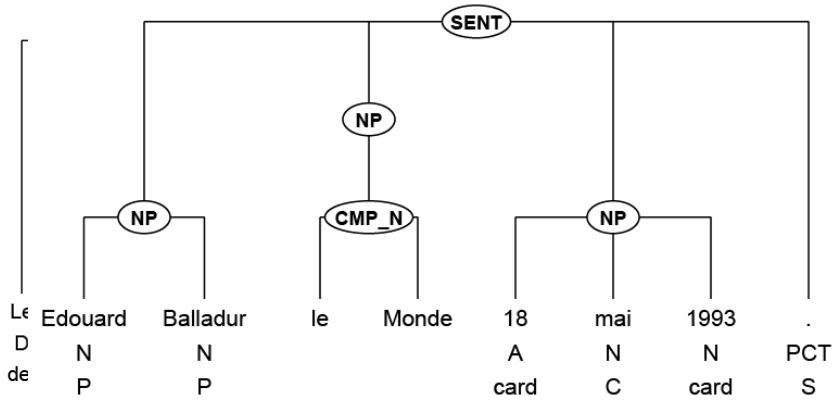
MWE: EN Proper Names in Penn Treebank



University of Zurich[UZH]

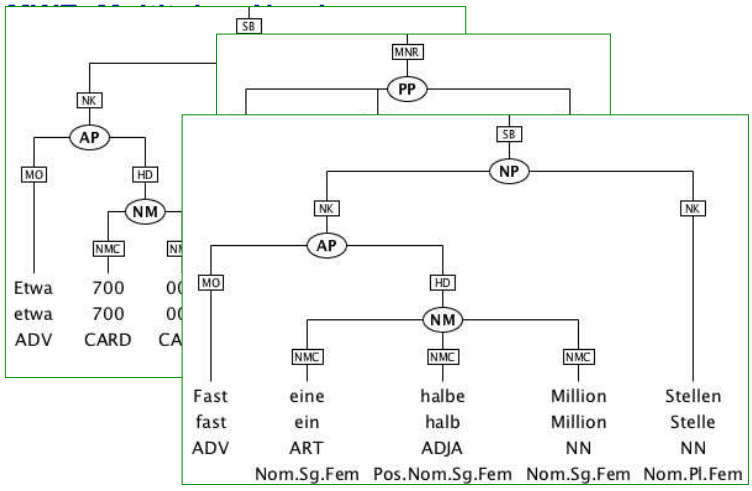
Institute of Computational Linguistics

MWE: FR Proper Names



University of Zurich[UZH]

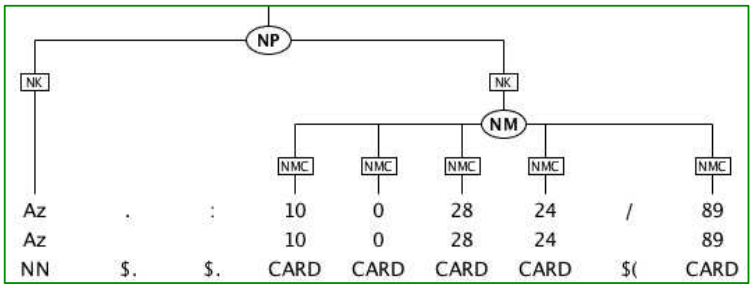
Institute of Computational Linguistics



University of Zurich[UZH]

Institute of Computational Linguistics

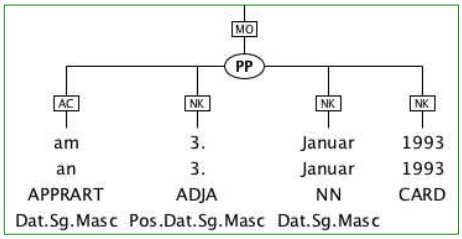
MWE: Multitoken Numbers

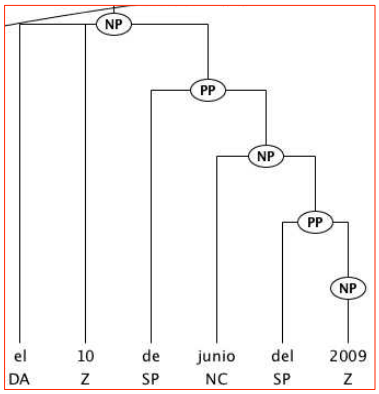


University of Zurich[UZH]

Institute of Computational Linguistics

MWE: Date Expressions

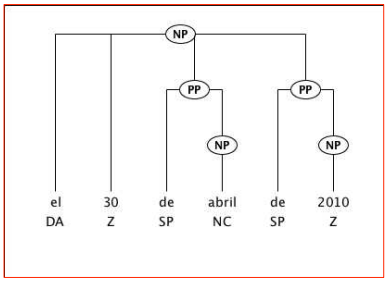


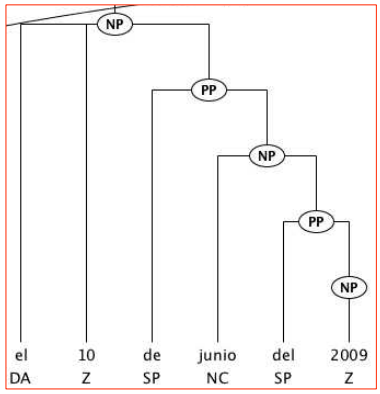


University of Zurich[UZH]

Institute of Computational Linguistics

MWE: Date Expressions

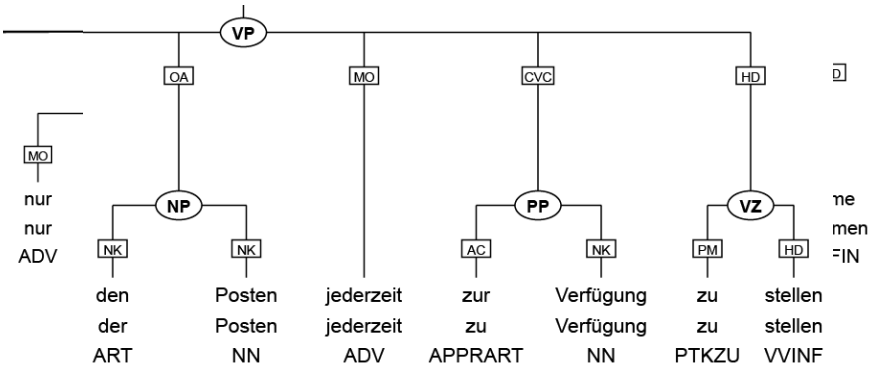




University of Zurich[UZH]

Institute of Computational Linguistics

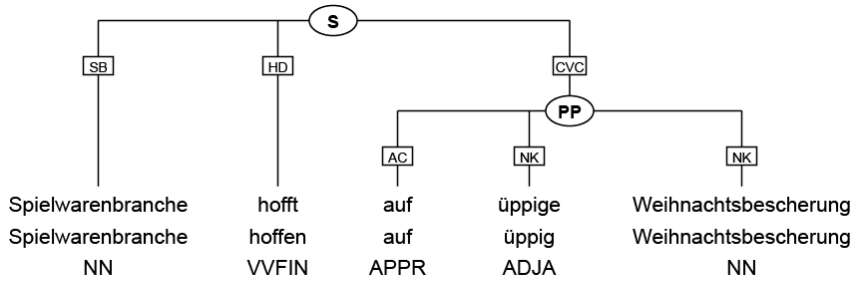
German Support Verb Constructions



University of Zurich[UZH]

Institute of Computational Linguistics

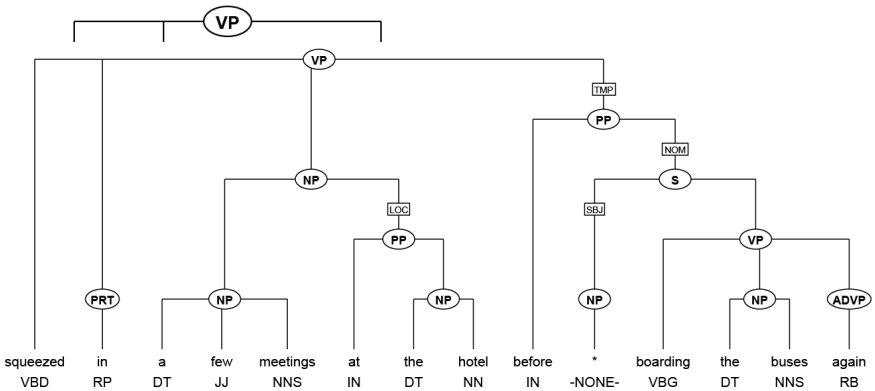
German Support Verb Constructions with ADJ ???



University of Zurich[UZH]

Institute of Computational Linguistics

English Particle Verbs



University of Zurich[UZH]

Institute of Computational Linguistics

English Particle Verbs

… in the Penn Treebank (WSJ 0-12): 24.618 trees = 629'798 tokens

PoS = particle (= RP): 1621 occurrences

category = particle (= PRT): 1608 occurrences (= 6.5% of all trees)

→ out of which **212** Verb + PRT pairs are discontiguous (13%)

→ out of which 175 occur with 1 intervening token

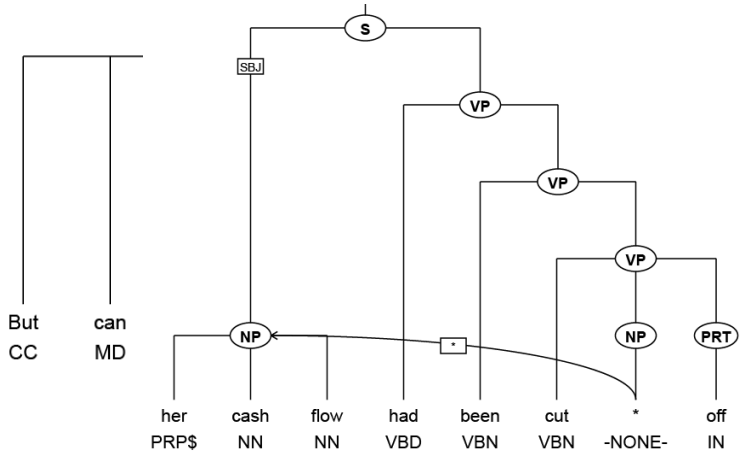
→ out of which 109 occur with the **empty** token intervening !!!

→ while the remaining 37 have two or more intervening tokens

University of Zurich[UZH]

Institute of Computational Linguistics

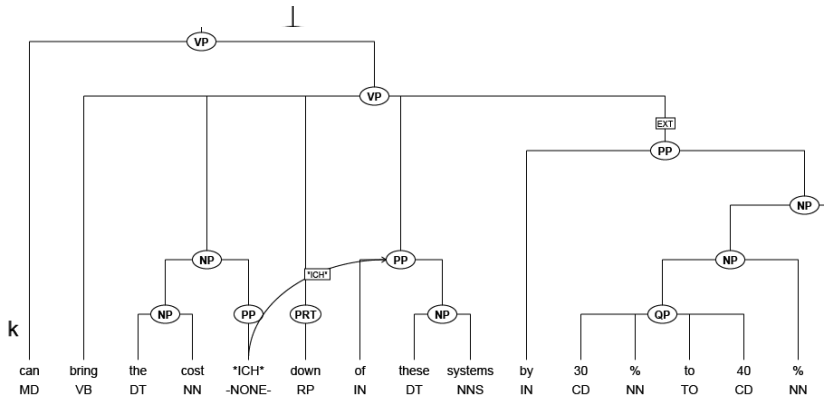
English Particle Verbs – Discontiguous with 1 Token



University of Zurich[UZH]

Institute of Computational Linguistics

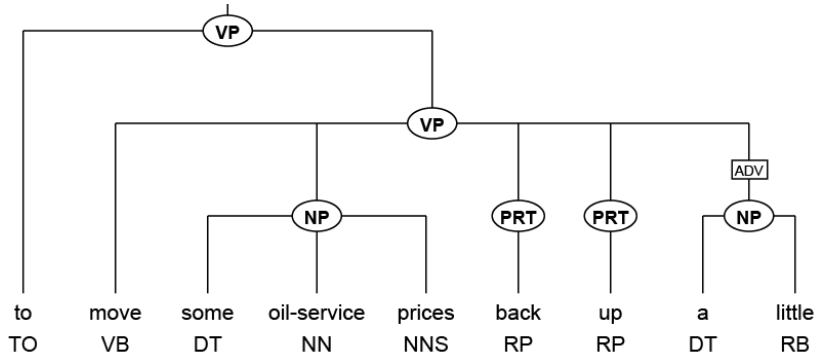
English Particle Verbs – Discontigous with >1 Tokens



University of Zurich[UZH]

Institute of Computational Linguistics

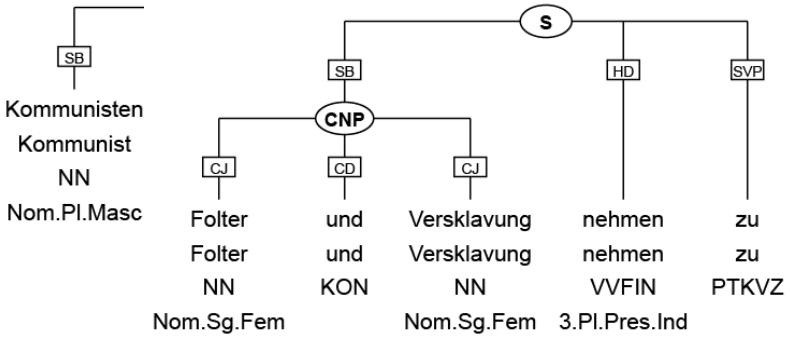
English Particle Verbs – Discontigous with 2 Particles



University of Zurich[UZH]

Institute of Computational Linguistics

German Verbs with Separated Prefixes



University of Zurich[UZH]

Institute of Computational Linguistics

German Verbs with Separable Prefixes

… in the TIGER Treebank: 50.474 trees = 888'299 tokens

PoS = separated verb prefix (= PTKVZ): 5074 occurrences in 4962 trees (= 9.8% of all trees)

→ out of which **355** Verb + PTKVZ pairs are adjacent (7%)

→ **550** occur with 1 intervening token

→ **632** occur with 2 intervening tokens

→ **502** occur with 3 intervening tokens

→ **2012** occur with 4-9 intervening tokens

→ while the remaining **853** have 10 or more intervening tokens

→ (max = 45 intervening tokens)

University of Zurich[UZH]

Institute of Computational Linguistics

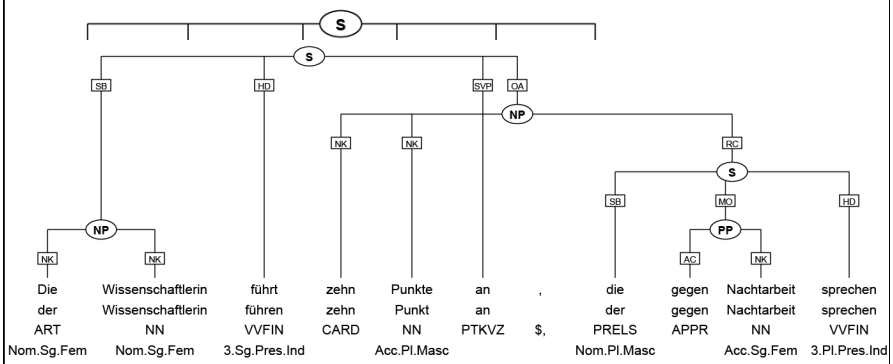
German Verb with Separable Prefixe (dist = 45)

Diese Staatsidee wurzelt in bürgerlich-demokratischen Emanzipationsvorstellungen und **weist** dem Staat eine aktive Rolle in der Steuerung wirtschaftlicher und gesellschaftlicher Abläufe in bezug auf eine – über den “ sozialstaatlichen Reparaturauftrag “ hinausweisende – Förderung einer größeren Gleichheit der Lebenschancen in bezug auf eine lebenslange Einkommenssicherung , Gesundheit , Kindererziehung , Wohnen und Bildung **zu**.

University of Zurich[UZH]

Institute of Computational Linguistics

German Verbs with Separated Prefixes



University of Zurich[UZH]

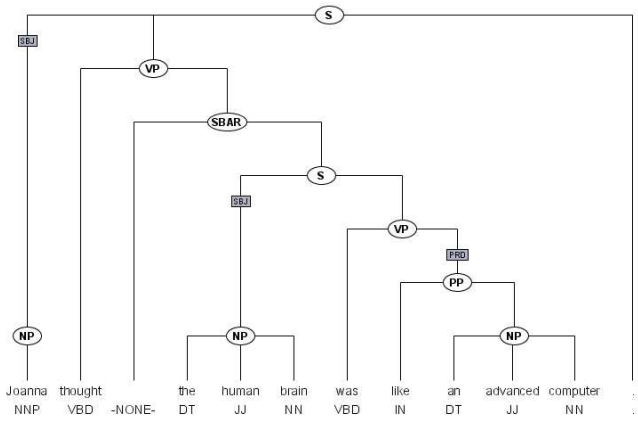
Institute of Computational Linguistics

Parallel Treebanks

University of Zurich[UZH]

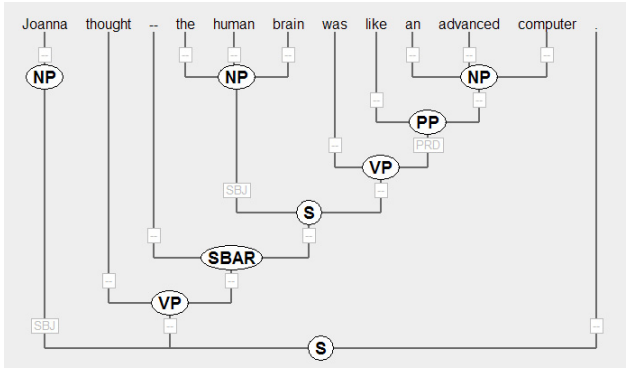
Institute of Computational Linguistics

English Syntax Tree

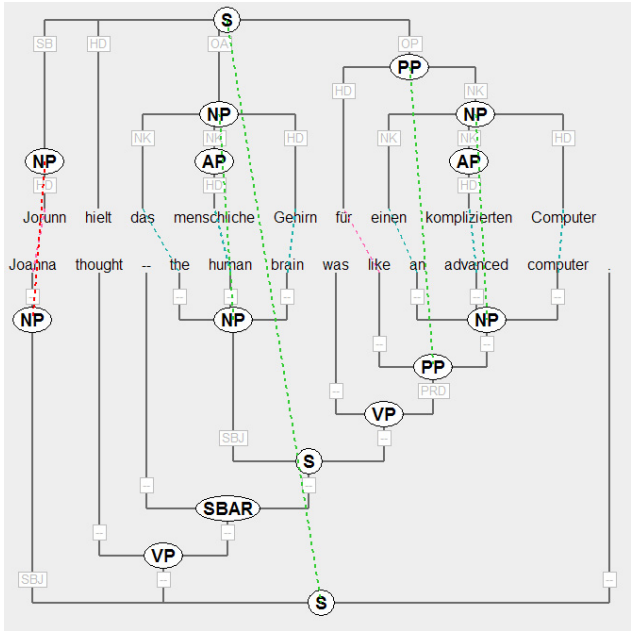


University of Zurich[UZH]

Institute of Computational Linguistics



DE – EN Alignment



SMULTRON

= Stockholm MULtilingual TReebank

- 1500 sentences in 3 languages (DE-EN-SV)
  - 500 from Jostein Gaarder's *Sophie's World*
  - 500 from Economy texts
    - ABB Quarterly report
    - Rainforest Alliance: Banana Certification Program
    - SEB Annual report
  - 500 from DVD player manual

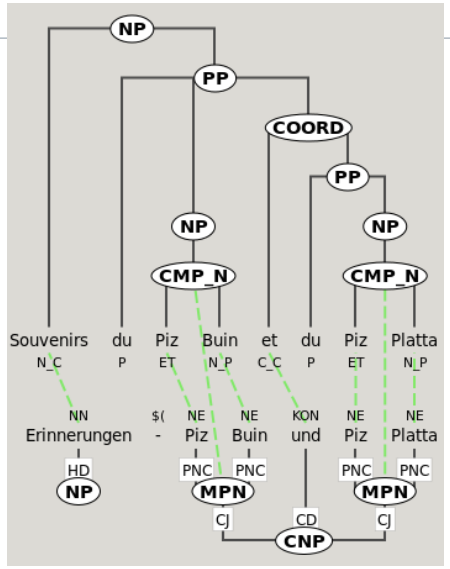
SMULTRON

- Plus: 1000 sentences in 2 languages (DE-FR)
  - Swiss Alpine Club mountaineering texts
- Plus: 4000 sentences in 2 languages (DE-ES)
  - agriculture, education, biography, etc.
- In preparation (ES - Quechua)
  - 500 trees for Quechua

University of Zurich[UZH]

Institute of Computational Linguistics

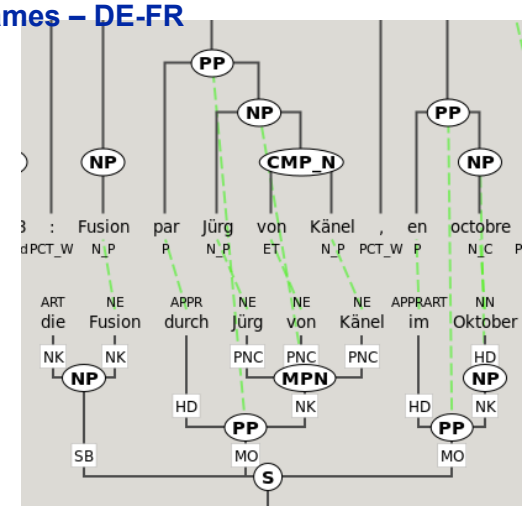
MWE: Proper Names – DE-FR

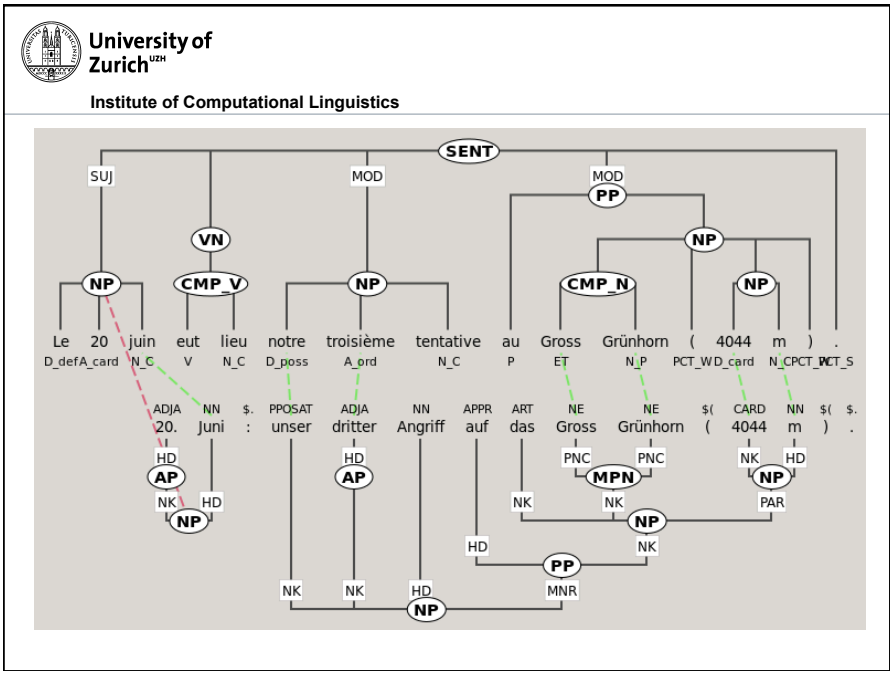


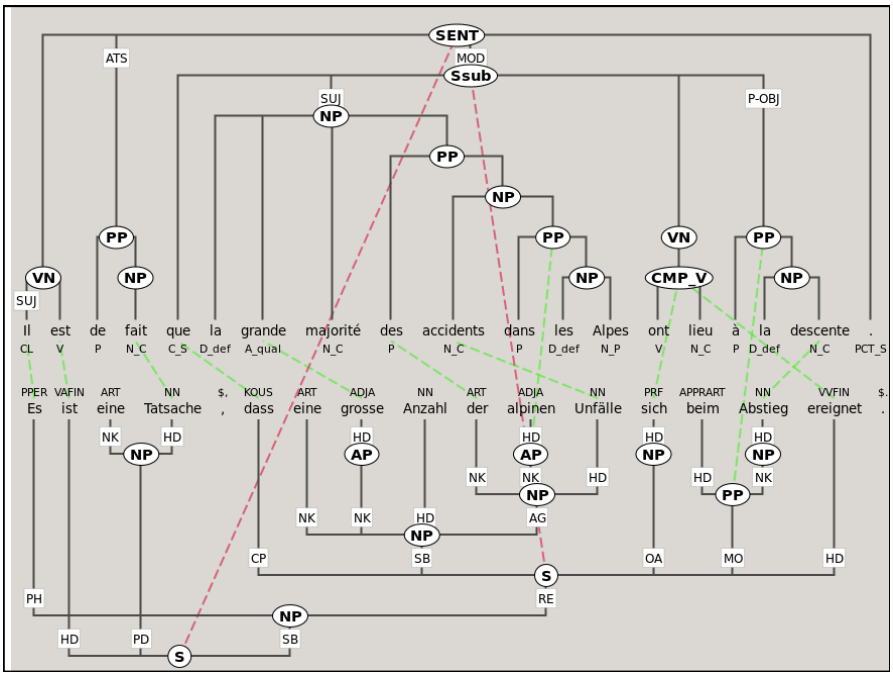
University of Zurich[UZH]

Institute of Computational Linguistics

MWE: Proper Names – DE-FR




University of Zurich[UZH]
Institute of Computational Linguistics
SENT
SUJ
MOD
MOD
PP
VN
NP
NP
CMP_V
NP
CMP_N
NP
Le 20 juin eut lieu notre troisième tentative au Gross Grünhorn (4044 m) .
D_def A_card N_C V N_C D_poss A_ord N_C P ET N_P PCT_W D_card N_C PCT_W PCT_S
ADJA NN $. PPOSAT ADJA NN APPR ART NE NE $(CARD NN $($.
20. Juni : unser dritter Angriff auf das Gross Grünhorn (4044 m) .
HD HD PNC PNC NK HD
AP AP MPN NP
NK HD NK HD NK NK PAR
NP NP
NP HD NK
PP
NK NK HD MNR
NP


SENT
ATS
MOD
Ssub
SUJ
P-OBJ
NP
PP
NP
PP
PP
VN
PP
VN
NP
NP
CMP_V
NP
SUJ
Il est de fait que la grande majorité des accidents dans les Alpes ont lieu à la descente .
CL V P N_C C_S D_def A_qual N_C P N_C P D_def N_P V N_C P D_def N_C PCT_S
PPER VAFIN ART NN $, KOUS ART ADJA NN ART ADJA NN PRF APPRART NN VVFIN $.
Es ist eine Tatsache , dass eine grosse Anzahl der alpinen Unfälle sich beim Abstieg ereignet .
NK HD HD HD HD HD HD
NP AP AP NP NP
NK NK NK HD HD NK
NP NP
NK NK HD AG
NP
CP SB OA MO HD
S
PH RE
NP
HD PD SB
S

University of Zurich[UZH]

Institute of Computational Linguistics

French MW prepositions

Given: The FR Le Monde Treebank with 20,500 trees

MW prepositions

consisting of 2 tokens: 1825 occurrences

près de, jusqu' à, plus de, lors de, afin de, …

consisting of 3 tokens: 1799 occurrences

par rapport à, en matière de, en raison de, …

consisting of 4 tokens: 1443 occurrences

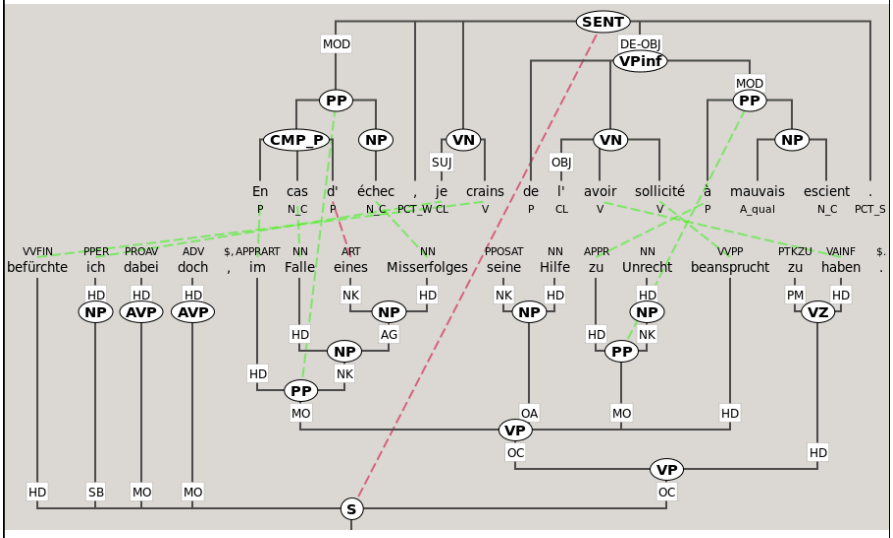
à la fin de, à la tête de, sous le nom de, au _ cours de, …

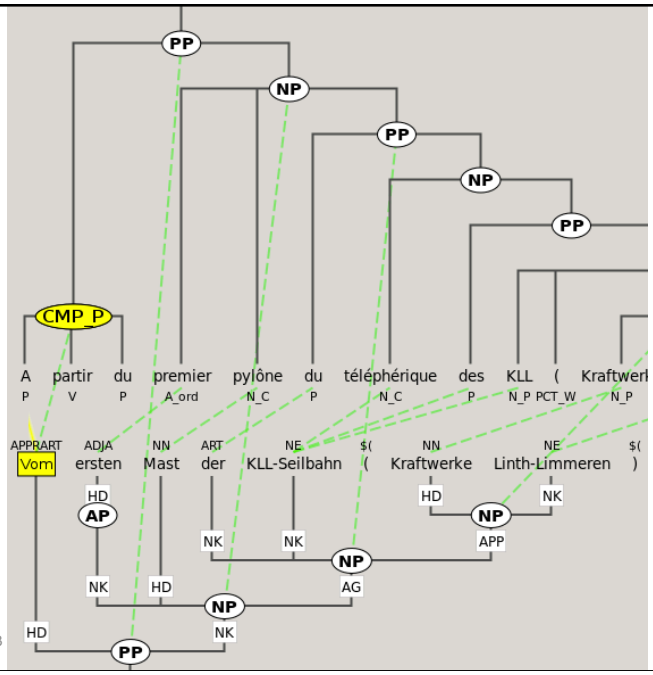
consisting of 5 tokens: 105 occurrences

de l' autre côté de, au _ – dessous de, …

consisting of 6 tokens: 74 occurrences

vis – à – vis de, de part et d' autre de, c' est - à - dire, …

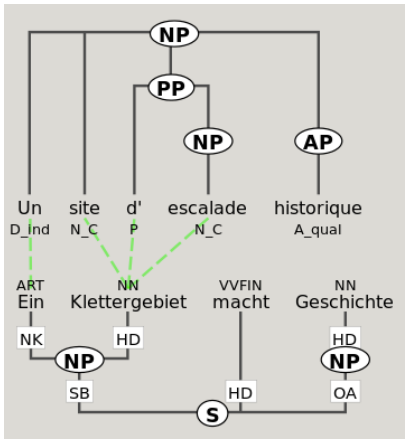




University of Zurich[UZH]

Institute of Computational Linguistics

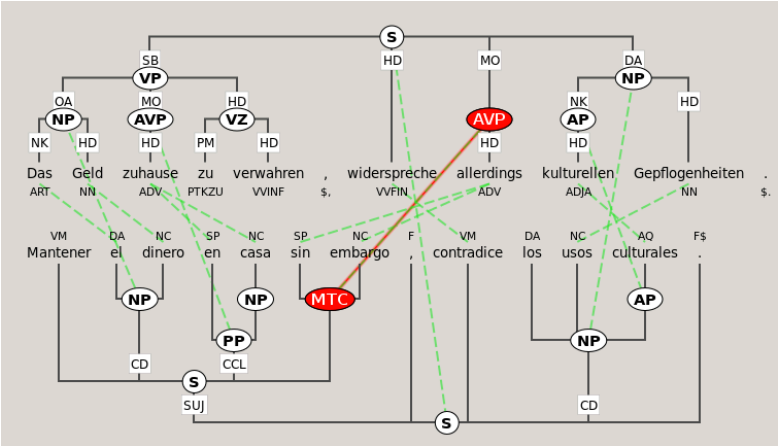
French compound nouns



University of Zurich[UZH]

Institute of Computational Linguistics

Spanish MW prepositions



University of Zurich[UZH]

Institute of Computational Linguistics

Summary

MWEs in Treebanks

- cover a wide range of phenomena
  - MW proper names
  - MW numbers
  - Date and time expressions
  - Support verb constructions
  - ES and FR MW prepositions
  - DE verbs with separated prefixes, EN particle verbs
- are handled very differently in treebanks across languages
- are under-researched

Suggestions for WG 4 Tasks

Perform **contrastive state-of-the-art survey**

… for as many treebanks as possible

Develop **recommendations** for

… retrospective annotation of MWEs in existing treebanks

… future annotation of MWEs in treebanks

… and other types of corpora

University of Zurich[UZH]
Institute of Computational Linguistics

Thank you!