Detecting Abstract Linguistic Properties through the Study of Corpus Data

Workshop on Historical Corpus Linguistics:
Methods and Applications

Anthony Kroch and Beatrice Santorini
University of Pennsylvania
June 2016

www.ling.upenn.edu/~kroch/handouts/saarland6-16.pdf

What is a morphosyntactically annotated corpus?

- morphological tagging
  case, gender, number features on nouns
  tense, mood, aspect features on verbs, etc.

- lemmatization
  word sense disambiguation
  spelling normalization

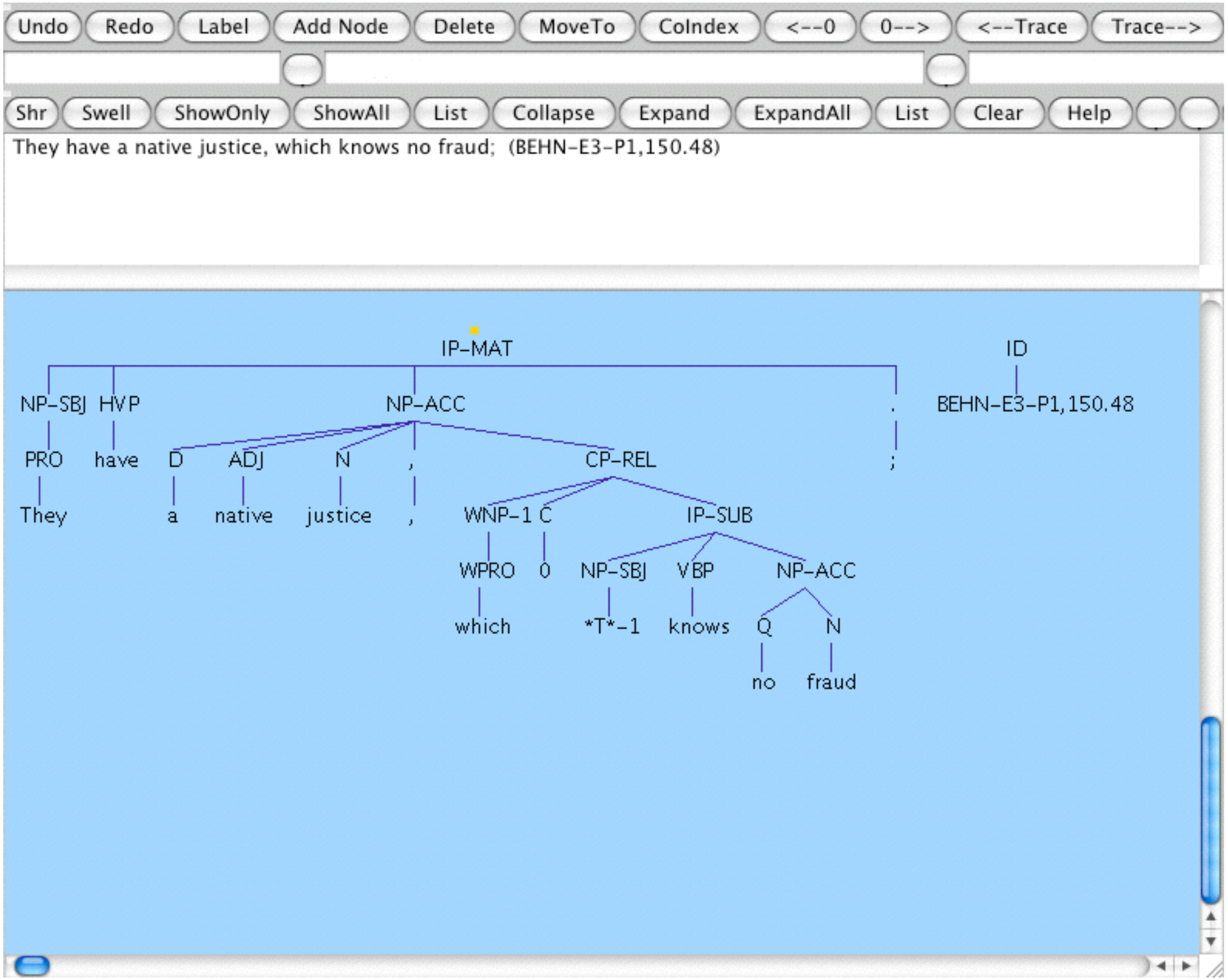
- part of speech tagging
  elementary syntactic functions

- syntactic parsing
  hierarchical structure of phrases/clauses
  grammatical function of phrases/clauses

An example sentence

```
((IP-MAT (NP-SBJ  (PRO They))
         (HVP have)
         (NP-ACC (D a)
                 (ADJ native)
                 (N justice)
                 (, ,)
                 (CP-REL (WNP-1 (WPRO which))
                         (C 0)
                         (IP-SUB  (NP-SBJ *T*-1)
                                  (VBP knows)
                                  (NP-ACC (Q no)
                                          (N fraud)))))
 (. ;))
(ID BEHN-E3-P1,150.48))
```

Legacy correction tool: CorpusDraw



The annotation task

- Annotation is multilevel and complex, so that using human effort for the whole job is impractical.
- At the same time, accuracy is crucial and unattainable at present with fully automated methods.
- In consequence, our parsed corpora are currently built by interleaving automated analysis with human correction of the output.

The future of annotated corpora

- Electronic text corpora of enormous size are becoming available; e.g. the Early English Books Online corpus, currently being digitized by the Text Creation Partnership, which will eventually amount to many billions of words.

- In consequence, methods of annotation accurate enough for research purposes that do not rely on human correction are badly needed.

- The development of such methods will depend on continuing advances in automated natural language processing.

Annotated Corpora Using the Penn Treebank Format

Parsed Corpora of Historical English

Together, these corpora total approximately
9.5 millon words of running text.

- York-Helsinki Parsed Corpus of Old English Poetry
- York-Toronto-Helsinki Parsed Corpus of Old English Prose
- Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English
- Penn-Helsinki Parsed Corpus of Middle English, 2nd edition (PPCME2)
- Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)
- York-Helsinki Parsed Corpus of Early English Correspondence (PCEEC)
- Penn Parsed Corpus of Modern British English, 2nd edition (PPCMBE2)

Other Languages

- Tycho Brahe Corpus, a parsed corpus of historical Portuguese
  - Charlotte Galves (University of Campinas, Brazil) and collaborators
- Modéliser le changement: les voies du français (Modelling change: the paths of French), a parsed corpus of historical French
  - France Martineau (University of Ottawa) and collaborators
- CORDIAL-SIN Corpus, a syntax-oriented corpus of Portuguese dialects
  - Ana Maria Martins (Centro de Linguística da Universidade de Lisboa) and collaborators
- Icelandic Parsed Historical Corpus (IcePaHC)
  - Eiríkur Rögnvaldsson (University of Iceland) and collaborators
- Word order and word order change in Western European languages (WOChWEL) Corpus, a growing parsed corpus of Old Portuguese
  - Ana Maria Martins and Sandra Pereira (Centro de Linguística da Universidade de Lisboa) and collaborators
- Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE), nearing completion
  - Christina Tortora (City University of New York) and collaborators
- P.S. Post Scriptum - A Digital Archive of Ordinary Writing (Early Modern Portugal and Spain), April 2017
  - Rita Marquilhas (Centro de Linguística da Universidade de Lisboa) and collaborators
- NINJAL Parsed Corpus of Modern Japanese (NPCMJ), under construction
  - Prashant Pardeshi (National Institute of Japanese Language and Linguistics) and collaborators

A Case Study: Detecting Stages in the transition from OV to VO

Data sources: English

- Anthony Kroch and Ann Taylor. 2000. Penn-Helsinki Parsed Corpus of Middle English, second edition (PPCME2).

Data sources: French

- France Martineau et al. 2005. Corpus du projet Modéliser le changement: les voies du français (MCVF).

- Anthony Kroch and Beatrice Santorini. 2016. Penn supplement to the MCVF corpus.

- Alexei Lavrentiev, Christiane Marchello-Nizia, Céline Guillot and Serge Heiden. 2014. BFM - Base de Français Médiéval [En ligne].

Data sources: Icelandic

- Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson and Eiríkur Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC).

Data sources: Yiddish

- Beatrice Santorini. 2008. Penn Yiddish Corpus.

Preliminaries

- Only non-finite VPs are considered to avoid interference from V-to-C and V-to-T.

- Modals are treated as auxiliary verbs in all three languages with reported data.

- Sentences in which the direct object moves further left than T are also excluded since the “in situ” position is not recoverable.

English

Evidence for VO word order in Early Middle English

(1) oðet he habe iȝetted ou al þet ȝe wulleð
'until he has granted you all that you want'
(CMANCRIW,I.68.229)

(2) þt he schulde in huden him ȝef he walde libben
'that he should hide himself if he would live'
(CMANCRIW,II.132.1744)

More evidence for VO word order in Early Middle English

(1) hwaso mei gan in
'whoever may go in'
(CMANCRIW,II.60.5)

(2) ha wes sone ibroht forð
'she was soon brought forth'
(CMKATHE, B.827)

More evidence for VO word order in Early Middle English

(1) worþy mennes sones þat sche myȝte han be maried to
'worthy men's sons that she might have been married to'
(CMAELR3-M23,33.189)

(2) þe terme, þe which hij ne shul nouȝt passe over
'the limit which he should not pass over'
(CMEARLPS-M2,125.5471)

Possible evidence for OV word order in Early Middle English

(1) þeos ne schulen neaver song singen song

'these should never sing songs'
(CMHALI,142.222)

(2) þat ne have noht here sinnes forleten here sinnes

'who have not there sinnes forsaken.'
(CMTRINIT,67.934)

More evidence for OV word order in Early Middle English

(1) al þe blodi sunnen þet ha is wið iwundet
'all the blodi sunnen that she is wounded with'
(CMANCRIW,I.62.202)

(2) sumþing þet god maȝe of arisen
'something that good may arise from'
(CMANCRIW,I.74.296)

Two-argument VPs: OOV word order

(1) Ne durste nauere gume nan oðerne ufele igreten
'Nor did a man ever dare to afflict evil on another'
(1200-BRUT,564.1322)

(2) þatt icc have ȝuw summ del nu spelledd offe
'which I have told you something of '
(CMORM-M1,1,221.1820)

Two-argument VPs: OVO word order

(1) For all þeo the habbeð any good idon me
'For all those who have done me any good'
(CMANCRIW,I.64.212)

(2) I sal yu lere þe dute of god
'I shall teach you the fear of God'
(CMBENRUL-M3,2.20)

Two-argument VPs: VOO word order

(1) oðet he habe iȝetted ou al þet ȝe wulleð
'until he has granted you all that you want'
(CMANCRIW,I.68.229)

(2) and wile ȝelden eche men his mede efter his werke
'and will pay each man his reward by his work'
(CMLAMB1-M1,143.310)

Distribution of Full DP Objects in Double Object Clauses in Early Middle English (<1420)

	IO>V	**V>IO**	**rate of IO scrambling**
DO>V	1	1	
V>DO	2	31	0.06
rate of DO scrambling		0.03	N=35

Chi-square: .731 (ns)

Expected rate of OOV based on rates of IO and DO scrambling	.03 x .06 = **0.002**
Actual rate of OOV	1/35 = **0.03**

Distribution of All Objects in Double Object Clauses in Early Middle English (<1420)

	IO>V	**V>IO**	**rate of IO scrambling**
DO>V	20	11	
V>DO	30	260	0.1034
rate of DO scrambling		0.0037	N=321

Chi-square: 62.498

Expected rate of OOV based on rates of IO and DO scrambling	.0037 **x** .1034 = **0.00038**
Actual rate of OOV	20/321 = **0.06**

French

VO & OV word order: modal + infinitive

(1) Je veul avoir mon loier
'I want to have my pay.'
(127X-CASSIDORUS-P,164.1546)

(2) Kar ne poeit le jur choisir le jur

'For he cannot choose the day.'
(116X-MARIE-DE-FRANCE-R,111.2262)

VO & OV word order: *avoir* + participle

(1) Rollant ad mis l' olifan a sa buche
'Roland raised the ivory horn to his mouth.'
(1100-ROLAND-V,133.1772)

(2) Li reis Marsilie out sun cunseill finet sun cunseill

'King Marsilla had adjourned his council.'
(1100-ROLAND-V,5.53)

Two-argument VPs: OOV word order

(1) Or ad Deus saint Thomas cel' ampole donee
'Now God gave Saint Thomas this phial'
(1173-becket-p-bfm,182.14984)

(2) ainsi pourroit Grace a Dieu querre
'In this way, he could ask God for grace'
(1190-BORON-R-PENN,7.88)

Two-argument VPs: OVO word order

(1) Tu auoiz dous choses amises al creator
'You had presented two things to the creator'
(1190-SBERNAN-P-BFM,10.325)

(2) Ancor uolt plus grant honor faire a nostre lum
'He wished to do our man an even great honor'
(1190-SBERNAN-P-BFM,37.1192)

Two-argument VPs: VOO word order

(1) Et Pilates a douné le cors Joseph
'and Pilate gave the body to Joseph'
(1210-BORON-P-PENN,24.230)

(2) É Deu ad dune le regne a Absalon tun fils
'and God has given the kingdom
to your son Absalom'
(1150-QUATRELIVRE-P-PENN,88.3317)

Distribution of Objects in Double Object Clauses in Early Old French (<1260)

	IO>V	**V>IO**	**rate of IO scrambling**
DO>V	11	6	
V>DO	9	55	0.14
rate of DO scrambling		0.10	N=81

Chi-square: 18.52

Expected rate of OOV based on rates of IO and DO scrambling	.14 x .10 = **0.014**
Actual rate of OOV	11/81 = **0.14**

Distribution of Objects in Double Object Clauses in late Old French (<1460)

	IO>V	**V>IO**	**rate of IO scrambling**
DO>V	2	17	
V>DO	31	176	0.15
rate of DO scrambling		0.09	N=226

Chi-square: .276

Expected rate of OOV based on rates of IO and DO scrambling	.15 x .09 = **0.013**
Actual rate of OOV	2/226 = **0.01**

Yiddish

VO & OV word order: modal+infinitive

(1) da velin mir vermisin di khasene
'Then we will ruin the wedding.'
(1615E-COURT,108.80)

(2) …ver nur kan zayn gezind farshiken zayn gezind

'who ever can send away his servants'
(1619W-LETTERS,.16)

VO & OV word order: *avoir*+participle

(1) …vau min hobn fergebin unzi zind
'where they have forgiven our sins'
(1704E-ELLUSH,.16)

(2) di hbn eyn yudn drmurt eyn yudn
?
'They murdered a Jew.'
(1465W-COURT,16.67)

Two argument VPs: OOV word order

(1) ikh hab den isral eyn tubh gtan
'I have done the Israelites a good turn'
(1579E-SHIR,10.60)

(2) un mustn imrdarn dem mtsraim ir fikh hitn
'and always had to guard the animals
for the Egyptians'
(1589E-ESTER,7.123)

Two argument VPs: OVO word order

(1) sukhr habn unzri bridr gigebn fil gelt
'Merchants gave our brothers much money'
(1692E-VILNA,217.134)

(2) drum hat er dem menshn gebn di turh …
'therefore has he the people given the Torah'
(1620E-LEVTOV1,41.47)

Two argument VPs: VOO word order

(1) hat gibrakht meyn oybrstn alirley shpetsirey
'[who] brought my boss all kinds of spices'
(1665W-COURT,221.246)

(2) mer haben unzer formuner gegeben meinem
stieffater tsvay hundert gulden
'our guardians gave my stepfather 200 guilders'
(1518W-GOETZ,.137)

Distribution of Objects in Double Object Clauses in early East Yiddish (<1800)

	IO>V	**V>IO**	**rate of IO scrambling**
DO>V	24	4	
V>DO	5	3	0.62
rate of DO scrambling		0.57	N=36

Chi-square: 2.14

Expected rate of OOV based on rates of IO and DO scrambling	.57 x .62 = **0.357**
Actual rate of OOV	24/36 = **0.667**

Distribution of Objects in Double Object Clauses in pre-contemporary East Yiddish (<1900)

	IO>V	**V>IO**	**rate of IO scrambling**
DO>V	10	1	
V>DO	3	8	0.27
rate of DO scrambling		0.11	N=22

Chi-square: 9.1

Expected rate of OOV based on rates of IO and DO scrambling	.11 x .27 = **0.030**
Actual rate of OOV	10/22 = **0.454**

Distribution of Objects in Double Object Clauses in contemporary East Yiddish (>1900)

	IO>V	**V>IO**	**rate of IO scrambling**
DO>V	2	6	
V>DO	6	25	0.19
rate of DO scrambling		0.19	N=39

Chi-square: .124

Expected rate of OOV based on rates of IO and DO scrambling	.19 x .19 = **0.037**
Actual rate of OOV	2/39 = **0.051**

Icelandic

Distribution of Objects in Double Object Clauses in pre-contemporary Icelandic (<1900)

	IO>V	**V>IO**	**rate of IO scrambling**
DO>V	41	27	
V>DO	8	192	0.04
rate of DO scrambling		0.12	N=268

Chi-square: 107.6

Expected rate of OOV based on rates of IO and DO scrambling	.12 x .04 = **0.005**
Actual rate of OOV	41/268 = **0.152**

Distribution of Objects in Double Object Clauses in contemporary Icelandic (>1900)

	IO>V	**V>IO**	**rate of IO scrambling**
DO>V	0	3	
V>DO	2	47	0.04
rate of DO scrambling		0.06	N=52

Chi-square: .127

Expected rate of OOV based on rates of IO and DO scrambling	.06 x .04 = **0.002**
Actual rate of OOV	0/52 = **0.000**

Finis