New Approaches to Syntactic Annotation: Constructing a Parsed Corpus of early High German

Workshop on Databases and Corpora in Linguistics
Stony Brook University

Richard Zimmermann
University of Geneva

October 17, 2014

1 The Geneva Corpus of early German

1.1 Overview

- The *Geneva Corpus of early German* (GeCeG) is currently under construction on a one-year graduate grant under the scholarly supervision of experts at the University of Pennsylvania
- a fully annotated and syntactically parsed corpus of Old and early Middle High German (800-1200 A.D.)
- searchable with CorpusSearch (Randell, 2004)
- a relatively small corpus, comprising c. 30 – 50,000 words
- for the foreseeable future, it will be the only available fully parsed resource for the earliest stages of the High German language

1.2 Aims of the talk

- an opportunity to experiment with annotation systems; new manual guided by considerations of user-friendliness and conformity to linguistic theory
- focus on syntactic annotation, not POS-tagging
- contrast with corpora PPCME (Middle English) and YCOE (Old English)
- discuss three general areas of parsing innovations: (i) the Headedness Principle, (ii) coping with direct speech and disfluencies, (iii) the `TAG` system

2 The Headedness Principle

2.1 Definition and Illustration

- a well-formedness conditions on syntactic structures

- inspired by representational theories of syntax like HSPG, LFG

- terminal = a <POS, form> pair (e.g. <N, house>)

- functional node = a labelled vertice without a form (e.g. subject)

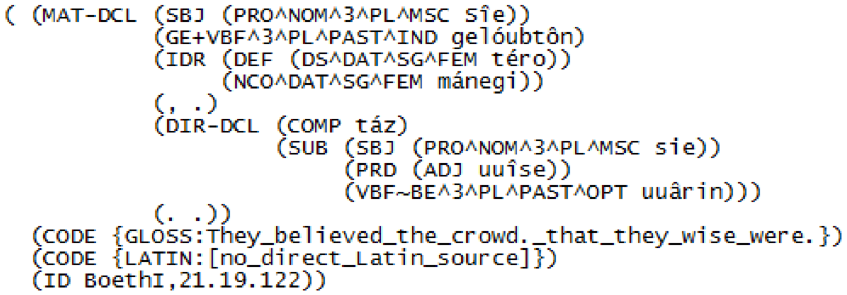
(1) **The Headedness Principle**

a. Completeness: Every functional node must immediately dominate a terminal. (There is a head)

b. Uniqueness: Every sister of a terminal must be a functional node. (The head is unique)

- heads of clauses are typically lexical verbs; other heads are more variable

(2)



```
( (MAT-DCL (SBJ (PRO^NOM^3^PL^MSC sîe))
           (GE+VBF^3^PL^PAST^IND gelóubtôn)
           (IDR (DEF (DS^DAT^SG^FEM téro))
                (NCO^DAT^SG^FEM mánegi))
           (, .)
           (DIR-DCL (COMP táz)
                    (SUB (SBJ (PRO^NOM^3^PL^MSC sie))
                         (PRD (ADJ uuîse))
                         (VBF~BE^3^PL^PAST^OPT uuârin)))
           (. .))
  (CODE {GLOSS:They_believed_the_crowd._that_they_wise_were.})
  (CODE {LATIN:[no_direct_Latin_source]})
  (ID BoethI,21.19.122))
```

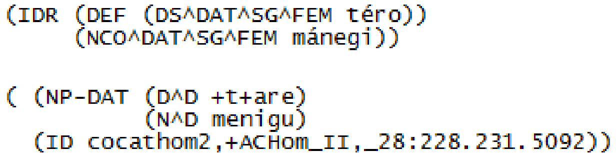
2.2 Advantages

2.2.1 Category Forms are redundant

- the Headedness Principle ensures that the form of a category (e.g. `NP`, `PP`, `ADJP`, etc.) can be unambiguously recovered from the head.

- e.g. contrast above *téro mánegi* (GeCeG) with Old English *þære menigu* ‘to the crowd’ (YCOE)

(3)



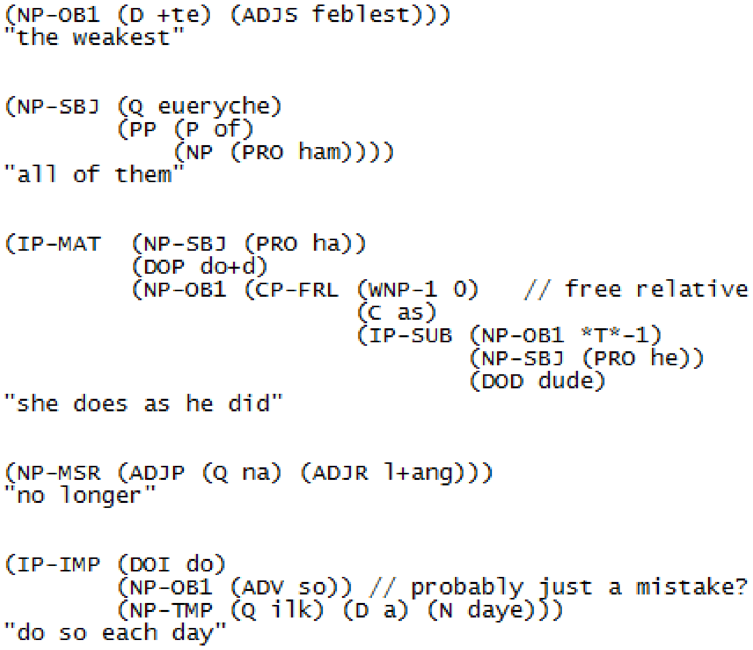
```
(IDR (DEF (DS^DAT^SG^FEM téro))
     (NCO^DAT^SG^FEM mánegi))


( (NP-DAT (D^D +t+are)
          (N^D menigu)
  (ID cocathom2,+ACHom_II,_28:228.231.5092))
```

- in the GeCeG, the `NP` category would be redundant because it follows directly from the fact that the head is a common noun

- in the GeCeG, the -DAT extension would be redundant because case is aready marked on the parts of speech of the category
- but the YCOE must mark NP because the head of the grammatical function is not identified unequivocally

- the meaning of category forms like NP is not always clear in early English corpora. They are simply convenient summaries of arguments and adjuncts that “feel” nominal
- e.g. in what sense are the following phases from the PPCME NPs?

(4)

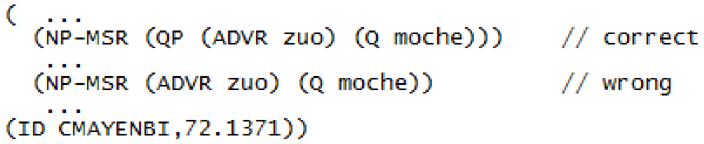


```
(NP-OB1 (D +te) (ADJS feblest)))
"the weakest"


(NP-SBJ (Q eueryche)
        (PP (P of)
            (NP (PRO ham))))
"all of them"


(IP-MAT  (NP-SBJ (PRO ha))
         (DOP do+d)
         (NP-OB1 (CP-FRL (WNP-1 0)   // free relative
                         (C as)
                         (IP-SUB (NP-OB1 *T*-1)
                                 (NP-SBJ (PRO he))
                                 (DOD dude)
"she does as he did"


(NP-MSR (ADJP (Q na) (ADJR l+ang)))
"no longer"


(IP-IMP (DOI do)
        (NP-OB1 (ADV so)) // probably just a mistake?
        (NP-TMP (Q ilk) (D a) (N daye)))
"do so each day"
```

2.2.2 Higher consistency

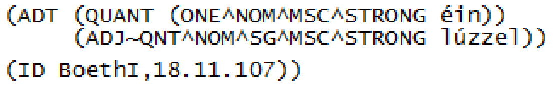
- the Headedness Principle makes the annotation process less error-prone
- e.g. modified quantifiers should form QPs in PPCME according to its manual; if nothing else occurs in the grammatical function, there is no obvious head; annotators sometimes prefer to have a head putting the quantifier immediately under the grammatical function producing a mistake

(5)



```
(  ...
  (NP-MSR (QP (ADVR zuo) (Q moche)))    // correct
   ...
  (NP-MSR (ADVR zuo) (Q moche))         // wrong
   ...
(ID CMAYENBI,72.1371))
```

- such a mistake is much harder to make in the GeCeG; e.g. the phrase *ein luzzel* ‘a little, somewhat’ functions as an adjunct on the clausal level, `ADT`, and requires a head by the Headedness Principle; no headless phrases are possible; easy to check automatically

(6) 

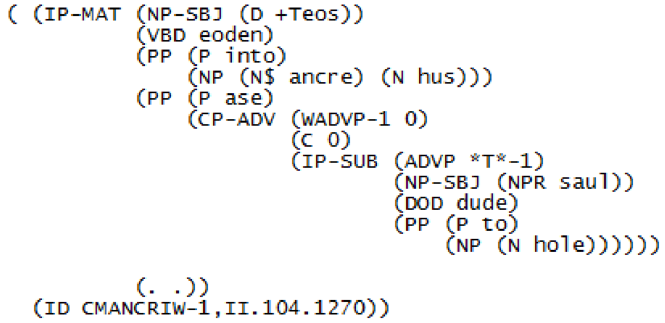
```
(ADT (QUANT (ONE^NOM^MSC^STRONG éin))
     (ADJ~QNT^NOM^SG^MSC^STRONG lúzzel))
(ID BoethI,18.11.107))
```

- similar problems are avoided elsewhere (e.g. PPCME measure phrases with adjectives etc.)

2.2.3 More information

- the Headedness Principle forces a more explicit functional annotation

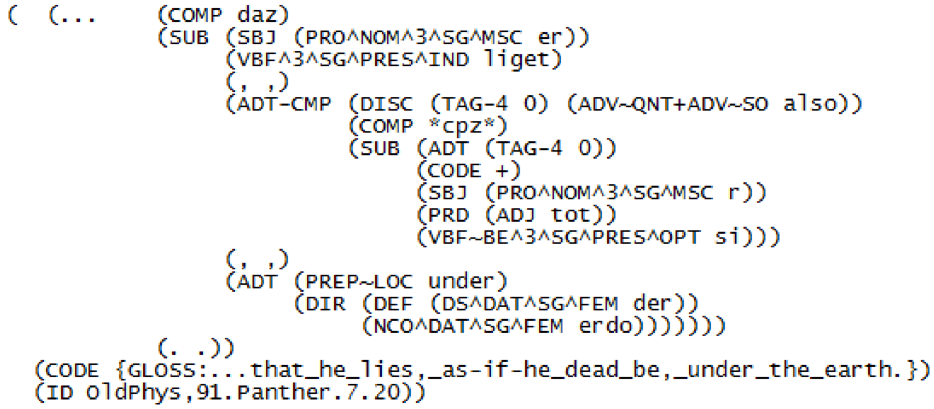
- e.g. the PPCME annotates a large variety of functions as `PP`s, including adjuncts, complements, subordinate clauses etc.

(7) 

```
( (IP-MAT (NP-SBJ (D +Teos))
          (VBD eoden)
          (PP (P into)
              (NP (N$ ancre) (N hus)))
          (PP (P ase)
              (CP-ADV (WADVP-1 0)
                      (C 0)
                      (IP-SUB (ADVP *T*-1)
                              (NP-SBJ (NPR saul))
                              (DOD dude)
                              (PP (P to)
                                  (NP (N hole))))))

          (. .))
  (ID CMANCRIW-1,II.104.1270))
```

- the GeCeG requires more detailed guidelines regarding argument structures; you need to annotate `PP`s explicitly as adjuncts, complements, predicates etc.

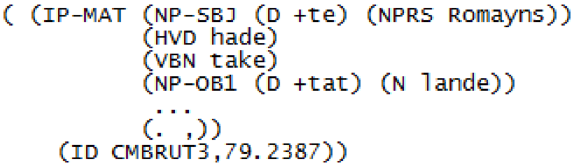
- subordinate clauses are basic functions with an additional extension, e.g. `ADT-CMP` is an adjunct, that is complex, i.e. clausal, namely comparative

(8) 

```
( (...     (COMP daz)
           (SUB (SBJ (PRO^NOM^3^SG^MSC er))
                (VBF^3^SG^PRES^IND liget)
                (, ,)
                (ADT-CMP (DISC (TAG-4 0) (ADV~QNT+ADV~SO also))
                         (COMP *cpz*)
                         (SUB (ADT (TAG-4 0))
                              (CODE +)
                              (SBJ (PRO^NOM^3^SG^MSC r))
                              (PRD (ADJ tot))
                              (VBF~BE^3^SG^PRES^OPT si)))
                (, ,)
                (ADT (PREP~LOC under)
                     (DIR (DEF (DS^DAT^SG^FEM der))
                          (NCO^DAT^SG^FEM erdo)))))))
           (. .))
  (CODE {GLOSS:...that_he_lies,_as-if-he_dead_be,_under_the_earth.})
  (ID OldPhys,91.Panther.7.20))
```

- e.g. auxiliary verbs; PPCME and YCOE annotate main and auxiliary verbs as daughters of the clause

(9)

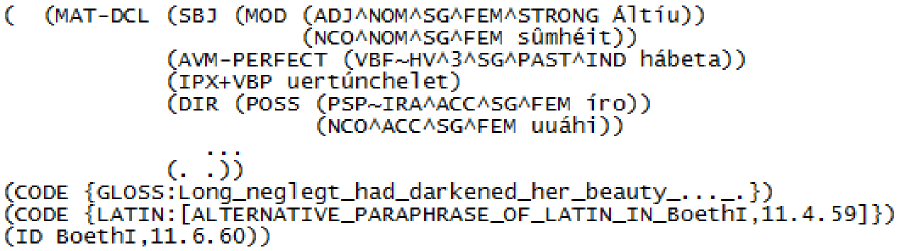


```
( (IP-MAT (NP-SBJ (D +te) (NPRS Romayns))
          (HVD hade)
          (VBN take)
          (NP-OB1 (D +tat) (N lande))
           ...
          (. ,))
    (ID CMBRUT3,79.2387))
```

- this would violate the Headedness Principle in the GeCeG

- special function called AVM for aspect, voice, modality, which hosts auxiliary verbs; subtypes like perfect, passive, modal

(10)



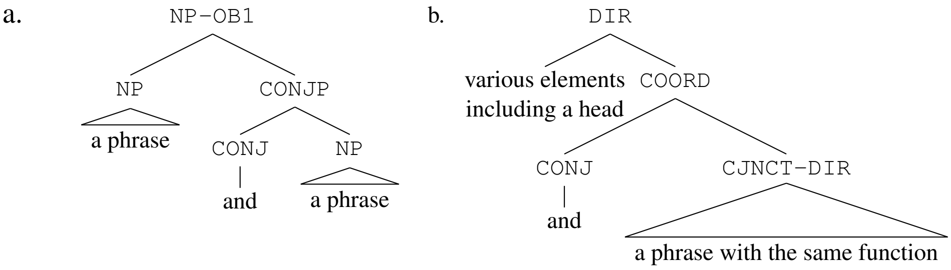
```
(  (MAT-DCL (SBJ (MOD (ADJ^NOM^SG^FEM^STRONG Áltíu))
                      (NCO^NOM^SG^FEM sûmhéit))
            (AVM-PERFECT (VBF~HV^3^SG^PAST^IND hábeta))
            (IPX+VBP uertúnchelet)
            (DIR (POSS (PSP~IRA^ACC^SG^FEM íro))
                       (NCO^ACC^SG^FEM uuáhi))
                  ...
            (. .))
(CODE {GLOSS:Long_neglegt_had_darkened_her_beauty_..._.})
(CODE {LATIN:[ALTERNATIVE_PARAPHRASE_OF_LATIN_IN_BoethI,11.4.59]})
(ID BoethI,11.6.60))
```

- similar annotation for negation

- the Headedness Principle forces a new annotation on coordinate structures; in the PPCME CONJP is sister of a phrase; the functional node does not have a head; in the GeCeG COORD is a sister of a the head of the functional node

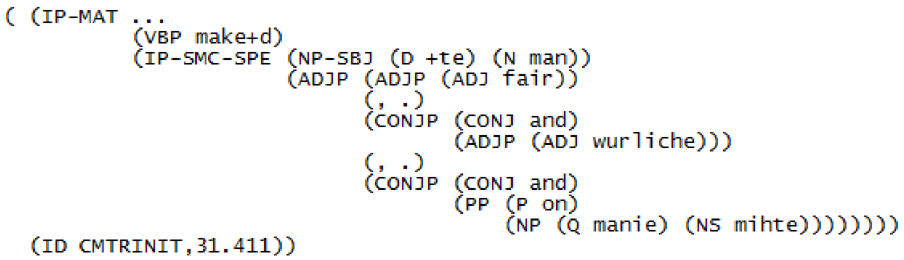
- prefix CJNCT on conjuncts

(11)



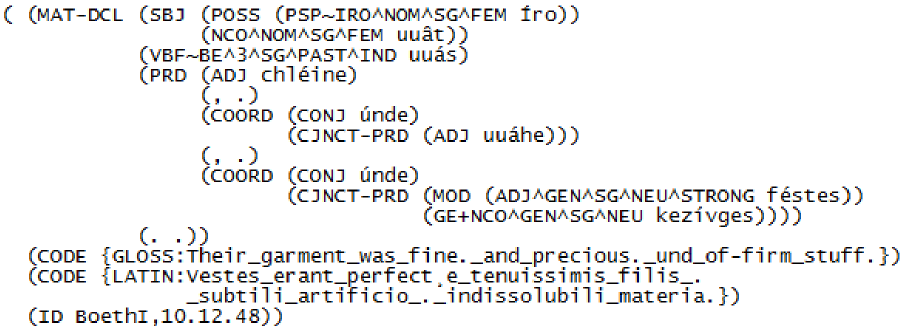
- this annotation scheme captures the idea that the function of conjuncts are identical while their forms may vary

(12)



```
( (IP-MAT ...
          (VBP make+d)
          (IP-SMC-SPE (NP-SBJ (D +te) (N man))
                      (ADJP (ADJP (ADJ fair))
                            (, .)
                            (CONJP (CONJ and)
                                   (ADJP (ADJ wurliche)))
                            (, .)
                            (CONJP (CONJ and)
                                   (PP (P on)
                                       (NP (Q manie) (NS mihte)))))))
  (ID CMTRINIT,31.411))
```

(13)



```
( (MAT-DCL (SBJ (POSS (PSP~IRO^NOM^SG^FEM íro))
               (NCO^NOM^SG^FEM uuât))
          (VBF~BE^3^SG^PAST^IND uuás)
          (PRD (ADJ chléine)
               (, .)
               (COORD (CONJ únde)
                      (CJNCT-PRD (ADJ uuáhe)))
               (, .)
               (COORD (CONJ únde)
                      (CJNCT-PRD (MOD (ADJ^GEN^SG^NEU^STRONG féstes))
                                 (GE+NCO^GEN^SG^NEU kezívges))))
          (. .))
  (CODE {GLOSS:Their_garment_was_fine._and_precious._und_of-firm_stuff.})
  (CODE {LATIN:Vestes_erant_perfect,e_tenuissimis_filis_.
              _subtili_artificio_._indissolubili_materia.})
  (ID BoethI,10.12.48))
```

2.2.4 Easier search scripting

- in general, the Headedness Principle provides great control over the structures a researcher might be interested in (e.g. easier to find arguments that are headed by a determiner vs. arguments that have a definite article and many other cases)

- the command `idoms` "immediately dominates" automatically identifies heads in the GeCeG

- more explicit annotations allow searching for specific constructions easily (e.g. just look for `AVM-PASSIVE` for all passive sentences) (e.g. `NEGAT` for all negations, but look for specific kinds of negations by form)

- coordinate structures are easier to search (directly look for conjuncts, no need for search command `idomsmod`)

3 Direct Speech and Parentheticals

3.1 Direct Speech Overview

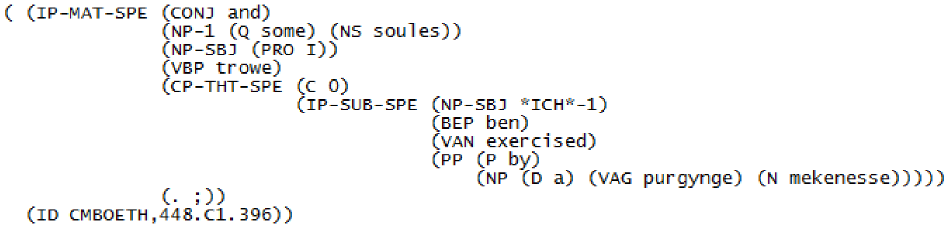
- direct speech is indicated with extension -SPE on clauses (IPs and CPs) in PPCME and YCOE

- result: complex labels such as IP-MAT-PRN-SPE

- arbitrary; why not on other phrases, fragments, foreign language?

- order of labels not always clear

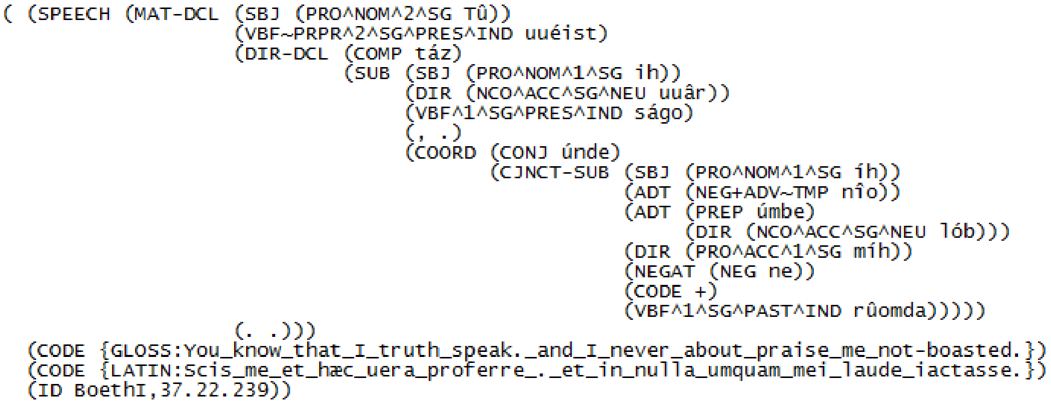
(14)



```
( (IP-MAT-SPE (CONJ and)
              (NP-1 (Q some) (NS soules))
              (NP-SBJ (PRO I))
              (VBP trowe)
              (CP-THT-SPE (C 0)
                          (IP-SUB-SPE (NP-SBJ *ICH*-1)
                                      (BEP ben)
                                      (VAN exercised)
                                      (PP (P by)
                                          (NP (D a) (VAG purgynge) (N mekenesse)))))
              (. ;))
  (ID CMBOETH,448.C1.396))
```

- the GeCeG separates out information on direct speech and parentheticals from the major clausal nodes; initial node called SPEECH; not regarded as a functional category; simply indicates scope of Speech

(15)



```
( (SPEECH (MAT-DCL (SBJ (PRO^NOM^2^SG Tû))
                   (VBF~PRPR^2^SG^PRES^IND uuéist)
                   (DIR-DCL (COMP táz)
                            (SUB (SBJ (PRO^NOM^1^SG ih))
                                 (DIR (NCO^ACC^SG^NEU uuâr))
                                 (VBF^1^SG^PRES^IND ságo)
                                 (, .)
                                 (COORD (CONJ únde)
                                        (CJNCT-SUB (SBJ (PRO^NOM^1^SG íh))
                                                   (ADT (NEG+ADV~TMP nîo))
                                                   (ADT (PREP úmbe)
                                                        (DIR (NCO^ACC^SG^NEU lób)))
                                                   (DIR (PRO^ACC^1^SG míh))
                                                   (NEGAT (NEG ne))
                                                   (CODE +)
                                                   (VBF^1^SG^PAST^IND rûomda)))))
                   (. .)))
  (CODE {GLOSS:You_know_that_I_truth_speak._and_I_never_about_praise_me_not-boasted.})
  (CODE {LATIN:Scis_me_et_hæc_uera_proferre_._et_in_nulla_umquam_mei_laude_iactasse.})
  (ID BoethI,37.22.239))
```

- apart from SPEECH nodes, there are also SPEECHEND nodes indicating that direct speech is interrupted within a token

(16)

```
((SPEECH (MAT-DCL (ADT (ADV~TMP Nû))
          (VBF~BE^3^SG^PRES^IND ist)
          (ADT (ADV~CNT áber))
          (ADT (ADV~CNT dóh))
          (ADT (ADV^R mêr)
          (SBJ (NCO^NOM^SG^FEM zît)
               (, .)
               (OBCTV (NCO^GEN^SG^NEU láchennis)))
          (, .)
          (SPEECHEND (COORD (CONJ Unde)
                      (CJNCT-MAT-DCL (ADT-SPR (SBJ (TAG-2 0))
                                          (COPULA *cop*)
                                          (PRD-NFN (SBJ (TAG-2 0))
                                                   (DIR (PRO^ACC^1^SG míh))
                                                   (ADT (ADV~LOC+PREP~TMP tára-nâh))
                                                   (ADT (GE+ADV cnôto))
                                                   (SPX+VBA^NOM^SG^FEM^STRONG ána-séhentíu)))
                                 (, .)
                                 (VBF~3^SG^PAST^IND frâgeta)
                                 (SBJ (TAG-2 0) (PRO^NOM^3^SG^FEM si))
                                 (, .)
                                 (SPEECH (DIR-DCL (COMP *cpz*)
                                          (MAT-QUE (NEGAT (NEG Ne))
                                                   (CODE +)
                                                   (AVM-PASSIVE (VBF~WRD^2^SG^PAST^IND uuúrte))
                                                   (SBJ (PRO^NOM^2^SG dû))
                                                   (ADT (PREP mít)
                                                   (DIR (POSS (PSP~MIN^DAT^PL^NEU mînemo))
                                                        (NCO^DAT^SG^NEU spúnne)))
                                                   (GE+VBP gesóuget)
                                                   (. ?)))))))))))
(CODE {GLOSS:Now_is_however_though_more_time._medicine._and_me_there-after_sternly_on-looking.
                                     _asked_she._not-became_you_with_my_breast-milk_suckled?})
(CODE {LATIN:Sed_tempus_est_inquit_medicinæ._Tum_uero_intenta_totis_luminibus_in_me_._inquit.
                                            Tune_es_ille_qui_quondam_nutritus_nostro_lacte?})
(ID BoethI,17.16.96))
```

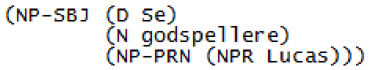
3.2 Parentheticals Overview

- the corpora PPCME and YCOE include an extension -PRN for parentheticals; annotated on clauses (quotatives, right node raising, asides, and others, e.g. bare reason adjuncts) as well as on other phrases (e.g. on NPs)

(17)

```
( (IP-MAT (CONJ bote)
          (NP-OB1 (PRO here)
                  (CP-REL (WNP-1 0)
                          (C +tat)
                          (IP-SUB (NP-SBJ *T*-1)
                                  (BED was)
                                  (VAN accused))))
          (, ,)
          (IP-MAT-PRN (NP-SBJ (D +tat))
                      (BED was)
                      (NP-OB1 (NPR Marie) (NPR Magdeleyne)))
          (, ,)
          (NP-SBJ (PRO he))
          (VBD (VBD asoylede) (CONJ and) (VBD excusyde))
          (. .))
  (ID CMAELR3,44.557))
```

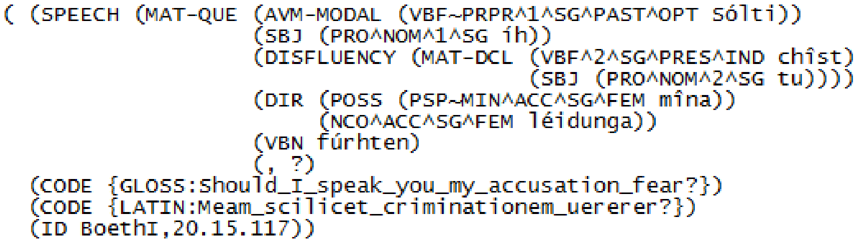
(18)



```
(NP-SBJ (D Se)
        (N godspellere)
        (NP-PRN (NPR Lucas)))
```

- clausal parentheticals and nominal parentheticals seem quite different
- the class of clausal parentheticals are not unified by an obvious criterion
- the GeCeG annotates clausal and nominal parenthetical structures differently
- clausal parentheticals are treated as a disfluency; DISFLUENCY nodes, just like SPEECH nodes, are conceptualized as scope-taking indicators; constituents marked as disfluent are sandwiched into the core clause, could be left out

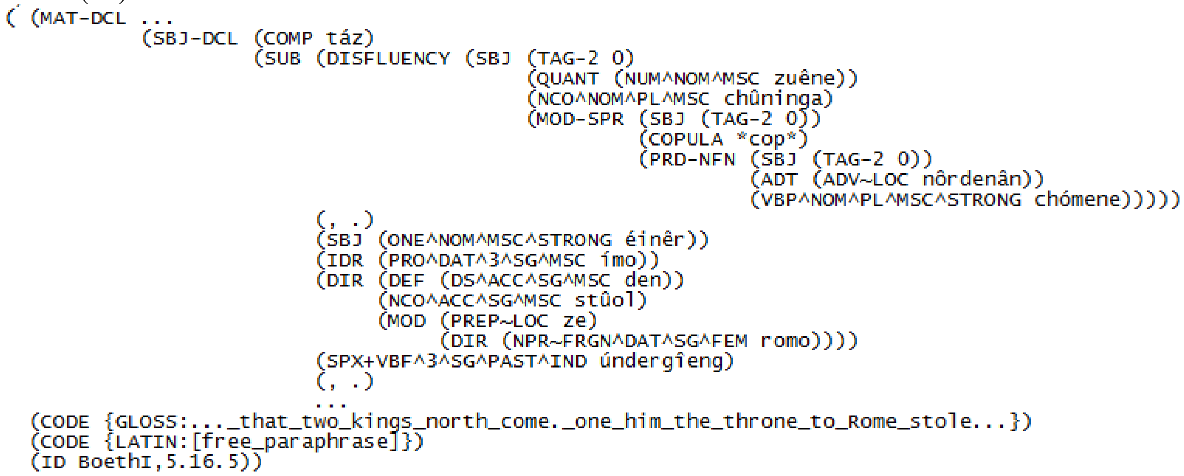
(19)



```
( (SPEECH (MAT-QUE (AVM-MODAL (VBF~PRPR^1^SG^PAST^OPT sólti))
                   (SBJ (PRO^NOM^1^SG íh))
                   (DISFLUENCY (MAT-DCL (VBF^2^SG^PRES^IND chîst)
                                        (SBJ (PRO^NOM^2^SG tu))))
                   (DIR (POSS (PSP~MIN^ACC^SG^FEM mîna))
                        (NCO^ACC^SG^FEM léidunga))
                   (VBN fúrhten)
                   (, ?)
  (CODE {GLOSS:Should_I_speak_you_my_accusation_fear?})
  (CODE {LATIN:Meam_scilicet_criminationem_uererer?})
  (ID BoethI,20.15.117))
```

- DISFLUENCY is also used for false starts, breaks, hesitations etc.

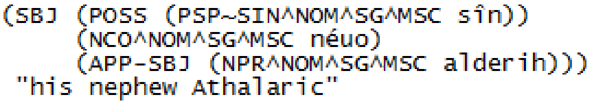
(20)



```
( (MAT-DCL ...
           (SBJ-DCL (COMP táz)
                    (SUB (DISFLUENCY (SBJ (TAG-2 0)
                                          (QUANT (NUM^NOM^MSC zuêne))
                                          (NCO^NOM^PL^MSC chûninga)
                                          (MOD-SPR (SBJ (TAG-2 0))
                                                   (COPULA *cop*)
                                                   (PRD-NFN (SBJ (TAG-2 0))
                                                            (ADT (ADV~LOC nôrdenân))
                                                            (VBP^NOM^PL^MSC^STRONG chómene)))))
                         (, .)
                         (SBJ (ONE^NOM^MSC^STRONG éinêr))
                         (IDR (PRO^DAT^3^SG^MSC ímo))
                         (DIR (DEF (DS^ACC^SG^MSC den))
                              (NCO^ACC^SG^MSC stûol)
                              (MOD (PREP~LOC ze)
                                   (DIR (NPR~FRGN^DAT^SG^FEM romo))))
                         (SPX+VBF^3^SG^PAST^IND úndergîeng)
                         (, .)
                         ...
  (CODE {GLOSS:..._that_two_kings_north_come._one_him_the_throne_to_Rome_stole...})
  (CODE {LATIN:[free_paraphrase]})
  (ID BoethI,5.16.5))
```

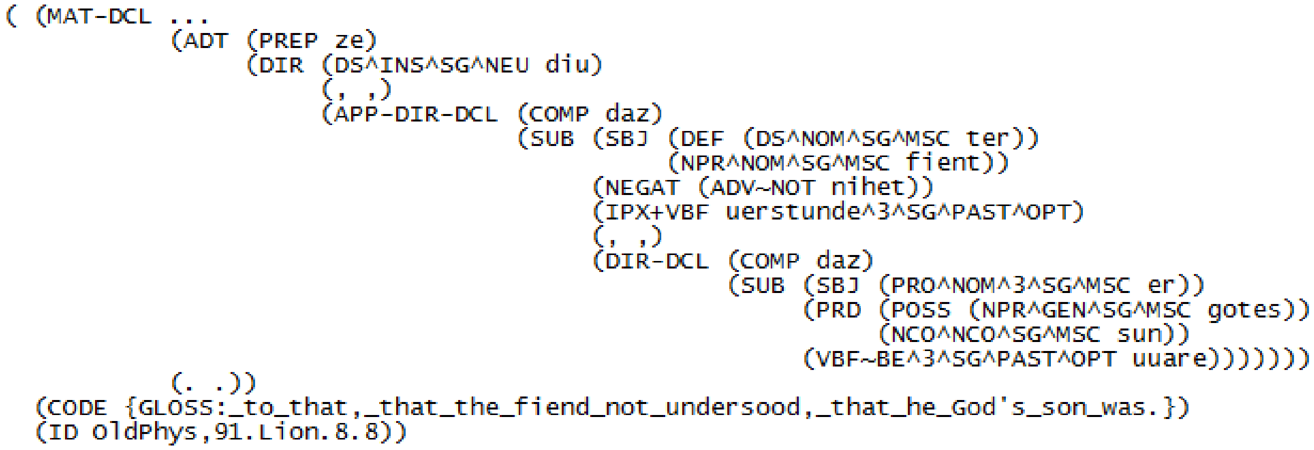
- nominal appositives are genuine functions, APP, which “compete” with the function of the mother node; an appositive must be appositive *on* something

(21)



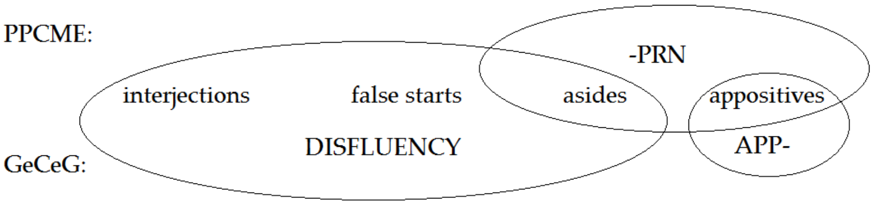
```
(SBJ (POSS (PSP~SIN^NOM^SG^MSC sîn))
      (NCO^NOM^SG^MSC néuo)
      (APP-SBJ (NPR^NOM^SG^MSC alderih))
 "his nephew Athalaric"
```

(22)



```
( (MAT-DCL ...
           (ADT (PREP ze)
                (DIR (DS^INS^SG^NEU diu)
                     (, ,)
                     (APP-DIR-DCL (COMP daz)
                                  (SUB (SBJ (DEF (DS^NOM^SG^MSC ter))
                                            (NPR^NOM^SG^MSC fient))
                                       (NEGAT (ADV~NOT nihet))
                                       (IPX+VBF uerstunde^3^SG^PAST^OPT)
                                       (, ,)
                                       (DIR-DCL (COMP daz)
                                                (SUB (SBJ (PRO^NOM^3^SG^MSC er))
                                                     (PRD (POSS (NPR^GEN^SG^MSC gotes))
                                                          (NCO^NCO^SG^MSC sun))
                                                     (VBF~BE^3^SG^PAST^OPT uuare))))))
           (. .))
  (CODE {GLOSS:_to_that,_that_the_fiend_not_undersood,_that_he_God's_son_was.})
  (ID OldPhys,91.Lion.8.8))
```

- summary:



3.3 Advantages

- easy to search: normally researchers aren't interested in differences between direct speech and "normal" tokens (`add_to_ignore: SPEECH | SPEECHEND`); in rare cases where direct speech is important, simply write normal search query with line `SPEECH doms <category of interest>`; same for `DISFLUENCY`

- appositives are also easy to search; search query simply says, "give me all elements that are appositive on some element, e.g. a subject"

- no inconsistent annotations like both `IP-MAT-PRN-SPE` as well as `IP-MAT-SPE-PRN`

- difference between asides and appositives theoretically plausible

- set of possible root nodes reduced; GeCeG only has `MAT` (matrix clauses, with subtypes declarative, question etc.), and fragments `FRAG`, i.e. `node: MAT*|FRAG` will find all tokens in the corpus; considerably simpler than other corpora

4 The TAG system

4.1 Overview

- new annotation system for movement / displacement / structure sharing / filler-gap constructions, called TAG system

- most corpora, including PPCME and YCOE, include various kinds of traces for different kinds of movements with a unique index for every local chain; basically standard model:

  ⋆-# A movement, rare because passives are not normally indicated

  ⋆ICH⋆-# A-bar movement, "interpret constituent here"

  ⋆T⋆-# A-bar movement, "trace", operator movement in relative clauses, questions etc.

- the GeCeG is different in two respects:
  - it employs only one displacement marker, called TAG-#
  - it is inserted *wherever* the displaced constituent is interpreted, i.e. it does not receive different indices for every local chain, but one index for all chains

- for A-movement: two grammatical functions will include a TAG marker (e.g. [(subject TAG-1 *Joe*) *seems to* [(subject TAG-1) *know all the answers*.]])

- for A-bar movememnt, displaced constituent is called DISC for "generic discourse function" (e.g. [(DISC TAG-1 *Bagels*) *I like* (object TAG-1)]

- resumptive elements are also included in the TAG system (e.g. (e.g. [(DISC TAG-1 *This woman*) *I like* (object TAG-1 *her*)]

- example: the consituent *Which soldiers* below is marked with TAG-1 and co-indexed with identical displacement markers in all relevant gaps

(23) [DISC TAG-1 Which soldiers] did the general convince TAG-1 [TAG-1 should scrub the floor [TAG-1 naked] [without warning TAG-1 [that a TV company would film TAG-1]]]?

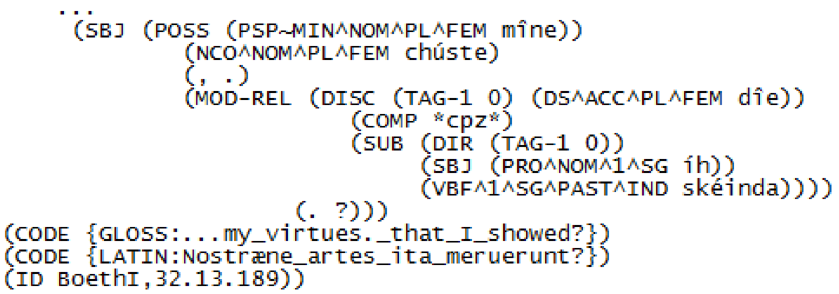
4.2 Uses of the TAG marker

- This annotation scheme allows a unified treatment of all filler-gap constructions: (i) A-bar and A movement operations,(ii) control (iii) multiple gaps involving more than one movement e.g. cyclic movement (for argument chains), (iv) left dislocations and other correlative / resumptive structures.

- examples:

- A-bar movement: relativization

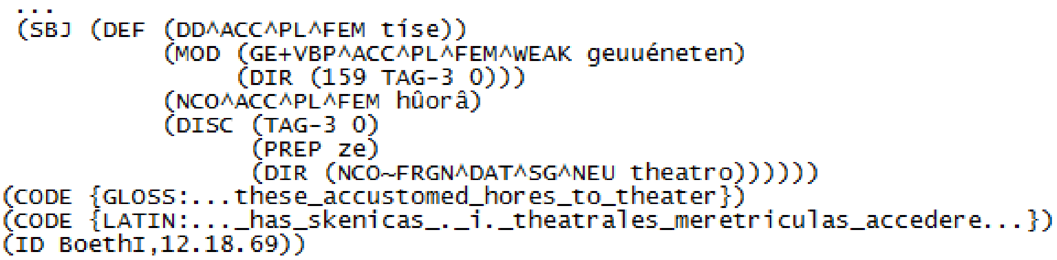
(24)



```
    ...
     (SBJ (POSS (PSP~MIN^NOM^PL^FEM mîne))
             (NCO^NOM^PL^FEM chúste)
             (, .)
             (MOD-REL (DISC (TAG-1 0) (DS^ACC^PL^FEM dîe))
                         (COMP *cpz*)
                         (SUB (DIR (TAG-1 0))
                              (SBJ (PRO^NOM^1^SG íh))
                              (VBF^1^SG^PAST^IND skéinda))))
                     (. ?)))
(CODE {GLOSS:...my_virtues._that_I_showed?})
(CODE {LATIN:Nostræne_artes_ita_meruerunt?})
(ID BoethI,32.13.189))
```

- A-bar movement: right dislocation

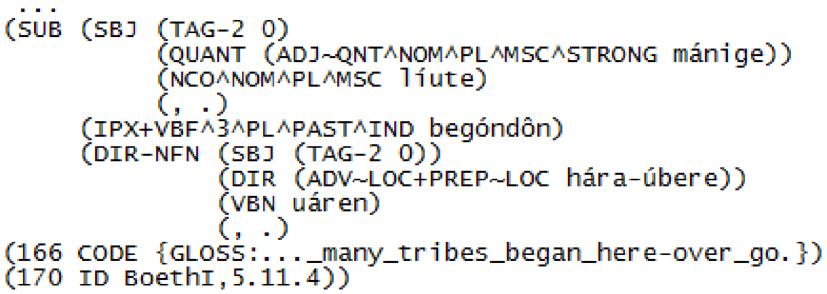
(25)



```
 ...
 (SBJ (DEF (DD^ACC^PL^FEM tíse))
           (MOD (GE+VBP^ACC^PL^FEM^WEAK geuuéneten)
                (DIR (159 TAG-3 0)))
           (NCO^ACC^PL^FEM hûorâ)
           (DISC (TAG-3 0)
                 (PREP ze)
                 (DIR (NCO~FRGN^DAT^SG^NEU theatro))))))
(CODE {GLOSS:...these_accustomed_hores_to_theater})
(CODE {LATIN:..._has_skenicas_._i._theatrales_meretriculas_accedere...})
(ID BoethI,12.18.69))
```

- A movement: subject raising

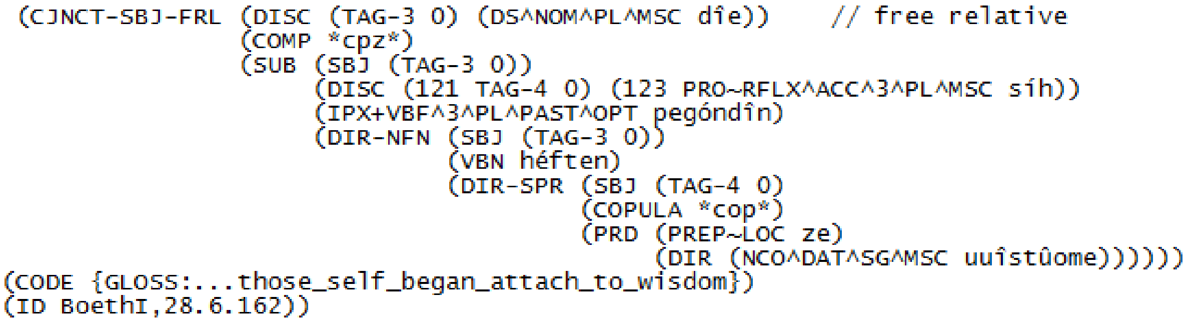
(26)



```
 ...
(SUB (SBJ (TAG-2 0)
          (QUANT (ADJ~QNT^NOM^PL^MSC^STRONG mánige))
          (NCO^NOM^PL^MSC líute)
          (, .)
     (IPX+VBF^3^PL^PAST^IND begóndôn)
     (DIR-NFN (SBJ (TAG-2 0))
              (DIR (ADV~LOC+PREP~LOC hára-úbere))
              (VBN uáren)
              (, .)
(166 CODE {GLOSS:..._many_tribes_began_here-over_go.})
(170 ID BoethI,5.11.4))
```

- more than one gap: cyclic movement

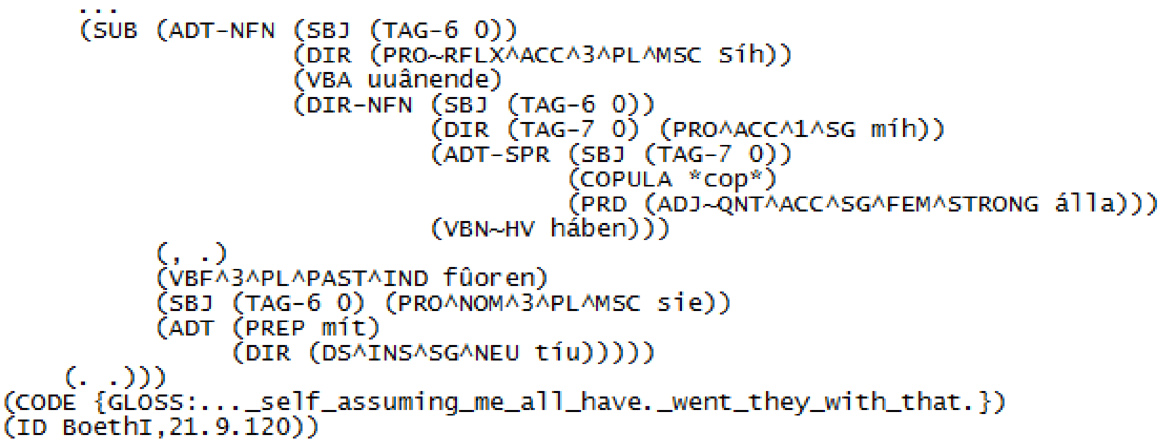
(27)



```
 (CJNCT-SBJ-FRL (DISC (TAG-3 0) (DS^NOM^PL^MSC dîe))    // free relative
                (COMP *cpz*)
                (SUB (SBJ (TAG-3 0))
                     (DISC (121 TAG-4 0) (123 PRO~RFLX^ACC^3^PL^MSC síh))
                     (IPX+VBF^3^PL^PAST^OPT pegóndîn)
                     (DIR-NFN (SBJ (TAG-3 0))
                              (VBN héften)
                              (DIR-SPR (SBJ (TAG-4 0)
                                       (COPULA *cop*)
                                       (PRD (PREP~LOC ze)
                                            (DIR (NCO^DAT^SG^MSC uuîstûome))))))
(CODE {GLOSS:...those_self_began_attach_to_wisdom})
(ID BoethI,28.6.162))
```

- more than one gap: a sequence of control structures

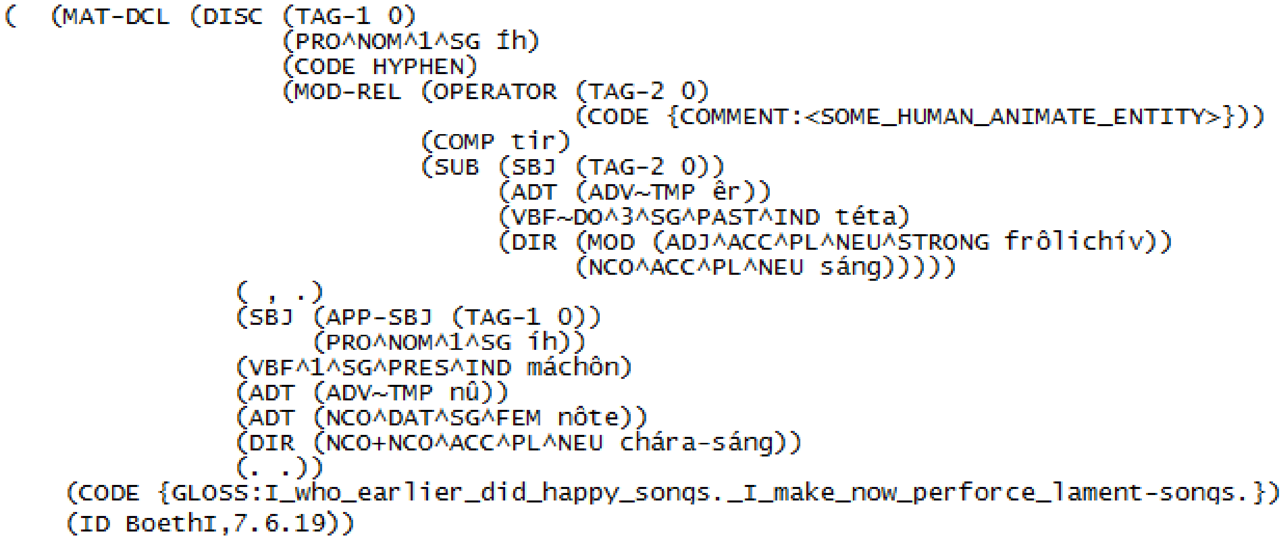
(28)



```
     ...
     (SUB (ADT-NFN (SBJ (TAG-6 0))
                   (DIR (PRO~RFLX^ACC^3^PL^MSC síh))
                   (VBA uuânende)
                   (DIR-NFN (SBJ (TAG-6 0))
                            (DIR (TAG-7 0) (PRO^ACC^1^SG míh))
                            (ADT-SPR (SBJ (TAG-7 0))
                                     (COPULA *cop*)
                                     (PRD (ADJ~QNT^ACC^SG^FEM^STRONG álla)))
                            (VBN~HV háben)))
          (, .)
          (VBF^3^PL^PAST^IND fûoren)
          (SBJ (TAG-6 0) (PRO^NOM^3^PL^MSC sie))
          (ADT (PREP mít)
               (DIR (DS^INS^SG^NEU tíu)))))
    (. .)))
(CODE {GLOSS:..._self_assuming_me_all_have._went_they_with_that.})
(ID BoethI,21.9.120))
```

- left-dislocation

(29)



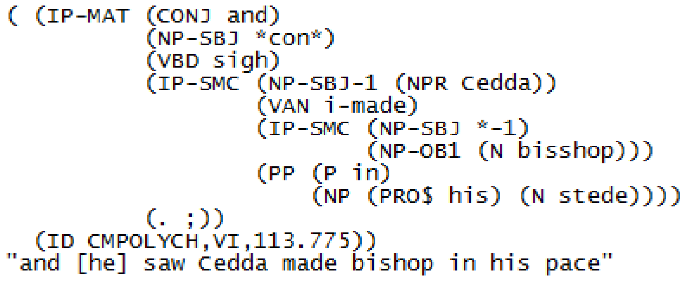
```
(  (MAT-DCL (DISC (TAG-1 0)
                  (PRO^NOM^1^SG îh)
                  (CODE HYPHEN)
                  (MOD-REL (OPERATOR (TAG-2 0)
                                     (CODE {COMMENT:<SOME_HUMAN_ANIMATE_ENTITY>}))
                           (COMP tir)
                           (SUB (SBJ (TAG-2 0))
                                (ADT (ADV~TMP êr))
                                (VBF~DO^3^SG^PAST^IND téta)
                                (DIR (MOD (ADJ^ACC^PL^NEU^STRONG frôlichív))
                                     (NCO^ACC^PL^NEU sáng)))))
               ( , .)
               (SBJ (APP-SBJ (TAG-1 0))
                    (PRO^NOM^1^SG íh))
               (VBF^1^SG^PRES^IND máchôn)
               (ADT (ADV~TMP nû))
               (ADT (NCO^DAT^SG^FEM nôte))
               (DIR (NCO+NCO^ACC^PL^NEU chára-sáng))
               (. .))
    (CODE {GLOSS:I_who_earlier_did_happy_songs._I_make_now_perforce_lament-songs.})
    (ID BoethI,7.6.19))
```

4.3 Advantages

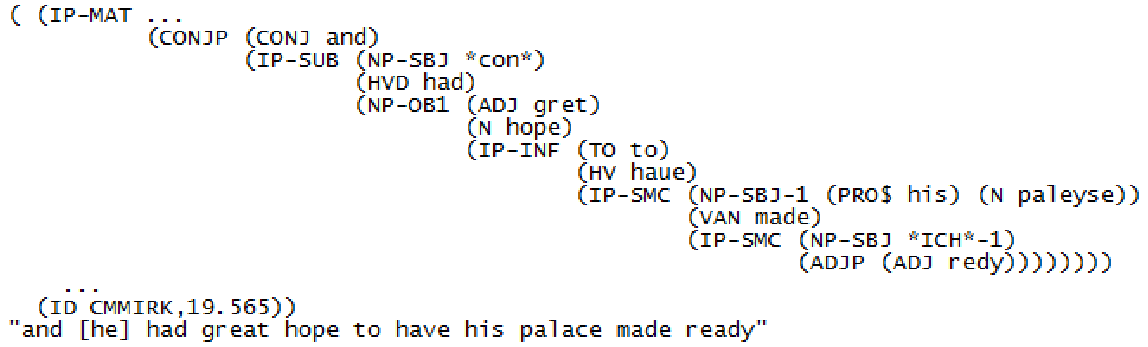
- a system with many different kinds of traces is more complicated and thus may lead annotators to make errors, especially in complex examples
- such mistakes are avoided in a system with just one displacement marker; results in more consistent annotation

- e.g.small clause containing another small clause

(30)



```
( (IP-MAT (CONJ and)
          (NP-SBJ *con*)
          (VBD sigh)
          (IP-SMC (NP-SBJ-1 (NPR Cedda))
                  (VAN i-made)
                  (IP-SMC (NP-SBJ *-1)
                          (NP-OB1 (N bisshop)))
                  (PP (P in)
                      (NP (PRO$ his) (N stede))))
          (. ;))
  (ID CMPOLYCH,VI,113.775))
"and [he] saw Cedda made bishop in his pace"
```



```
( (IP-MAT ...
          (CONJP (CONJ and)
                 (IP-SUB (NP-SBJ *con*)
                         (HVD had)
                         (NP-OB1 (ADJ gret)
                                 (N hope)
                                 (IP-INF (TO to)
                                         (HV haue)
                                         (IP-SMC (NP-SBJ-1 (PRO$ his) (N paleyse))
                                                 (VAN made)
                                                 (IP-SMC (NP-SBJ *ICH*-1)
                                                         (ADJP (ADJ redy)))))))
    ...
  (ID CMMIRK,19.565))
"and [he] had great hope to have his palace made ready"
```

- first example raises lower subject with A-movement `*-1`; second example with A-bar movement `ICH-1`

- the GeCeG would necessarily require a subject function of the higher small clause (since all top-level categories require a subject) leaving no room for a generic discourse label `DISC`. The subject of the lower small clause would be unified with the subject of the higher clause with the usual TAG marker rather than one of numerous potential traces. This would automatically force the correct GeCeG analogue of the PPCME annotation.

- GeCeG's `TAG` system is simpler than the PPCME system:
  - no need for left dislocation (`LFD`) and resumptive (`-RSP`) labels
  - primitives of different kinds of movements are identified and separated: `DISC` vs. unification with a grammatical function (=A-bar vs. A movement), presence and absence of overt material in filler (= "ordinary" movement vs. resumption), place of gap (local vs. long distance dependency) etc. and do not need to be annotated with different labels

- explicit annotation of subjects in all clauses; control is not made explicit in other corpora

5 Conclusion

- GeCeG is only a small project, but offers an opportunity to experiment with and improve existing corpus manuals
- result of suggested improvements lead to an annotations scheme with, I believe, unprecedented detail
- the suggested system is simpler than syntactic annotation in other comparable corpora and less error-prone
- the core principles can be applied universally; specific names of categories will change, but guidelines for headedness, extra-syntactic scope markers like direct speech, and for movement can be applied to all languages

References

Randell, Beth. 2004 *CorpusSearch 2*.

Kroch, Anthony and Taylor, Ann. 2000. *Penn-Helsinki Parsed Corpus of Middle English*. Department of Linguistics, University of Pennsylvania. http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3 (Accessed 10 April 2013)