Morphology

Joanna Blaszczak[1], *Stefanie Dipper*[1], *Gisbert Fanselow*[1], *Shinishiro Ishihara*[1], *Svetlana Petrova*[2], *Stavros Skopeteas*[1], *Thomas Weskott*[1], *Malte Zimmermann*[1]

University of Potsdam ([1]) and Humboldt University Berlin ([2])

The guidelines for morphological annotation contain the layers that are necessary for understanding the structure of the words in the object language: morphological segmentation, glossing, and annotation of part-of-speech.

1 Preliminaries

The guidelines for these layers follow existing recommendations in language typology and norms for the creation of language corpora. The glossing guidelines belong to the paradigm of guidelines that has arisen on the basis of *Eurotyp* (König et al. 1993), being more closely related to the conventions of the *Leipzig Glossing Rules* (see Bickel et al. 2002). The guidelines for morphological categories combine the practices recommended in *Eurotyp* with norms that have been established for the morphological annotation of corpora such as *EAGLES* (Leech & Wilson 1996) and *STTS* (Schiller et al. 1999).

Interdisciplinary Studies on Information Structure 07 (2007): 55–94
Dipper, S., M. Götze, and S. Skopeteas (eds.):
Information Structure in Cross-Linguistic Corpora

2 Layer Declaration

Table 2: Layers

Layer	Abbreviation
morphemic segmentation	MORPH
morpheme-to-morpheme translation	GLOSS
part of speech	POS

3 Layer I: Morphemic Segmentation (MORPH)

3.1 Introduction

The layer of morphemic segmentation (sometimes referred to as morphemic transcription) indicates morpheme boundaries. It contains a copy of the original text and makes use of special characters like hyphens, dots, etc. to segment words into morphemes.

Instructions for the use of this layer:

(1) English

<WORDS>	The	wolf	jumps	out	of	the	building.
<MORPH>			jump-s				

The proposed guidelines are based on *Leipzig Glossing Rules* (see Bickel et al. 2002).

3.2 Tagset declaration

Table 3: Tagset declaration for morphemic segmentations

tag	meaning	see in:
<new cell>	word boundary	§3.3.1
-	morpheme boundary	§3.3.2
=	clitic boundary	§3.3.3
_	union of sublexical components	§0
0	zero affix	§3.3.6

3.3 Instructions

3.3.1 Word boundaries

Words are given in separate cells in Exmaralda (otherwise separated through spaces).

(2) English

<WORDS>	the	children	work
<MORPH>	the	children	work

Instructions for the identification of word boundaries:

- If the object language has an orthographical representation that indicates word boundaries, then annotate the word boundaries indicated in the local orthography.
- If the orthographical representation in the object language indicates sublexical units (usually syllables) instead of words, then see §0.

3.3.2 Morpheme boundaries

Morphemes are separated by a hyphen:

(3) English

<WORDS>	Peter	works
<MORPH>	Peter	work-s

Inflection

- If the morpheme boundaries in the object language are transparent, then they should be indicated in the morphemic transcription. This holds especially for agglutinative languages, but also for morphemes that may be easily distinguished in fusional languages.

(4) English

<WORDS>	Peter	works
<MORPH>	Peter	work-s

- If the morpheme boundaries in the object language are not transparent, then do not indicate boundaries in cases where it is not feasible to establish some uncontroversial conventions. This holds especially for fusional languages. In the morphemic translation, these cases must be treated as shown in §4.4.3.

(5) English

<WORDS>	children
<MORPH>	children

(6) German

<WORDS>	entbrannt
<MORPH>	entbrannt
<GLOSS>	conflagrant

Word formation

- If the stems of a compound can be easily separated and the semantics of the compound can be compositionally derived by the unification of the semantics of the individual roots, then the analytical representation is preferred. Note that in contrast to some other current practices, the stems contained in the compound are separated by a hyphen (not by a plus sign):

(7) German

<WORDS>	Bürgersteig
<MORPH>	Bürger-steig
<GLOSS>	citizen-path

(8) Japanese

<WORDS>	gengogaku
<MORPH>	gengo-gaku
<GLOSS>	language-study

- Compositional morphemes are also separated by a hyphen and are indicated as such in the morphemic translation:

(9) German

<WORDS>	Legehenne
<MORPH>	Leg-e-henne
<GLOSS>	lay-0-hen(F)

- If the internal structure of compounds and derivatives displays difficulties in the object language (in terms of identification of the morpheme boundaries or in terms of semantic compositionality), then do not indicate the internal structure of the word.

(10) German

<WORDS>	Erdbeere
<MORPH>	Erdbeere
<GLOSS>	strawberry

3.3.3 Clitic boundaries

Clitic boundaries are indicated by an equal sign. They may be tokenized with their phonological target as in example (18). In other cases, it might be preferrable to tokenize the clitic separately, e.g. when the orthographical transcription in the <WORDS> layer requires separate tokens for the clitic and its target (see example (19) below):

(11) German

<WORDS>	wie	geht’s
<MORPH>	wie	geht=s
<GLOSS>	how	go:3.SG=it

Instructions for the identification of clitics: Clitics are phonologically weak (unstressed) elements that need a host in the form of a phonologically strong (stressed) element on which they (mostly in their reduced form) cliticize, e.g., *kommste* (= *kommst du*), *s'Fenster* (= *das Fenster*)

- For elements like *zum*, *am*, *ins*, *vom* (German), *au*, *des*, *aux* (French), see §4.4.4.
- In languages which provide an opposition between clitic and emphatic (personal, relative, etc.) pronouns or auxiliaries, clitics are identified through the use of the clitic boundary “=”:

(12) Greek

<WORDS>	to	thélo
<MORPH>	to=	thél-o
<GLOSS>	3.SG=	want-1.SG

(13) Greek

<WORDS>	aftó	thélo
<MORPH>	aftó	thél-o
<GLOSS>	3.SG	want-1.SG

(14) English

<WORDS>	he	‘s	leaving
<MORPH>	he	=s	leav-ing

(15) English

<WORDS>	he	is	leaving
<MORPH>	he	is	leav-ing

3.3.4 Union of sublexical components

This rule applies especially in languages in which blank spaces in the orthography do not always indicate word boundaries. Sublexical components of one word are put in one cell and are connected by an underscore:

(16) Vietnamese

<WORDS>	tiểu thuyết
<MORPH>	tiểu_thuyết
<GLOSS>	roman

The original form is one orthographical form in Vietnamese. Blank spaces in Vietnamese are orthographically ambiguous: they denote both word boundaries and syllable boundaries. Many words contain more than one syllable, which may be assigned only a common translation (a syllable-by-syllable translation is not possible). In morphemic segmentation, syllable boundary is represented by blank space.

3.3.5 Special characters

Special characters, i.e. non-alpha-numerical characters, such as -, %, ‘, “,), etc., that are used in orthographic representations (that may be used in WORDS) are left out at the layer of morphemic segmentation, see examples (17)-(18).

(17) German

<WORDS>	das	“Pünktchen”
<MORPH>	das	Pünkt-chen
<GLOSS>	DEF:N.SG.NOM	point-DIM

Note that the hyphen has different meaning in the two layers of example (18): at the layer WORD it is an orthographic symbol, and at the layer MORPH it encodes morpheme boundaries.

(18) German

<WORDS>	die	“Pünktchen”-Partei
<MORPH>	die	Pünkt-chen-Partei
<GLOSS>	DEF:F.SG.NOM	point-DIM-party

3.3.6 Zero morphemes

The indication of zero morphemes is sometimes part of the morphemic segmentation. Since a morphemic analysis in terms of zero morphemes is not theory neutral, we recommend avoiding the use of zeroes in the database. If a project needs this kind of information for its data, the standard symbol ‘0’ is recommended (note that ‘0’ is also used in glossing, compare (57)).

(19) German

<WORDS>	die	Lehrer
<MORPH>	die	Lehrer-0
<GLOSS>	DEF:NOM.PL	teacher-PL

4 Layer II: Morphemic Translation (GLOSS)

4.1 Introduction

The layer of morphemic translation identifies the lexical meaning or grammatical function of individual morphemes as they are segmented at the layer of morphemic transcription. This section includes:

- rules for morpheme-to-morpheme translation;
- the list of tags for the recommended glosses.

4.2 Related standards

The proposed guidelines are based on *Leipzig Glossing Rules* (see Bickel et al. 2002) and Eurotyp (see König et al. 1993). In particular, a basic list of abbreviations is adopted from LGR – and if not available in this standard from Eurotyp (see König et al. 1993); further tags for terms that are not available in these standards and are needed for our corpus have been introduced in our document.

4.3 Tagset declaration

The symbols used at the MORPH layer are replicated at the GLOSS layer. In addition to these symbols (see §3.2), some symbols are only used in the GLOSS:

Table 4: Conventions for morphemic translation

tag	meaning	see in:
x:y	x and y are different morphemes with non-segmentable boundaries	§4.4.4; 4.4.5
x.y	x and y are semantic components of the same morpheme	§4.4.4; 4.4.5
x_*n*	all x_*n* are parts of the same discontinuous morpheme	§4.4.3
x/y	x and y are alternating meanings/meaning components	§4.4.6
{x}	x is a feature not realized in this context	§4.4.6
[x]	x is non-overtly encoded	§4.4.6; 0
XXX	grammatical meaning	§4.4.8

4.4 Instructions

4.4.1 Isomorphism between GLOSS and MORPH

Symbols introduced at the layer of morphemic segmentation for the indication of boundaries (§3.2) are also used obligatorily in morpheme translations in a one-to-one relation. For exceptions to the general principle of isomorphism see §4.4.2-0.

- word boundaries

(20) German

<WORDS>	heute	morgen
<MORPH>	heute	morgen
<GLOSS>	today	morning

- morpheme boundaries

(20) English

<WORDS>	works
<MORPH>	work-s
<GLOSS>	work-3.SG

- clitic boundaries

(21) German

<WORDS>	wie	geht’s
<MORPH>	wie	geht=s
<GLOSS>	how	go:3.SG=3.SG.NOM

4.4.2 Non-Isomorphism: Sublexical components

In case the morphemic transcription contains more than one sublexical components (indicated by an underscore; see §0), they correspond to one unit at the GLOSS layer.

(22) Vietnamese

<WORDS>	tiểu thuyết
<MORPH>	tiểu_thuyết
<GLOSS>	roman

4.4.3 Non-Isomorphism: Discontinuity

Discontinuous morphemes are indicated by repeating the gloss in each part of the morpheme. The parts of the discontinuous morpheme are indicated through the index ‘_*n*’. In infixation, the discontinuous morpheme is the root:

(23) Tagalog

<WORDS>	bili
<MORPH>	bili
<GLOSS>	buy

<WORDS>	bumili
<MORPH>	b-um-ili
<GLOSS>	buy_1-A.FOC-_1

In circumfixation, the discontinuous morpheme is the affix:

(24) Tuwali Ifugao, Philippines

<WORDS>	baddang
<MORPH>	baddang
<GLOSS>	help

<WORDS>	kabaddangan
<MORPH>	ka-baddang-an
<GLOSS>	NMLZ_1-help-_1

The same logic applies to cases like the particle verbs in German, where the particle can be separated from the verb and can occur like an independent word:

(25) German

<WORDS>	ich	fange	mit	dem	Studium	an
<MORPH>	ich	fange	mit	dem	Studium	an
<GLOSS>	1.SG	start:1.SG_1	with_1	DEF:DAT.N	study[DAT.N]	_1

<WORDS>	weil	ich	mit	dem	Studium	anfange
<MORPH>	weil	ich	mit	dem	Studium	anfange
<GLOSS>	because	1.SG	with	DEF:DAT.N	study[DAT.N]	start:1.SG

4.4.4 Non-Isomorphism: Non-indicated boundaries

If the original form contains different morphemes that are not segmented (at the MORPH layer), then a colon is used in the gloss:

(26) German

<WORDS>	geht
<MORPH>	geht
<GLOSS>	go:3.SG

Special instructions for non-indicated boundaries:

- Morpheme boundaries that may not be easily identified in a theory neutral way, are not indicated (see §3.3.2):

(27) German

<WORDS>	ging
<MORPH>	ging
<GLOSS>	go:PAST:1.SG

- In the case of portmanteau morphemes (i.e. morphemes that fuse more than one grammatical functions), it usually makes no sense to indicate boundaries in the morphemic transcription; however, the different grammatical functions can be read off the GLOSS layer:

(28) French

<WORDS>	au
<MORPH>	au
<GLOSS>	to.DEF.SG.M

4.4.5 Non-Isomorphism: Complex glosses

If the morphemic translation contains more than one gloss, the glosses are separated by periods:

(29) Polish

<WORDS>	ciastko
<MORPH>	ciastko
<GLOSS>	cake:SG.NOM.N

Special instructions for complex glosses:

- Amalgamated grammatical information in fusional languages is translated through complex glosses:

(30) Polish

<WORDS>	ciastko
<MORPH>	ciastko
<GLOSS>	cake:SG.NOM.N

- Person and number combinations are treated as complex glosses:

(31) German

<WORDS>	geht
<MORPH>	geht
<GLOSS>	go:3.SG

- Lexical information that may not be translated by a single element in the translation language is treated as a complex gloss:

(32) Hawaian

<WORDS>	ulua
<MORPH>	ulua
<GLOSS>	old.man

- In complex glosses conveying grammatical information the following orders are used:

  NOMINAL INFLECTION

  {gender}.{number}.{case} (for nouns, adjectives, and determiners)
  The order of these categories corresponds to the cross-linguistically preferred order for the realization of the corresponding morphemes.

(33) Polish

<WORDS>	ciastko
<MORPH>	ciastko
<GLOSS>	cake:N.SG.NOM

(34) Spanish

<WORDS>	mojigata
<MORPH>	mojigata
<GLOSS>	prude:F.SG.NOM

(35) Spanish

<WORDS>	una
<MORPH>	una
<GLOSS>	INDEF:F.SG.NOM

PRONOMINAL INFLECTION

{person}.{number}.{gender}.{case}

The idea of this order is to start the GLOSS with the information which identifies the paradigms as they are commonly presented in grammars, e.g. “2nd singular”, “3rd singular masculine”; the relational information, i.e. case, comes at the end of the GLOSS.

(36) German

<WORDS>	du
<MORPH>	du
<GLOSS>	2.SG.NOM

(37) German

<WORDS>	ihm
<MORPH>	ihm
<GLOSS>	3.SG.M.DAT

(38) German

<WORDS>	wir
<MORPH>	wir
<GLOSS>	1.PL.NOM

- Elements denoting person/number are decomposed into their semantic features if they are personal pronouns (i.e., if they belong to a syntactically identifiable paradigm that structures person/number oppositions in the object language):

(39) German

<WORDS>	sie
<MORPH>	sie
<GLOSS>	3.SG.NOM.F

<WORDS>	mir
<MORPH>	mir
<GLOSS>	1.SG.DAT

<WORDS>	wir
<MORPH>	wir
<GLOSS>	1.PL.NOM

- If the categorial status of these elements is not different from simple nouns, then their meaning is rendered by the English translation:

(40) Japanese

<WORDS>	kanojo
<MORPH>	kanojo
<GLOSS>	she

VERB INFLECTION

{aspect}.{voice}.{finiteness}.{tense}.{mood}.{person}.{gender}.
{number}

(41) Ancient Greek

<WORDS>	lusaímēn
<MORPH>	lusaímēn
<GLOSS>	unbind:PFV.MID.PST.OPT.1.SG

The conventions for the order of morphological categories only hold for complex morpheme glosses, which contain more than one piece of grammatical information. Otherwise, the GLOSS corresponds to the actual order of morphemes.

(42) Turkish

<WORDS>	bilmiyorum
<MORPH>	bil-m-iyor-um
<GLOSS>	know-NEG-PROG-1.SG

4.4.6 Non-isomorphism: Alternative meanings

If a given grammatical or lexical morpheme has different meanings (that are activated in different contexts; in cases of either polysemy or homonymy), we recommend that only the context-relevant meaning is given:

(43) German

<WORDS>	vom	Jahr
<MORPH>	vom	Jahr
<GLOSS>	from:DEF.SG.DAT.N	year[DAT.SG]

(44) German

<WORDS>	das	Band
<MORPH>	das	Band
<GLOSS>	DEF:N.SG.NOM	tape[NOM.SG]

<WORDS>	der	Band
<MORPH>	der	Band
<GLOSS>	DEF:M.SG.NOM	volume[NOM.SG]

If in particular parts of the corpus you wish to indicate the ambiguity of particular morphemes which is resolved in syntactic context, then you may set the further alternatives in curly brackets:

(45) German

<WORDS>	vom	Jahr
<MORPH>	vom	Jahr
<GLOSS>	from:DEF.SG.DAT.N	year[DAT]{/NOM/ACC}

(46) German

<WORDS>	das	Band
<MORPH>	das	Band
<GLOSS>	DEF:N.SG.NOM	tape[DAT]{/volume[DAT]}

Complex examples of homonymy of case morphemes:

(47) Greek

<WORDS>	kaló
<MORPH>	kaló
<GLOSS>	good{N.{NOM/ACC}.SG/M.ACC.SG}

4.4.7 Non-isomorphism: Non-overtly encoded meaning

The German word *Frau* ‘woman’ consists of only one lexical morpheme, but it also contains information about grammatical number. Thus, the glossing:

(48) German

<WORDS>	Frau
<MORPH>	Frau
<GLOSS>	woman

is incomplete, because the word *Frau* ‘woman’ in contrast to *Frauen* ‘women’ also includes the information ‘singular’. If non-overtly encoded information should be stored, use square brackets:

(49) German

<WORDS>	Frau
<MORPH>	Frau
<GLOSS>	woman[SG]

Instructions for the annotation of non-overtly encoded information:

- If the non-overtly encoded category is the unmarked category, then our recommendation is to not indicate it in the gloss. The following rules may be postulated as default:

(50) Lack of voice in the gloss for a verb implies “active”.
Lack of number in the gloss for a noun implies “singular”.
Lack of tense in the gloss for a verb implies “present”.
Lack of case in the gloss for a noun implies “absolutive” in an ergative system.

These rules are language-specific: Lack of number morpheme indicates ‘singular’ in some languages, whereas in other languages it shows ‘general number’, lack of tense/aspect morpheme indicates ‘present’ in some languages, whereas in other languages it indicates ‘imperfective’, lack of case morpheme indicates absolutive in some languages, in some languages accusative, in some languages nominative, etc. That means the rules under (50) should be respectively postulated for every language.

- If a category which is treated cross-linguistically as unmarked is encoded through paradigmatic opposition and not through the lack of a morpheme, then this category is given in the gloss:

(51) Modern Greek

<WORDS>	neró
<MORPH>	neró
<GLOSS>	water:SG.NOM.N

<WORDS>	near
<MORPH>	near
<GLOSS>	water:PL.NOM.N

(52) Modern Greek

<WORDS>	gráfo
<MORPH>	gráfo
<GLOSS>	write:ACT.PRS.IND.1.SG

4.4.8 Tags

Table 4: Tags for glosses

tag	term
0	Element without semantic content or syntactic function
1	First person
2	Second person
3	Third person
A	Agent-like argument of canonical transitive verb
ABL	Ablative
ABS	Absolutive
ACC	Accusative
ACT	Active
ALL	Allative
ANTIP	Antipassive
APPL	Applicative

tag	term
ART	Article
BEN	Benefactive
CAUS	Causative
CLF	Classifier
COMPR	Comparative
COM	Comitative
COMP	Complementizer
COMPL	Completive
COND	Conditional
COP	Copula
DAT	Dative
DECL	Declarative
DEF	Definite
DEM	Demonstrative
DIM	Diminutive
DIREV	Direct evidential marker
DIST	Distal (long distance from deictic center)
DISTR	Distributive
DU	Dual
DUR	Durative
ERG	Ergative
EXCL	Exclusive
EXPEV	Evidential marker for personal experience
F	Feminine
FILL	Break filler
FOC	Focus

tag	term
FUT	Future
GEN	Genitive
HAB	Habitual
IMP	Imperative
INCL	Inclusive
IND	Indicative
INDF	Indefinite
INF	Infinitive
INS	Instrumental
INTR	Intransitivizer
IPFV	Imperfective
IRR	Irrealis
ITER	Iterative
LOC	Locative
M	Masculine
MED	Medial (medial distance from deictic center)
MID	Middle (voice which excludes passive voice)
N	Neuter
NEG	Negative
NMLZ	Nominalizer
NOM	Nominative
NON	Negativelly defined categories
OBJ	Object
OBL	Oblique
P	Patient-like argument of canonical transitive verb
PASS	Passive

tag	term
PFV	Perfective
PL	Plural
POSS	Possessive
POT	Potential
PRF	Perfect
PRS	Present
PROG	Progressive
PROH	Prohibitive
PROX	Proximal (short distance from deictic center)
PST	Past
PTCP	Participle
PURP	Purposive
Q	Question particle/marker
QUOT	Quotative
RECP	Reciprocal
REFL	Reflexive
REL	Relative
REP	Reportative evidential marker
RES	Resultative
S	Single argument of canonical intransitive verb
SBJ	Subject
SBJV	Subjunctive
SG	Singular
SUPERL	Superlative
TOP	Topic
TR	Transitivizer

4.4.9 Special instructions

- Negatively defined categories may be rendered with the abbreviation NON. The scope of the negation operator is indicated through parentheses, e.g. NON(SG) non-singular, NON(FUT) non-future, NON(3.SG) non-third-singular.

(53) Dyirbal

<WORDS>	balgan
<MORPH>	balgan
<GLOSS>	hit.NON(FUT)

(54) English

<WORDS>	drink
<MORPH>	drink
<GLOSS>	drink.NON(3.SG)

- This tag is only used if the language possesses a category, which is negatively defined. Negatively defined terms are not used for the indication of polysemy. Thus:

(55) Modern Greek

<WORDS>	neró
<MORPH>	neró
<GLOSS>	water:SG.{NOM/ACC}

may not be rendered as in (56):

(56) Modern Greek

<WORDS>	neró
<MORPH>	neró
<GLOSS>	water:NON(PL).NON(GEN)

- The tag ‘0’ is used for elements that lack semantic content. Note that the layer “morphemic translation (GLOSS)” contains the meaning or syntactic function of the elements of the layer “morphemic segmentation”. Elements that do not have such a function are rendered as ‘0’s. E.g. in French questions, there is a liaison particule as in *que se passe-t-il?*. The *t* in this example has no semantic value, it is only there as liaison between a vowel ending verb and a vowel initial pronoun. The gloss of this element looks as follows:

(57) French

<WORDS>	que	se	passe-t-il
<MORPH>	que	se	passe-t-il
<GLOSS>	what	REFL.3.SG	happen:3.SG-0-3.SG.M

- The use of lexical verbs as auxiliaries for the formation of inflectional forms is not indicated in gloss. The gloss contains the lexical meaning of the verb. The special use of the verb in this case is indicated at the POS layer.

(58) French

<WORDS>	ai	aimé
<MORPH>	ai	aimé
<GLOSS>	have:1.SG	love:PTCP.PRF
<POS>	VAUX	VLEX

- Complex verbal aspects like ‘aorist’ should be decomposed, e.g. Modern Greek aorist is glossed as ‘PFV.PAST’ in indicative mood and as ‘PFV’ in non-indicative moods.

(59) Modern Greek

<WORDS>	fáe
<MORPH>	fáe
<GLOSS>	eat:IMPR.PFV.2.SG
<TRANS>	Eat!

(60) Modern Greek

<WORDS>	éfaje
<MORPH>	éfaj-e
<GLOSS>	eat:PFV.PAST-3.SG
<TRANS>	he/she/it has eaten

- Break fillers are elements like "hmmm…", "äh…", etc. These elements are glossed as 'FILL'.

(61) German

<WORDS>	ich	gehe	...hmm...	ins	Kino	.
<MORPH>	ich	gehe	hmm	in=s	Kino	
<GLOSS>	1.SG	go:1.SG	FILL	in:DEF:ACC.SG.N	cinema [ACC.SG.N]	
<TRANS>	I am going to the cinema.					

5 Layer III: Part of Speech (POS)

5.1 Introduction

The layer "part of speech" indicates the grammatical categories of words. The general principle behind part of speech categorization in these guidelines is syntax-oriented. The idea is not to establish language specific categories, but to provide categorial information which is relevant for syntax. For instance, the word *walk* in English may be used as a noun or a verb. Rather than establishing a new category which captures all possible functions, e.g., "V/N" for *walk*, we recommend specifying the categorial information which is relevant in that context:

(62) English

<WORDS>	the	walk
<POS>	DET	N

(63) English

<WORDS>	to	walk
<POS>	PTC	VLEX

5.2 Tagset declaration

Similar to STTS, tag names for parts of speech are organized in a hierarchical manner: The first letter(s) indicate the superordinate category, e.g. N for ‘noun’, and subsequent letters denote subclasses, e.g. NCOM for ‘common noun’.

Table 5: List of tags for part of speech

tag	term
A	adjective
ADV	adverb
AT	attributive
CLF	classifier
COOR	coordinating conjunction
DET	determiner
N	noun
NCOM	common noun
NPRP	proper noun
P	preposition/postposition
PRON	pronoun
PRONDEM	demonstrative pronoun
PRONEXPL	expletive pronoun
PRONINT	interrogative pronoun
PRONPOS	possessive pronoun
PRONPRS	personal pronoun
PRONQUANT	quantifier

PRONREL	relative pronoun
PRONRFL	reflexive pronoun
PTC	particle
SU	substantive
SUB	subordinating conjunction
SUBADV	adverbial subordinating conjunction
SUBCOM	complementizer
V	verb
VAUX	auxiliary verb
VCOP	copula verbs
VDITR	ditransitive verb
VINTR	intransitive verb
VLEX	lexical verb
VMOD	modal verb
VN	verbal noun
VTR	transitive verb
CLIT	clitic form
FULL	full form

If a part of speech has some subclasses, as, e.g., in the case of ‘nouns’ which may be further divided into ‘common nouns’ and ‘proper nouns’, then it is recommended to choose one level of categorization, i.e. either annotate every noun just as ‘N’, or make the distinction between ‘NCOM’ and ‘NPRP’ every time. The same also holds for verbs, pronouns, etc.

(64) English, annotation of supercategories

<WORDS>	Peter	bicycle
<POS>	N	N

(65) English, annotation of subcategories

<WORDS>	Peter	bicycle
<POS>	NPRP	NCOM

5.3 Specific instructions

5.3.1 Nouns

General case

(66) English

<WORDS>	water
<POS>	N

Subclasses

- proper nouns:

(67) English

<WORDS>	Peter
<POS>	NPRP

- common nouns:

(68) English

<WORDS>	house
<POS>	NCOM

5.3.2 Verbs

General case

(69) English

<WORDS>	sleep
<POS>	V

Subclasses

The following subclasses of verbs may be used according to the function of the verb in certain contexts, i.e. the verb be would be annotated as VCOP in *be happy* and VAUX in *be destroyed.* Similarly, the German verb *wollen* ‘want’ would be annotated as VMOD in *ich will gehen* ‘I want to go’ and as VLEX in *ich will ein Eis* ‘I want ice-cream’.

- modal verbs:

(70) English

<WORDS>	can
<POS>	VMOD

- auxiliary verbs:

(71) English

<WORDS>	have
<POS>	VAUX

- copula verbs:

(72) English

<WORDS>	be
<POS>	VCOP

- lexical verbs:

(73) English

<WORDS>	walk
<POS>	VLEX

The annotation of part of speech follows the syntactic function of the verb. I. e., the verb *haben* in German may be a transitive verb if it is used with a direct object, or an auxiliary verb when it is used for the formation of perfect tenses.

(74) German

<WORDS>	Hunger	haben
<MORPH>	hunger	have:INF
<GLOSS>	NCOM	VLEX

(75) German

<WORDS>	gegessen	haben
<GLOSS>	eat:PRF.PTCP	have:INF
<POS>	VLEX	VAUX

- transitivity

  It is possible to distinguish between intransitive, transitive, and ditransitive verbs by using the following glosses:

(76) English

<WORDS>	sleep
<POS>	VINTR

(77) English

<WORDS>	buy
<POS>	VTR

(78) English

<WORDS>	give
<POS>	VDITR

5.3.3 Adjectives

(79) Spanish

<WORDS>	aburrido
<GLOSS>	boring
<POS>	A

5.3.4 Adverbs

(80) English

<WORDS>	soon
<POS>	ADV

(81) English

<WORDS>	where
<POS>	ADV

So called pronominal adverbs in German are also annotated as ADV:

(82) German

<WORDS>	darüber
<GLOSS>	there:over
<POS>	ADV

(83) German

<WORDS>	hierüber
<GLOSS>	here:over
<POS>	ADV

(84) German

<WORDS>	worüber
<GLOSS>	where:over
<POS>	ADV

(85) German

<WORDS>	dessentwegen
<GLOSS>	DEM:M.GEN.SG:because.of
<POS>	ADV

(86) German

<WORDS>	meinetwegen
<GLOSS>	1.SG.GEN:because.of
<POS>	ADV

5.3.5 Adpositions

Including all types of X-positions:

(87) English

<WORDS>	behind	the	house
<POS>	P	DET	NCOM

(88) English

<WORDS>	two	years	ago
<POS>	DET	NCOM	P

5.3.6 Determiners

Determiners include articles and numerals used as determiners (see §0; §5.3.8). They do not include demonstratives or quantifiers (cf. 5.3.8).

(89) English

<WORDS>	the
<POS>	DET

5.3.7 Conjunctions

All types of subordinators are annotated as SUB:

(90) English

<WORDS>	if
<POS>	SUB

<WORDS>	that
<POS>	SUB

<WORDS>	when
<POS>	SUB

If you need to indicate complementizers or adverbial subordinating conjunctions separately, then use the corresponding tags:

(91) English

<WORDS>	when
<POS>	SUBADV

(92) English

<WORDS>	that
<POS>	SUBCOM

Coordinating conjunctions are annotated as COOR:

(93) English

<WORDS>	and
<POS>	COOR

5.3.8 Pronouns

- personal pronouns:

(94) English

<WORDS>	you
<POS>	PRONPRS

- interrogative pronouns:

(95) English

<WORDS>	who
<POS>	PRONINT

- demonstrative pronouns:

(96) English

<WORDS>	this
<POS>	PRONDEM

Notice that German displays a demonstrative pronoun that is in most cases homonymous to the definite article.

(97) German

<WORDS>	Das	ist	es	.
<GLOSS>	this:N.SG.NOM	be:3.SG	3.SG.NOM	
<POS>	PRONDEM	VCOP	PRONPERS	

- reflexive pronouns:
  This category should be used only if the language possesses pronouns which are always used as reflexives, e.g. the English reflexive pronouns (not the German pronouns of the type *ich schäme mich*, where the ambiguity personal/reflexive is resolved in the argument structure of the given verb).

(98) English

<WORDS>	myself
<POS>	PRONRFL

- possessive pronouns:

(99) English

<WORDS>	your
<POS>	PRONPOS

- relative pronouns:

(100) English

<WORDS>	which
<POS>	PRONREL

- expletive pronouns:

  Expletive pronouns (also called “impersonal pronouns”, “pleonastic pronouns”) are pronouns which do not have any meaning but are syntactically required, as for instance:

(101) English

<WORDS>	there	is	a	man	.
<POS>	PRONEXPL	V	DET	N	

(102) German

<WORDS>	es	riecht	nach	Erdbeeren	.
<GLOSS>	3.SG	smell:3.SG	to	strawberry:DAT.PL	
<POS>	PRONEXPL	V	P	N	

(103) German

<WORDS>	es	regnet	.
<GLOSS>	3.SG	rain:3.SG	
<POS>	PRONEXPL	V	

We also use PRONEXPL for pre-field *es* in German. The difference between *es* in (101)-(103) and *es* in (104) is encoded at the syntactic layer:

(104) German

<WORDS>	es	kamen	drei	Sportler	.
<GLOSS>	3.SG	come:3.PL	three	sportsman[PL]	
<POS>	PRONEXPL	V	DET	N	

- quantifiers:

  The properties of quantifiers are described in detail in the semantics guidelines.

(105) German

<WORDS>	jeder
<GLOSS>	every.one:M.SG.NOM
<POS>	PRONQUANT

(106) German

<WORDS>	jeder	Mann
<GLOSS>	every:M.SG.NOM	man
<POS>	PRONQUANT	NCOM

(107) German

<WORDS>	alle
<GLOSS>	all:PL.NOM
<POS>	PRONQUANT

If you need to differentiate between substantive and attributive paradigms of pronouns, then use the following tags (append SU and AT respectively). Substantive pronouns replace the whole NP, attributive ones function as a determiner:

(108) English

<WORDS>	yours
<POS>	PRONPOSSU

(109) English

<WORDS>	your
<POS>	PRONPOSAT

5.3.9 Particles

(110) German

<WORDS>	ja
<GLOSS>	yes
< POS >	PTC

Interjections are also annotated as particles:

(111) German

<WORDS>	oh
<GLOSS>	oh
<POS>	PTC

5.3.10 Special instructions

Clitic vs. full forms

If a language makes a difference between clitic and full forms in a given category, then append the tags 'FULL' and 'CLIT'. E.g.,

(112) Croatian

<WORDS>	jesam	sam
<MORPH>	be:1.SG	be:1.SG
<GLOSS>	VAUXFULL	VAUXCLIT

(113) Modern Greek

<WORDS>	eména
<GLOSS>	1.SG.ACC
<POS>	PRONPRSFULL

<WORDS>	me
<GLOSS>	1.SG.ACC
<POS>	PRONPRSCLIT

Numerals

Numerals are treated as members of broader syntactic categories (for the explicit marking of numerals, use the Semantic Annotation Layer QuP):

- cardinal numerals in English are treated as determiners;
- ordinal numerals in English are treated as adjectives;
- adverbial numerals in English are treated as adverbs.

(114) English

<WORDS>	two
<POS>	DET

<WORDS>	second
<POS>	A

<WORDS>	twice
<POS>	ADV

Discontinuity

Similar to discontinuous morphemes (see §4.4.3), discontinuous elements are indicated by indices also in the POS layer:

(115) English

<WORDS>	either	John	or	Mary
<POS>	COOR_1	NPRP	_1	NPRP

(116) German

<WORDS>	ich	fange	jetzt	an
<MORPH>	ich	fange	jetzt	an
<GLOSS>	1.SG	start:1.SG_1	now	_1
<POS>	PRONPRS	VLEX_1	ADV	_1

(117) German

<WORDS>	um	unseres	Vaters	willen
<POS>	P_1	PRONPOS	NCOM	_1

6 References

Bickel, Balthasar, Bernard Comrie, and Martin Haspelmath. 2002. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. Leipzig: MPI for Evolutionary Anthropology & University of Leipzig (http://www.eva.mpg.de/lingua/files/morpheme.html).

König, Ekkehard (with Dik Bakker, Öesten Dahl, Martin Haspelmath, Maria Koptjevskaja-Tamm, Christian Lehmann, Anna Siewierska). 1993, *EUROTYP Guidelines*. European Science Foundation Programme in Language Typology.