3 The Morphosyntactic Tagging

3.1 Principles behind the Tag Set

The principles behind the morphosyntactic tagset have been extensively discussed in Ejerhed, Källgren, Wennstedt & Åström (1992).

3.2 Definition of the tags

3.2.1 The SUC Morphosyntactic Tagset

The morphosyntactic tagset used in SUC was designed at an early stage of the project. It was documented in Ejerhed, Källgren, Wennstedt & Åström (1992) and was used to tag the 300 000-word subpart of SUC published on the first ECI CD-ROM (1993). Since then the tagset has only undergone a few changes, most notably the introduction of a part-of-speech tag for verbal particles, PL.

The tagset is based upon that used by the SWETWOL (Karlsson 1992) with some modifications, e.g. in the subclassification of adverbs and in the relative order between elements. Throughout the project, the first part of the tagging of SUC texts was done in co-operation with the Department of General Linguistics at Helsinki, where the words were given SWETWOL tags. The tags were then transduced into SUC tags at the Department of Linguistics at Umeå.

There also exists a one-to-one mapping between the SUC tags and the more compact Parole tags (see 'taggtabell' at http://spraakbanken.gu.se/lb/parole for translation schemes in both directions). Mostly this mapping is quite straightforward. One exception is the treatment of participles, which SUC regards as a separate part-of-speech while Parole classifies them as adjectives.

Another difference is that the SUC format gives the text word at the beginning of the line and the base form at the end of it, while Parole has it the other way round. This makes SUC, with its verticalized running text, infinitely more readable to a human eye. Texts in Parole format are not supposed to be read by humans. Still, many scholars working with texts, e.g. doing analysis on discourse level, have to look at the texts with markup and read them as best they can. That is why we have kept this order between the elements in the SUC format of the corpus.

SUC 2.0 presently exists in three sgml-versions, two with SUC tags and one with Parole tags. Two of them are TEI-conformant, while the third has a local sgml-format not strictly TEI-conformant. The text words and their analyses are the same for all three versions. Below is a presentation of the current morphosyntactic SUC tags. For each part-of-speech tag the corresponding part of the Parole tag is also given. (Table 14)

3.2.2 The Structure of SUC Tags

Each SUC tag contains a part-of-speech label. (cf. Table 12). For many parts-of-speech, this is all there is of morphosyntactic information. For others, the part-of-speech tag is followed by one or more feature values for various properties of the tagged word. (cf. Table 13). Last in all tags, complex or simple, comes the base form of the word. In the SGML-format used, this will look like in the example below, the plural form arenor 'arenas' of the noun arena. (<w> is the SGML-tag used for words in SUC, <ana> stands for analysis, <ps> for part-of-speech, <m> for morphosyntactic information and <b> for base form.)

```
<w>arenor<ana><ps>NN<m>UTR PLU IND NOM<b>arena</w>
```

Code	Swedish category	Example	English translation
AB	Adverb	*inte*	Adverb
DT	Determinerare	*denna*	Determiner
HA	Frågande/relativt adverb	*när*	Interrogative/Relative Adverb
HD	Frågande/relativ determinerare	*vilken*	Interrogative/Relative Determiner
HP	Frågande/relativt pronomen	*som*	Interrogative/Relative Pronoun
HS	Frågande/relativt possessivt pronomen	*vars*	Interrogative/Relative Possessive
IE	Infinitivmärke	*att*	Infinitive Marker
IN	Interjektion	*ja*	Interjection
JJ	Adjektiv	*glad*	Adjective
KN	Konjunktion	*och*	Conjunction
NN	Substantiv	*pudding*	Noun
PC	Particip	*utsänd*	Participle
PL	Partikel	*ut*	Particle
PM	Egennamn	*Mats*	Proper Noun
PN	Pronomen	*hon*	Pronoun
PP	Preposition	*av*	Preposition
PS	Possessivt pronomen	*hennes*	Possessive
RG	Grundtal	*tre*	Cardinal number
RO	Ordningstal	*tredje*	Ordinal number
SN	Subjunktion	*att*	Subjunction
UO	Utländskt ord	*the*	Foreign Word
VB	Verb	*kasta*	Verb

Table 12. The 22 part-of-speech categories in SUC. The fairly mnemonic 2-letter-code is followed by the Swedish category name and a typical word where it can apply. English translations of category names in the last column.

In Table 13 below, all the morphosyntactic features used are given along with their possible values. The parts-of-speech to which each feature can be applied are also specified. Several parts-of-speech do not have any morphosyntactic features. Somewhat longer descriptions can be found in Ejerhed et al. (1992).

Feature	**Value**	**Legend**	**Parts-of-speech where feature applies**
Gender	UTR	Uter (common)	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	NEU	Neuter	
	MAS	Masculine	
Number	SIN	Singular	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	PLU	Plural	
Definiteness	IND	Indefinite	DT, (HD, HP, HS), JJ, NN, PC, PN, (PS, RG, RO)
	DEF	Definite	
Case	NOM	Nominative	JJ, NN, PC, PM, (RG, RO)
	GEN	Genitive	
Tense	PRS	Present	VB
	PRT	Preterite	
	SUP	Supinum	
	INF	Infinite	
Voice	AKT	Active	
	SFO	S-form (passive or deponential)	
Mood	KON	Subjunctive (Sw. konjunktiv)	
Participle form	PRS	Present	PC
	PRF	Perfect	
Degree	POS	Positive	(AB), JJ
	KOM	Comparative	
	SUV	Superlative	
Pronoun form	SUB	Subject form	PN
	OBJ	Object form	
	SMS	Compound (Sw. sammansättningsform)	All parts-of-speech

Table 13 . Morphosyntactic features (with 3-letter-values) and the parts-of-speech to which they apply. (Parentheses indicate that a feature only applies to some members of the p-o-s or that not all the values of a feature are applicable.)

3.2.3 Three tags that are not normal morphosyntactic tags: SMS, UO, AN

The tag SMS has a peculiar status. Originally (and as described in Ejerhed et al. 1992) it was meant for the particular forms that a handful of Swedish nouns have in compounds. Historically that is a case form and it can appear in isolation only in connection with conjoined compounds, as in *kvinno- och mansgrupper* 'women's and men's groups'. It is disturbing to call the form *kvinno-* either nominative or genitive and even more disturbing to evoke some obsolete early Swedish case. Instead we chose to indicate an omitted component through a hyphen in a contracted, conjoined compound and call it SMS (for Sw. *sammansättning* 'compound'). The human annotators, however, soon started to use it for all kinds of conjoined compounds, not only for nouns that do not have a special compound form (*café- och biovagn* 'café and movie train', SUC-text HE03) but also in constructions such as *över- och bottenvåning* 'upper and ground floor' (SUC-text KK34) with conjoined non-flexional adverbs. This actually turned out to be smart, as there have occurred, e.g. conjoined compounds involving verbs, where it is not possible to decide which form of the verb is being used (ex. *sov- och liggvagnar* 'wagon-lits and couchettes', lit. 'sleep- and liewagons', SUC-text HE03; *sov-* would, on morphological grounds, otherwise be classified as an imperative or a preterite). The tag SMS is thus allowed in all parts-of-speech on strings that end in a hyphen and are the first part of a compound whose second part comes later.

Examples of SMS-tagged items in the corpus follow.

aa04:

```
<w>skörde-<ana><ps>NN<m>UTR - - SMS<b>skörd</w>
<w>och<ana><ps>KN<b>och</w>
<w>upparbetningsmetoder<ana><ps>NN<m>UTR PLU IND NOM<b>upparbetningsmetod</w>
```

aa10:

```
<w>torsk-<ana><ps>NN<m>UTR - - SMS<b>torsk</w>
<w>och<ana><ps>KN<b>och</w>
<w>plattfiskyngel<ana><ps>NN<m>NEU PLU IND NOM<b>plattfiskyngel</w>
```

hb11:

```
<w>låg-<ana><ps>JJ<m>POS UTR - - SMS<b>låg</w>
<d>,<ana><ps>MID<b>,</d>
<w>mellan-<ana><ps>AB<m>SMS<b>mellan</w>
<w>och<ana><ps>KN<b>och</w>
<w>högstadiet<ana><ps>NN<m>NEU SIN DEF NOM<b>högstadium</w>
```

Foreign words and expressions are surrounded by the SGML-tag <foreign> (cf. 4.3.10) but they also have the part-of-speech tag UO (Sw. *utländskt ord* 'foreign word') without any further subclassification. The same holds for all foreign words, no matter if they can be expected to be wellknown by the reader or not. (Ex. cf03 and cc01 respectively.) Foreign names are not marked <foreign>, only <name>, and all words within such a <name>-tag are tagged PM. This may seem strange in languages where a normal reader easily can tell what is a proper noun and what is not (Ex. cb03) but in order to be consistent across languages and not have the analysis dependent on the annotator's knowledge of foreign languages, we have chosen the more simplistic solution. [3]

cf03:

```
<w n=1538>sant<ana><ps>AB<m>POS<b>sant</w>
<w n=1539>genuin<ana><ps>JJ<m>POS UTR SIN IND NOM<b>genuin</w>
<w n=1540>och<ana><ps>KN<b>och</w>
<d n=1541>"<ana><ps>PAD<b>"</d>
<foreign lang=en>
<w n=1542>basic<ana><ps>UO<b>basic</w>
</foreign>
<d n=1543>"<ana><ps>PAD<b>"</d>
```

cc01:

```
<foreign lang=el>
<w n=2152>kouroi<ana><ps>UO<b>kouroi</w>
</foreign>
```

cb03:

```
<name type=work>
<w n=2129>Howards<ana><ps>PM<m>GEN<b>Howard</w>
<w n=2130>end<ana><ps>PM<m>NOM<b>end</w>
</name>
```

An abbreviation can consist of one or more words. The part-of-speech of an abbreviation is decided from the syntactic function of the expression, not of the single words in it. Abbreviations have AN added as a morphological feature but are also surrounded by <abbr>-tags (cf. 4.3.12). The base form is either the abbreviation itself (cd03) or a spelling out of it (fa05). The choice of base form for each abbreviation is rather unsystematic in SUC, but for all occurrences of an abbreviation the base form is always the same. Initials in proper names are not treated as abbreviations.

cd03:

```
<w n=155>etc<ana><ps>AB<m>AN<b>etc</w>
```

fa05:

```
<w n=94>dvs<ana><ps>AB<m>AN<b>det_vill_säga</w>
```

[3] The attributes *lang* to foreign and *type* to name will be discussed in 4.3.10 and 4.3.6. The values used here are 'en' for an English word, 'el' for a Spanish word and 'work' for the name of a film.

3.2.4 A Comparison between the Part-of-Speech Tags used in SUC and Parole

The overview in Table 14 gives the part-of-speech tags in alphabetic order, the full grammatical term (only in Swedish), the corresponding tag fragment in the Parole system, the relevant SUC morphosyntactic features if any, and sometimes a short description of how the tag is used where this is not felt to be more or less self-evident. As for the values that the <m>-features can take, see Table 13.

Tables for converting entire suctags to corresponding Parole tags and vice versa can be found at http://spraakbanken.gu.se/lb/parole.

Here is just an example. The word *arenor*, which with suctags is:

```
<w n=1414>arenor<ana><ps>NN<m>UTR PLU IND NOM<b>arena</w>
```

(cf. 3.2.2) will with Parole tags be:

```
<w lem='arena' msd='NCUPN@IS' n=1414>arenor</w>.
```

SUC							**Parole**	**Example**
AB	adverb, icke-komparerbart						**RG**	*inte*
	Komparerbart	POS						*fullkomligt*
		KOM						
		SUV						
DT	Determinerare	NEU	SIN	DEF			**D**	*denna*
		UTR	PLU	IND				
		MAS		IND/DEF				
		UTR/NEU						
HA	frågande/relativt adverb						**RH**	*när*
HD	frågande/relativ determinerare	UTR	SIN	IND			**DH**	*vilken*
		NEU	PLU					
		UTR/NEU						
HP	frågande/relativt pronomen	NEU	SIN	IND			**PH**	*som, vad*
		UTR	PLU					
		UTR/NEU						
HS	frågande/relativt possessivt pronomen	DEF					**PE**	*vars*
IE	infinitivmärke						**C (CIS)**	*att*
IN	interjection						**I**	*ja*
JJ	adjektiv	POS	NEU	SIN	IND	NOM	**A**	*glad*
		KOM	UTR	PLU	DEF	GEN		
		SUV	UTR/NEU	SIN/PLU	IND/DEF			

			MAS					
KN	konjunktion						**C (CCS)**	*och*
NN	substantiv	NEU	SIN	IND	NOM		**N**	*hotell*
		UTR	PLU	DEF	GEN			
PC	particip							
	perfekt	PRF	NEU	SIN	IND	NOM	**A (AF)**	*utsänd*
			UTR	SIN/PLU	DEF	GEN		
			UTR/NEU	PLU	IND/DEF			
			MAS					
	presens	PRS	UTR/NEU	SIN/PLU	IND/DEF	NOM	**A (AP)**	*talande*
						GEN		
PL	partikel						**Q**	*under*
PM	egennamn	NOM					**N (NP)**	*Mats*
		GEN						
PN	pronomen	NEU	SIN	IND	SUB		**P**	*hon*
		UTR	PLU	DEF	OBJ			
		UTR/NEU	SIN/PLU	SUB/OBJ				
		MAS						
PP	preposition						**S (SP)**	*av*
PS	possessiva pronomen	NEU	SIN	DEF			**P (PS)**	*hennes*
		UTR	PLU					
		UTR/NEU	SIN/PLU					
RG	grundtal	NOM					**M (MC)**	*tre*
		GEN						
		NEU	SIN	IND	NOM			*ett, en*
		UTR			GEN			
RO	ordningstal	NOM					**M (MO)**	*tredje*
		GEN						
		MAS	SIN	IND/DEF	NOM			*förste, andre*
						GEN		
SN	subjunktion						**C (CS)**	*att*
UO	utländskt ord						**X (XF)**	*the*
VB	verb	KON	IMP	AKT			**V**	*ger*
			INF	SFO				
			PRS					
			PRT					
			SUP					

(Table 14. Caption on next page.)

Table 14. An overview of the SUC word tags. Swedish grammatical terms are given for the p-o-s tags, but not for their relevant morphological features, which were explained in table 13. The two rightmost columns contain the corresponding p-o-s-parts of Parole tags and examples of typical Swedish words where a tag is applicable. [4]

Delimiters also have "part-of-speech" tags that show their function. (Cf. 4.2.2)

SUC		**Parole**	**Example**
MAD	major delimiter	FE	.
MID	minor delimiter	FI	,
PAD	pairwise delimiter	FP	(
PAD	pairwise delimiter	FP	)

[4] For corresponding grammatical terms in English and a full discussion of the morphosyntactic features see Ejerhed et al. (1992).