2 Tagset

(Author: Adam Przepiórkowski; modified: 2 October 2011)

Each morphosyntactic tag is a sequence of colon-separated values, e.g.: *subst:sg:nom:m1* for the segment *chłopiec* 'boy'. The first value, e.g., *subst*, determines the *grammatical class* (cf. §2.2), while the values that follow it, e.g., *sg*, *nom* and *m1*, are the values of grammatical categories (cf. §2.1) appropriate for that grammatical class.

2.1 Grammatical categories

The following table presents the repertoire of grammatical categories used in the National Corpus of Polish:

Number: (2 values)		
singular	*sg*	*oko*
plural	*pl*	*oczy*
Case: (7 values)		
nominative	*nom*	*woda*
genitive	*gen*	*wody*
dative	*dat*	*wodzie*
accusative	*acc*	*wodę*
instrumental	*inst*	*wodą*
locative	*loc*	*wodzie*
vocative	*voc*	*wodo*
Gender: (5 values)		
human masculine (virile)	*m1*	*papież, kto, wujostwo*
animate masculine	*m2*	*baranek, walc, babsztyl*
inanimate masculine	*m3*	*stół*
feminine	*f*	*stuła*
neuter	*n*	*dziecko, okno, co, skrzypce, spodnie*
Person: (3 values)		
first	*pri*	*bredzę, my*
second	*sec*	*bredzisz, wy*
third	*ter*	*bredzi, oni*
Degree: (3 values)		
positive	*pos*	*cudny*
comparative	*com*	*cudniejszy*
superlative	*sup*	*najcudniejszy*
Aspect: (2 values)		
imperfective	*imperf*	*iść*
perfective	*perf*	*zajść*
Negation: (2 values)		
affirmative	*aff*	*pisanie, czytanego*
negative	*neg*	*niepisanie, nieczytanego*
Accentability: (2 values)		
accented (strong)	*akc*	*jego, niego, tobie*
non-accented (weak)	*nakc*	*go, -ń, ci*
Post-prepositionality: (2 values)		
post-prepositional	*praep*	*niego, -ń*
non-post-prepositional	*npraep*	*jego, go*
Accommodability: (2 values)		
agreeing	*congr*	*dwaj, pięcioma*
governing	*rec*	*dwóch, dwu, pięciorgiem*
Agglutination: (2 values)		
non-agglutinative	*nagl*	*niósł*
agglutinative	*agl*	*niosł-*
Vocalicity: (2 values)		
vocalic	*wok*	*-em*
non-vocalic	*nwok*	*-m*
Fullstoppedness: (2 values)		
with full stop	*pun*	*tzn*
without full stop	*npun*	*wg*

2.2 Grammatical classes

The scope of traditional parts of speech such as verb, noun, numeral or pronoun is fuzzy and, hence, controversial. For example, are gerundial forms such as *picie* 'drinking' and *palenie* 'smoking' verbs (they have the category of aspect and they are productively related to verbal forms such as *pić* 'to drink' and *palić* 'to smoke'), or are they nouns (they decline for case, and they have the lexical category of gender)? Are ordinal numerals such as *piąty* 'fifth' numerals (semantically, they are numerals), or are they adjectives (they have adjectival inflection)? Are adjectival pronouns such as *taki* 'such' pronouns (semantics) or adjectives (inflection)?

Grammatical classes used in the National Corpus of Polish are more precisely delimited and, overall, finer-grained than traditional parts of speech. The classes assumed here are based on the notion of *flexeme*, narrower than the notion of *lexeme*.

The following table contains the rough morphosyntactic characteristics of all flexemic classes assumed in the present tagset. The symbol ⊕ in the table means that, for a given flexemic class, a given grammatical category is a morphological category (flexemes belonging to this class normally inflect for that category), while the symbol ⊙ means that the category is a lexical category (for each flexeme belonging to this class, all forms of that flexeme have the same value of that category, although that value may differ between flexemes, as in the case of the gender of nouns).

	number	case	gender	person	degree	aspect	negation	accentability	post-prep.	accom.	agglt.	vocalicity	fullstop.
noun	⊕	⊕	⊙										
depreciative form	⊙	⊕	⊙										
main numeral	⊙	⊕	⊕							⊕			
collective numeral	⊙	⊕	⊙							⊕			
adjective	⊕	⊕	⊕		⊕								
ad-adj. adjective													
post-prep. adjective													
predicative adjective													
adverb					⊕								
pronoun (non-3rd person)	⊙	⊕	⊕	⊙				⊕					
pronoun (3rd person)	⊕	⊕	⊕	⊙				⊕	⊕				
pronoun SIEBIE		⊕											
non-past form	⊕			⊕		⊙							
future BYĆ	⊕			⊕		⊙							
agglut. BYĆ	⊕			⊕		⊙						⊕	
l-participle	⊕		⊕			⊙					⊕		
imperative form	⊕			⊕		⊙							
impersonal form						⊙							
infinitive						⊙							
adv. contemp. prtcp.						⊙							
adv. anter. prtcp.						⊙							
gerund	⊕	⊕	⊙			⊙	⊕						
adj. act. prtcp.	⊕	⊕	⊕			⊙	⊕						
adj. pass. prtcp.	⊕	⊕	⊕			⊙	⊕						
winien-like verb	⊕		⊕			⊙							
predicative													
preposition		⊙											
coord. conjunction													
subord. conjunction													
particle-adverb													
abbreviation													⊕
bound word													
interjection													
punctuation													
alien													
unknown form													

The following table provides the information about base forms for all grammatical classes, as well as the abbreviations of these classes as used in the National Corpus of Polish.

flexeme	**abbreviation**	**base form**	**example**
noun	*subst*	singular nominative	*profesor*
depreciative form	*depr*	singular nominative form of the corresponding noun	*profesor*
main numeral	*num*	inanimate masculine nominative form	*pięć, dwa*
collective numeral	*numcol*	inanimate masculine nominative form of the main numeral	*pięć, dwa*
adjective	*adj*	singular nominative masculine positive form	*polski*
ad-adjectival adjective	*adja*	singular nominative masculine positive form of the adjective	*polski*
post-prepositional adjective	*adjp*	singular nominative masculine positive form of the adjective	*polski*
predicative adjective	*adjc*	singular nominative masculine positive form of the adjective	*zdrowy, ciekawy*
adverb	*adv*	positive form	*dobrze, bardzo*
non-3rd person pronoun	*ppron12*	singular nominative	*ja*
3rd-person pronoun	*ppron3*	singular nominative	*on*
pronoun SIEBIE	*siebie*	accusative	*siebie*
non-past form	*fin*	infinitive	*czytać*
future BYĆ	*bedzie*	infinitive	*być*
agglutinate BYĆ	*aglt*	infinitive	*być*
l-participle	*praet*	infinitive	*czytać*
imperative	*impt*	infinitive	*czytać*
impersonal	*imps*	infinitive	*czytać*
infinitive	*inf*	infinitive	*czytać*
contemporary adv. participle	*pcon*	infinitive	*czytać*
anterior adv. participle	*pant*	infinitive	*czytać*
gerund	*ger*	infinitive	*czytać*
active adj. participle	*pact*	infinitive	*czytać*
passive adj. participle	*ppas*	infinitive	*czytać*
winien	*winien*	singular masculine form	*powinien, rad*
predicative	*pred*	the only form of that flexeme	*warto*
preposition	*prep*	the non-vocalic form of that flexeme	*na, przez, w*
coordinating conjunction	*conj*	the only form of that flexeme	*oraz*
subordinating conjunction	*comp*	the only form of that flexeme	*że*
particle-adverb	*qub*	the only form of that flexeme	*nie, -że, się*
abbreviation	*brev*	the full dictionary form	*rok, i tak dalej*
bound word	*burk*	the only form of that flexeme	*trochu, oścież*
interjection	*interj*	the only form of that flexeme	*ech, kurde*
punctuation	*interp*	the only form of that flexeme	*;, ., (,]*
alien	*xxx*	the only form of that flexeme	*cool , nihil*

unknown form	*ign*	the only form of that flexeme	