

Specification for morphosyntactic tagging of UGC texts

<u>Principles of tagging</u>	<u>2</u>
<u>Nouns</u>	<u>4</u>
<u>Numerals</u>	<u>4</u>
<u>Adjectives</u>	<u>5</u>
<u>Adverbs</u>	<u>7</u>
<u>Pronouns</u>	<u>7</u>
<u>Verbs</u>	<u>8</u>
<u>Prepositions</u>	<u>9</u>
<u>Conjunctions</u>	<u>10</u>
<u>Burks</u>	<u>10</u>
<u>Interjections</u>	<u>10</u>
<u>Qubs</u>	<u>11</u>
<u>Abbreviations</u>	<u>11</u>
<u>Foreign items (aliens)</u>	<u>12</u>
<u>Specific solutions</u>	<u>13</u>
<u>Proper names</u>	<u>13</u>
<u>Graphemic representations</u>	<u>13</u>
<u>Special expressions (internet)</u>	<u>13</u>
<u>Symbols</u>	<u>14</u>
<u>Spelling errors</u>	<u>15</u>

Principles of tagging

Given the adopted set of tags and selected tools, our principles of tagging are based on the ones used in the course of annotation of the National Corpus of Polish (NKJP¹). To facilitate the tagging process, the following subsections related to individual parts of speech contain extensive fragments of the aforementioned book, supplemented with additional explanations or comments, if necessary.

The following table provides the information about base forms for all grammatical classes, as well as the abbreviations used for these classes as used in the NKJP (see: <http://nkjp.pl/poliqarp/help/ense2.html>).

grammatical class	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

¹ See: Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski i Barbara Lewandowska-Tomaszczyk (eds) *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw, 2012.

future być	bedzie	infinitive	być
agglutinate być	aglt	infinitive	być
I-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	

Nouns

Nouns (subst) are those flexemes that are inflected by case and – with the exception of pluralia tantum (e.g., *spodnie, wujostwo*) and singularia tantum (e.g., *co, to*) – by number ;

As for the grammatical category of gender, every noun is assigned to a gender value. In order to help assign a particular value of gender, the following contexts were used as tests:

Widzę jednego ... z tych, których lubię.	m1
Widzę jednego ... z tych, które lubię.	m2
Widzę jeden	m3
Widzę jedno	n
Widzę jedną	f

For pluralia tantum, the forms that combine with masculine verb forms in the past tense were treated as nouns of m1 gender (e.g., *wujostwo*), while the remaining ones were treated as nouns of n gender (e.g., *skrzypce, sanie, pomyje*).

It is also worth mentioning that the value of gender of segments syntactically dependent on the noun was assigned in accordance with the value of the noun's gender, e.g., in the sentence *Trzy małe pieski spały* all four forms are assigned to m2 gender.

Numerals

Cardinal numerals (num) were singled out as a class of flexemes with a fixed number (always plural) and inflected by case and gender. Therefore, this class does not include traditional numerals inflected by number, i.e. ordinal (*drugi, tysięczny*), multipliers (*dwukrotny, tysiąckrotny*), etc. - we consider them as adjectival lexemes. In addition to the typical cardinal numbers (*dwa, pięć, dziewięćset*), the following were also considered as cardinal numbers: *oba, obydwa*, because they inflect similar to the numeral *dwa*. All these numerals can be called proper numerals. The cardinal numerals, with slightly different syntactic characteristics, and therefore incorrect, were considered to be flexemes such as *tyle, ile, wiele* (a także *więcej i najwięcej*), *kilka*, because they inflect similar to the numeral *sto*, as well as defective flexemes such as *dużo, mało* (and also *mniej, najmniej*) and *trochę*.

Numerals also have a grammatical category of accommodation, which determines the type of relationship between a numeral form and a noun form. Based on the value of this category, word forms of numeral flexemes can be divided into two accommodation classes. The first includes forms that combine with nouns with a different case - the governing forms *governing*; the second one includes those that combine with nominal forms of the same case value - the agreeing forms *agreeing*.

Collective numerals (numcol) were grouped into a separate class as they constitute distinct flexemes of numeral lexemes. Collective numeral flexemes have a fixed value of the number category (always pl), and a gender category value of n or m1, depending on the gender of the noun, e.g. *pięcioro dzieci, pięcioro skrzypiec* (both n), *dwoje wujostwa* and *sześcioro studentów* (both m1).

Similar to cardinal numerals, collective numerals have the category of accommodation, but the distribution of governing and agreeing forms is slightly different in this case. Numerals

written in digits are treated as cardinal, ordinal or collective numbers: an interpretation depends on reading a given digit (or of given digits) in context.

Adjectives

There are 4 classes of adjectival flexemes (adj, adja, adjc and adjp). Adjectives (adj) inflect by number, case, gender and (not all) by degree. The degree of non-gradable adjectives was determined as equal (pos). Adjectival adjectives (adja) and predicative adjectives (adjc) are classes of uninflected and dependent flexemes, occurring only in specific constructions. Adjectival adjectives are flexemes containing forms such as *polsko* in *polsko-niemiecki*. Predicative adjectives are a small, closed class of flexemes such as *zdrów*, *ciekaw* or *godzien*; they are characteristic in that they occur only in predicative constructions (e.g., *jestem ciekaw*, but not **ciekaw człowiek*). Let's look at some examples of tags:

Przyszedł do mnie pewien pan. – *pewien* adj:sg:nom:m1:pos (=jakiś).

Nie jestem pewna tego. – *pewny* adj:sg:nom:f:pos.

Jak możesz być pewien Marka? – *pewny* adjc.

Jak możesz być pewny Marka? – *pewny* adj:sg:nom:m1:pos.

Post-prepositional adjectives (adjp) are also a class of uninflected and dependent flexemes, occurring only in specific (although not uncommon) constructions, i.e. in prepositional phrases such as, for example, *bez mała*, *co gorsza*, *do późna*, *od niedawna*, *na prawo*, *po ludzku*, *po prostu*, *po swojemu*, *po wojskowemu*, *w lewo*, *za młodu*, *z bliska*, *z rzadka*.

Supplementary information:

It is often difficult to decide whether a given form is an adjective or a participle (notably active participle)

We can use the following hints:

- Słowność (plWordNet): does such a lemma even exist as a "qualitative adjective" and does it occur as a qualitative adjective in the sense found in a sentence under scrutiny
- Paraphrase: if you can paraphrase it in the form: '*taki, który* + verb from which the adjective/participle form is derived', then we are (rather ;-)) dealing with a participle
- structures with active participle built according to the scheme: *on* + *jest* + ..., e.g.. *on jest piszący*, are impossible (with some exceptions)

A fragment of instructions for the adjective in Słowność (plWordNet):

How to distinguish between an adjective and a participle? Participles denote actual situations, i.e. when something happens at a certain moment / interval of time, while adjectives denote features of entities that are in a sense permanent or constant. This is clarified by the following principle:

Temporal expressions, e.g., *teraz*, *w tej chwili*, *wczoraj* or *na długo*, combine with participles, but not with gerunds. We can verify combination patterns of these expressions with the tested words (meanings/senses).

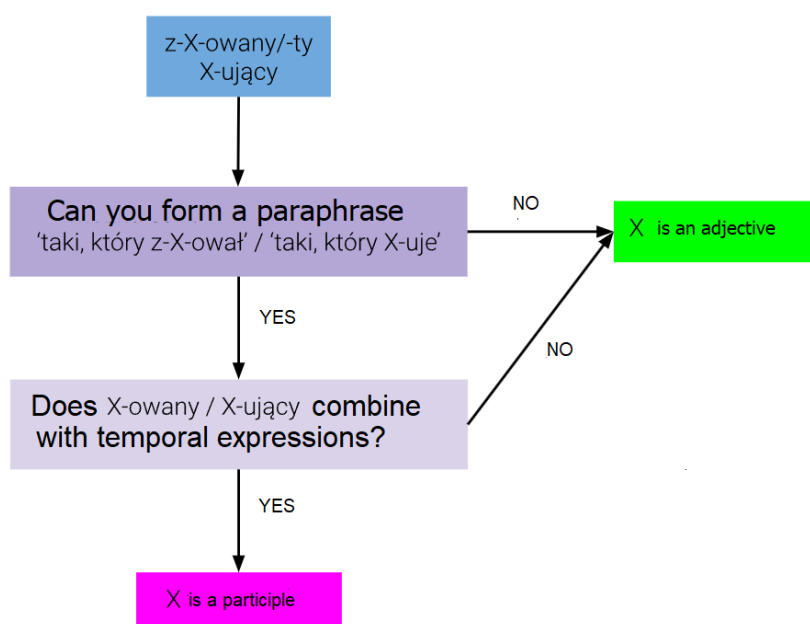
In the course of the test, it is possible to verify paraphrases of linguistic expressions that contain words suspected of being participles, where they were replaced with the verb from which they were formed. Verbs express actual situations, so it should be possible to substitute a participle with a verb and in this way create a correct sentence with one of the aforementioned adverbs. Remember that the meaning of the tested word can not change when paraphrased according to the following scheme:

z-X-owany / z-X-ty = taki, który ktoś z-X-ował
X-ący = taki, który X-uje

In the case of paraphrases, any semantic addition indicates lexicalization of the participle and its transition to the class of adjectives.

The above principle is particularly useful in the case of perfective participles. For example, *otwarty* = *taki, który ktoś otworzył*. Thus, *otwarty kubeł* = *kubeł, który ktoś otworzył*. We can also connect the open participle with temporal expressions: *kubeł był wczoraj otwarty*. Meanwhile, we have such uses of the word *otwarty* that cannot be connected with the verb using the above paraphrases: we have, for example, *pojemnik otwarty* = *taki, który nie ma wieczka, pokrywki*.

From among the two rules described above, we give priority to the first one, i.e. the one related to temporal expressions. The testing scheme should therefore look as follows:



Obviously enough, the participle should look like a participle, i.e. it can only end in *-ny*, *-ty*, *-ące* (in singular, masculine).

Adverbs

Adverbs (adv) belong to a class of flexemes that, for the most part, inflect by degree and/or are derived from an adjective. Thus, adverbs are properly gradable words, e.g. *cicho*, *ciszej*, *najciszej*, and descriptively gradable, e.g. *gorzko*. This class also includes gradable words not necessarily derived from adjectives: *bardzo*, *bardziej*, *najbardziej*. The degree of non-gradable adjectival adverbs, similarly to the corresponding adjectives, is set as equal (adv:pos). In addition, this class includes other (non-gradable, non-adjectival) lexemes traditionally regarded as adverbs, e.g., *gdzie*, *gdzieś*, *wczoraj*, *tam*, *wraz*. Such adverbs were tagged as (adv) without specifying their degree value.,

The criterion distinguishing adverbs from qubs was primarily a distributional one. Adverbs cannot directly modify nouns (in contrast to gerunds, which can do so) or proper numerals (but they can do so incorrectly, e.g. *bardzo* in *bardzo dużo wyjątków*): they usually modify verb forms, verbal forms (gerunds), adjectives and adverbs.

Pronouns

The class of non-third person pronouns (ppron12) includes four flexemes: *ja*, *my*, *ty*, *wy*. Each element of this class has a fixed number (sg for *ja* and *ty*, pl for *my* and *wy*) and person (pri for *ja* and *my*, sec for *ty* and *wy*), but it is inflected by case, gender (without differentiation of forms) and accentuation (in the case of those forms for which the opposition of accentuation exists).

The third person pronoun (ppron3) is a one-element inflectional class containing the flexeme with the base form *on*. This flexeme has a fixed person (ter) and is inflected by case, gender and number as well as by the category of accentuation and postposition, e.g.,:

<i>jego</i>	<i>on</i> ppron3:sg:gen.acc:m1.m2.m3:ter:akc:npraep
<i>go</i>	<i>on</i> ppron3:sg:gen.acc:m1.m2.m3:ter:nakc:npraep
<i>niego</i>	<i>on</i> ppron3:sg:gen.acc:m1.m2.m3:ter:akc:praep
<i>-ń</i>	<i>on</i> ppron3:sg:gen.acc:m1.m2.m3:ter:nakc:praep

The gender of the pronoun ppron12 and ppron3 was tagged as the gender of the noun to which it refers. For example, in the sentence *Mam dość tego kota i jego głupoty* the segment *jego* was tagged as m2 (not: f!). Thus, these pronouns differ from the traditionally understood first and second person possessive forms, which are treated here as adjectival forms, e.g. *moje* in *moje palto*. Such possessive forms agree in gender with the modified noun. Hence, traditional possessive pronouns break down into adjectives (*mój*, *wasz* and other 1st or 2nd person possessive pronouns) and personal pronouns in the genitive (3rd person forms: *jego*, *jej*, *ich*).

There is also a one-element class *siebie*: its only component is the flexeme *siebie*, which is inflected only by case, and it contains three forms with the following possible values of this grammatical category:

1. *siebie*: *siebie*:acc.gen

2. *sobie*: siebie:dat.loc
3. *sobą*: siebie:inst.

Supplementary information:

In principle, as in the case of tagging the National Corpus of Polish (NKJP), practice does not fully align with theory. We should rather stick to the latter, because our tools should be compatible with it. The mechanism recently introduced to Inforex enables one to omit both the accentuation and post-preposition categories. These two categories should only be used in the genitive, dative and accusative case of the pronoun *on*:

<http://sgjp.pl/leksemy/#90012/on>

Verbs

The class of flexemes inflected by number and person includes the following:

- finite non-past forms – fin;
e.g., *zjemy* tagged as *zjeść:fin:pl:pri:perf*, *mówię* tagged as *mówić:fin:sg:pri:imperf*, *jest* tagged as *być:fin:sg:ter:imperf*;
- future forms of *być* – *bedzie*;
e.g., *będą* tagged as *być:bedzie:pl:ter:imperf*;
- agglutinative forms – aglt;
e.g., *m* tagged as *być:aglt:sg:pri:imperf:nwok*, *ście* tagged as *być:aglt:pl:sec:imperf:nwok*;
- imperative forms – impt;
e.g., *zabijmy* tagged as *zabić:impt:pl:pri:perf*, *twórz* tagged as *tworzyć:impt:sg:sec:imperf*.

As a rule, imperatives have a paradigm limited to the forms of the 2nd person singular and the 1st and 2nd person plural. Agglutinative forms (aglt) include *-(e)m*, *-(e)s*, *-(e)śmy*, *-(e)ście*. The presence of the letter *e* in the aforementioned segments enables one to distinguish vocalic forms of agglutinants. The value of their aspect category is always *imperf*. It is worth noting that they can be attached not only to pseudo-participles (or to elements of the *winien* class), but also to forms from other classes, e.g. *Głupiś*, *Abyśmy wszyscy zdrowi byli*, *Pięknieś go podsumował*. Such words do not have features of agglutinative forms and are tagged separately (as adjectives etc.) from the agglutinative forms that follow them. Traditionally understood forms of the past tense and conditional mood are divided into the form of the pseudo-participle or the *winien* class, the *qub* *by* and the agglutinative form of the verb *być*. For example, the word *kupiłbym* is treated as follows:

kupił *kupić:praet:sg:m1:perf*,
by *by:qub*,
m *być:aglt:sg:pri:imperf:nwok*

Gerunds (*ger*) are verb forms inflected by case and, in some cases, by number. The problem of distinguishing between gerunds from nouns has not been fully solved, yet there are certain criteria that points towards interpreting contested forms as gerunds:

co-occurrence of arguments "inherited" from the verb, e.g., *pomaganie mu, więzienie go, czy twierdzenie, że...*;

co-occurrence of typical adverbials of time, e.g., *bieganie (przez) 2 godziny*;

adverb modification, e.g., *robienie czegoś szybko*;

presence of agent of an activity in a phrase such as, for example, *przejęcie władzy przez rebeliantów*.

Since we consider gerunds and participles (i.e. contemporary and anterior adverbial participles as well as active and passive adjectival participles) as verb forms, they have the category of aspect, and their basic form is the infinitive. The forms of contemporary and active participles (pcon and pact) are characterized by an imperfective aspect (e.g. *pisząc, piszący*), while anterior participles (pant) are characterized by a perfective aspect (e.g. *napisawszy*). Passive participles and gerunds can have any value of the aspect category (*pisany/napisany, pisanie/napisanie*).

Other flexemes include the impersonal forms (imps), e.g., *pisano, pito*, and the infinitive forms (inf), e.g., *pisać, pić*. These are not inflected verbal classes as they only have the dictionary category of aspect.

Pseudo-participles (praet), e.g., *pisał, piła*, inflect by number and gender, but not by person. Some forms of some pseudo-particles also have a morphological category of agglutination, e.g. *nieść*:

niosł- agl

niósł nagl

However, the forms which are not distinguished with respect to the possibility of adding an agglutinative form, e.g., *niosła* or *czytał*, do not have the category of agglutination.

The verbs *powinien* and *winien* have untypical inflection, which is why such flexemes belong to a separate class *winien*. The elements of these flexemes function as present tense forms, although they have a structure and features similar to pseudo-participles (yet no possible differentiation due to agglutination is visible in their case). Personal forms are created by adding agglutinative forms. The past tense and conditional forms are analytical.

In addition, a separate class of predicatives (pred) was introduced. This class consists of, among others, the flexemes *brak, trzeba, czas, warto, słyszać, widać, strach* etc., which inflect only analytically (e.g. *było warto, warto, warto by, będzie warto*). The word *to* was also categorized as a predicative (e.g. in the sentence *Ala to dziewczyna*).

Prepositions

Prepositions (prep) are uninflected flexemes which function as a linking device and require a specific case, as indicated in the preposition tag, e.g. prep:dat for the form *ku* in the expression *ku niemu*. The basic test for prepositionality of uninflected flexemes is their possible combination with post-prepositional forms (praep) of pronouns: if the forms of a given flexeme can be combined with post-prepositional forms of pronouns, then we deal with a prepositional flexeme, e.g., *naprzeciwko nich*, a nie **naprzeciwko ich*.

Some prepositions, e.g. *w, z* and *przed*, take the values of the "grammatical category" of vocality, e.g., the tag for *we* in the expression *we wtorek* is prep:acc:wok.

As a rule, prepositions combine with nouns, adjectival forms (*Wyglądał na zmęczonego*) or numerals (*Przyszli we trzech*). It has also been agreed that prepositions can combine with segments tagged as adverbs, e.g., *na pewno*, *co najmniej*, *na trzeźwo*, *na zawsze*, or *Pomalował dom na brązowo* and *Ugotował jajko na twardo*.

Conjunctions

Coordinating conjunctions (conj) include uninflected flexemes which function as a linking device and do not require a specific case; they connect elements with equivalent functions in a sentence, e.g., *i*, *lecz*, *oraz*, *ani*, *więc*. The same tag was also used for the components of discontinuous conjunctions, where each segment was tagged as conj. Here are two examples of such discontinuous conjunctions:

nie tylko. . . , lecz/ale także/też/również/i. . . ;
(zarówno). . . , jak i/też/również. . .

In addition, a class of subordinating conjunctions (comp) was distinguished. It includes uninflected items which function as a linking device, do not require a specific case, and introduce a subordinate clause, e.g., *że*, *aby*, *bowiem*.

Burks

Burks (burk) are uninflected items whose distribution is limited to the close vicinity of another specific segment, e.g., *trochu* in *po trochu* (but not in *prostu* or *polsku* in *po prostu* or *po polsku* – since adjectives *prosty* and *polski* exist, they were tagged as adjp), *naprzeciwka* in the expression *z naprzeciwka*, *oścież* in *na oścież* etc., but also – as an exception – *zamian* in the expression *w zamian*, although a nominal segment *zamian* also exists (yet it has a different case than the ones required by the preposition *w*). This is a relatively small class of lexemes; parts of proper names do not belong to it. For example, both *Burkina* and *Faso* in the sentence *Zobaczyłem na własne oczy Burkina Faso* were tagged as subst:sg:acc:n (with the head forms *Burkina* and *Faso*). Most burks are in fact phrases with an isolated segment, which used to be a fully-fledged noun (e.g. *po omacku*, *z oddali*, *na pohybel*), although this does not have to be the rule (cf. *z kretesem*, where *kretes* has not existed in the Polish language).

Interjections

Interjections (interj) have been understood very narrowly for the reason that in fact any word can be used with a pragmatic function of an interjection. Thus, interjections include the following:

- those segments that can only be used in such a function, e.g., *ach*, *och*, *oj*, *psiakrew*;
- as an exception, forms which can be interpreted as having functions unrelated to interjections, e.g., *a* (it can also be a conjunction or an abbreviation), but not *tak*, *stop* or *cholera*;

- onomatopoeic segments, such as *muu* i *kukuryku*.

The basic form of interjections is the form from the text, which also applies to words used as interjections, e.g., *Pomocy! Ratunku!*, where the base form is [pomocy] and [ratunku] respectively.

Qubs

Qubs (qub) include a semantically and distributionally inconsistent class of uninflected lexemes (except for particles *-ż/-że, z/ze* „inflected” by vocality). Qubs are mainly those segments which can modify words of different classes, including nouns. So *nawet* is a qub because the sequences *nawet on, nawet po pijaku, nawet wczoraj, nawet poprosił, nawet biały* etc. are correct; however, *bardzo* is not a qub because it does not modify nouns. It was precisely the criterion of noun modification that made it possible to distinguish between qubs and adverbs, also in cases where the morphological form of the word would rather indicate an adverb. That is why the word *głównie* was tagged as a qub, cf. *Lubi czytać książki, głównie te z górnej półki / po niemiecku / dobre / kryminały*. The correct uses of the word *prawdopodobnie* were also consistently distinguished: we could have a gradable adverb here (= *wiarygodnie* ‘credibly’), cf. *Zabrzmiało to dość prawdopodobnie. Zabrzmiało to prawdopodobniej, niż wczoraj. Zabrzmiało to najprawdopodobniej ze wszystkich twoich wypowiedzi.*, but a qub will occur more often (= *przypuszczalnie, zapewne* / ‘supposedly, probably’), cf. *Zrobił to prawdopodobnie Marek. Spotkamy się prawdopodobnie wiosną.*

The following words, among others, were also tagged as qubs:

- *się* (all occurrences);
- *niech* (all occurrences);
- *nie* (as negation);
- partykuły *-li, -ż/-że, -by*;
- modifiers of quantity with *z* (*z 5 kilo, około (około tysiąc osób przyszło), koło (koło tuzina), gdzieś (gdzieś ze dwieście od razu wyszło), bodaj, aż, dokładnie, ponad* itp.);
- segments such as *za, zbyt, dość, dosyć* etc. in contexts *za/zbyt/dość/dosyć dużo/często/wysoki*;
- all uses of a segment *już* (*Studia już skończyłem, 4 tysiące akapitów to już jest coś; Już, już!*) and *jeszcze* (*To jeszcze nie koniec, Jeszcze czego!*).

Abbreviations

Abbreviations (brev), but not acronyms, were treated as a separate grammatical class. For this class, the special category of requiring/non-requiring a dot was proposed, with tags such as pun (requires a dot, e.g., for *dr* in the sentence *Widziałem się z dr. Kowalskim*) and npun (dot not required), e.g., in the sentence *Dr Kowalski przybył*.

Information about whether a dot is required or not is independent from whether it actually occurs in the text. For example, in the sentence *To są różne zwierzęta, np konie.* the abbreviation *np* should be tagged as brev:pun.

Generally, abbreviations differ from acronyms: when reading the text we use the abbreviations’ full forms (acronyms are read by “spelling” individual letters), e.g., “dr” (abbreviation) will be read as “doktor”, but “USC” (acronym) will be read as “u es ce”., among other differences.

Foreign items (aliens)

Foreign items (xxx) are segments of foreign origin which are parts of longer quotes or utterances, and therefore they do not interact directly with Polish segments in the text. For example, each of the first three segments in the sentence *My czeczinskije bojewiki – mówią o sobie z dumą dwaj 13-latkowie* was tagged as xxx.

However, if they were single words or very short sequences of foreign words interacting grammatically with Polish words, then they would be annotated like Polish words: all segments in this short sequence would be given the same tag, namely the one appropriate for a given position in the sentence and for the main element of that group of foreign items. The head forms of such segments were the same segments.

In case of any doubts related to gender, an arbitrary decision to use the tag m3 was made.

For example, in the sentence *Kupiła Washington Post*, both *Washington* and *Post* are tagged as subst:sg:acc:m3. Obviously enough, such shorter sequences could belong to different grammatical classes, cf. *warunkiem sine qua non jest ustalenie* – adj, *metoda in vitro* – adj, *zapłodnić in vitro* – adv, *ocenić in plus* – adv, *różnice in plus* – adj.

Apart from compound conjunctions, this was the second case where we went beyond one segment and annotated segments forming a larger whole using a single tag.

Specific solutions

Proper names

In the case of proper names, our automatic tagger often cannot indicate the correct base form and instead assigns the tag (ign). In such situations, the task of the annotator is to select "Add interpretation option" and enter the correct base form and appropriate tags. We use capital letters then, e.g.,

Citibanku → Citibank

mBankiem → mBank

When determining the type of foreign proper names, we can take into consideration language use (also in the original language) and applicable general rules:

- Based on the ending of the name, e.g. the consonant ending of the word is typical of masculine gender in Polish, hence *Guiyang* is inflected as, for example, *Elk*. (Jan Grzenia)
- Based on context, which is a common word with a generic meaning, e.g. *Hefei* [pronounced in Polish: hefej] may be masculine (by analogy to *Poraj*), but since the ending poses more difficulty we can assume that the context (the city of Hafei) implies a neuter gender. (Jan Grzenia)
- It is customary to leave in the nominative form in the neuter gender (as neutral) those names whose ending makes it impossible or difficult to match them with a particular inflectional paradigm, e.g., *Fidži*, *Peru*, *Kobe*, *Kioto*, *Vincennes*, *Essen*, *Baltimore*. (Adam Wolański)

Graphemic representations

Words written in capital letters (e.g., for emotional reasons) and which are not proper names are reduced to lowercase letters in the course of lemmatization (i.e. when we introduce the base form), e.g.,

CZŁOWIEKU → człowiek

Special expressions (internet)

On the Internet, we can find special expressions which require specific solutions. The first group is emoticons. We use the [interj] tag to annotate them. If the tokenizer divided the emoticon into smaller fragments, we use the [interj] tag for the first segment, and [blank] tag for subsequent segments, e.g.,

: -)

- : ← interj | base form is :
- - ← blank | base form is -
-) ← blank | base form is)

The second group includes:

1. links
2. nicknames
3. tags
4. fragments of programming code

In their case, we determine the number and gender in an arbitrary manner (singular, masculine inanimate gender (m3)), and for the grammatical case we conduct a contextual grammatical analysis and make a decision regardless of the form (which is most often the base form), e.g.,

Czy są tu jacyś fani #bieganie albo #rower?

bieganie - [subst:sg:**gen**:m3]

Symbols

Symbolic product markings occur especially in texts about electronics, e.g. Nokia **N 80**, Sony **E100**...

We use somewhat conventional solutions in their tagging, which consists in the following:

- base = form found in the text (e.g., N, 80, E100...)
- grammatical class = subst (exception: num for tokens that consist only of digits)
- number = sg (exception : digital symbol - it takes the value pl (e.g., 80))
- the same case as in the modified fragment (e.g., Nokia, Sony...)
- gender = n (exception: the same gender as in the head element for tokens that are all digits)

Spelling errors

We correct the words by tagging them as if there was no error. This mainly applies to typos, the absence of Polish characters etc., which follows that we deal with words that are not recorded in dictionaries, e.g., *zo/w* (instead of *żółw*), *zonaczony* (instead of *oznaczony*).

One of the errors of normalization involves not replacing the so-called "dotted" word with the full form, e.g., leaving the form "k...a" instead of providing the full form. Such tokens are "glued" when tagging: we use the (blank) tag and, at the same time, we provide the full base form and conduct morphosyntactic interpretation of the first token, e.g.,

k [kurwa - interj]

. [. - blank]

. [. - blank]

. [. - blank]

a [a - blank]