Morphological Analysis of the Slovak National Corpus

Lucia Gianitsová
Department of the Slovak National Corpus, Linguistics Institute of Ľ. Štúr
Slovak Academy of Sciences, Bratislava

1. Basis of a morphological analysis of the Slovak National Corpus

A question of morphological (or morphosyntactic) analysis has been a key problem for natural language processing (NLP) for several years. Automatic morphological annotation is a useful tool especially with regard to the corpus data processing. In this respect morphological annotation has been considered also during the development of the Slovak National Corpus (SNC). Theoretical aspects of morphological analysis and its application in corpus tagging associated with the morphological tagset preparation for the manual tagging of SNC were outlined by M. Forróová and A. Horák (2003, in press). Annotation, generally understood as the process of adding some information to texts, is undoubtedly a convenient tool – in spite of different views – in verification of linguistic theories, but also in carrying out various lexicographic projects.

Morphological analysis is generally understood as an assignment of a base form (lemmatization), classification of words into grammatical and semantic classes and assignment of grammatical categories to words in texts (in a form of tags). Generally, for a language-competent person this kind of analysis is not difficult, on the other hand for computer processing it is a hard-nut task (Foróová – Horák, 2003). At the same time we take into account the problem of formal homonymy resulting from automatic morphological annotation; this problem requires a subsequent disambiguation. From the very beginning we were aware of the fact that a set of morphological tags should have represented language properties of a text, in other words, somehow it should have interpreted a text. We should decide whether the proposed tagset would result in a new formal description of language, or if we reflect to the current existing linguistic descriptions, and tried to formalise them. Foróová and Horák (2003) point to 7 maxims proposed by G. Leech, providing regularity of annotation, and guarantee that annotation wouldn't result in misinterpretations of corpus data. We would like to emphasize a maxim of accessibility and a maxim of consensus of academic theories, theoretical "neutrality" which was predominantly taken into consideration when preparing a tagset.

The central issue can be formulated as follows: to which extent we can refer to the traditional grammatical descriptions of Slovak morphology when preparing a lemmatization and a tagset? We considered relevant to take into account a systemic description made by academic *Morfológia slovenského jazyka* (lit. Morphology of the Slovak language, MSJ, 1966)**,** eventually some other works dealing with morphology (Oravec – Bajzíková – Furdík, 1984; Dvonč, 1984).

The conflict between representatives of traditional grammatical categories and a possibility of automatic language processing is reflected also in approach of morphological tagset for SNC. M.

Foróová and A. Horák (2003) have already pointed to vagueness of criteria for morphological classification. They brought attention to a complex of morphologic, syntactic, lexical and semantic properties of words serves as a criterion for their classifying into parts of speech (MSJ, 1966; Oravec – Bajzíková – Furdík, 1984), and within the framework of morphology linguists traditionally point to the lexical-grammatical categories such as intention, aspect or grade**.**

At the level of automatic processing this approach brings out numbers of difficulties. Obviously, an alternative approach could be applied. In the field of corpus linguistics it is possible to observe and analyse various approaches to morphological annotation operating in text annotations of national corpora. M. Forróová and A. Horák (2003) were influenced by considering advantages and disadvantages of morphological tagsets of the Czech National Corpus (CNC), Multext-East Project and a corpus built in the Institute of Computer Science of the Polish Academy of Sciences (IPI PAN). They conclude that SNC annotation should not be based only on a formal approach. This approach led authors of the CNC tagset to specify actually 74 possible SubPOS values including 19 pronoun values (Hajič, 2000; Hana – Hanová, 2002; Forróvá – Horák, 2003). Similarly in IPI PAN tagset there are 29 grammatical classes (for details cf. Forróová – Horák, 2003).

When comparing the above-mentioned approaches with tagset types (linguistically optimistic tagset types and linguistically pessimistic tagset types)[1] specified in Forróová – Horák (2003), for the purpose of tagging in SNC the formal-grammatical principle was designated. However, this principle is characterized by some specific features with regard to the domestic Slovak linguistic tradition.

A complete morphological tagset proposal was introduced on October 24, 2003 in Bratislava[2]. Later on it was disputed on November 10, 2003.[3] Moreover, then the concept of morphological annotation has been subject to several changes and attesting its validity. After a manual annotation of the first text samples a need to re-value some parts of the tagset arose. We took into account tagging of real texts in the SNC database and requirements and demands of real corpus users (including possible users). Current tagset version can be found on the SNC website.[4] In the following parts we will point out only some general features of morphological annotation of SNC and explain some particular problems and their possible solutions, influenced also by approaches to text tokenization. That is why at first a brief attention should be paid to the problem of tokenization.

2. Tokenization as a basis for morphological annotation of SNK

[1] According to Foróová and Horák (2003), linguistically optimistic tagset types include implementation of a maximum number of grammatical categories, disambiguation based on syntactic rules, this approach is represented by e.g. V. Petkevič and K. Oliva), on the other hand linguistically pessimistic tagset types represent a compromise between linguistic and „engineering-based" approach, eventually accommodation of tagset contents to the mathematical model of a tagger: representatives are Hajič`s tagset and tagset of Multext-East Project.

[2] Contribution to the international conference *Slovko – Slavic languages and their computer processing*; see Forróová – Garabík – Gianitsová – Horák – Šimková, 2003.

[3] Internal seminar meetings in SNC, see http://korpus.juls.savba.sk/aktivity/archiv/.

[4] http://korpus.juls.savba.sk/korpus/biblioteka/publikacie/Tagset-aktualny.pdf

Approaches to morphological annotation as well as a tagset proposal are derived especially from approaches to tokenization, e.g. identification of the smallest text units (tokens) which are not equivalent with neither words nor grammatical forms. Tokens are usually defined as chains of characters between two spaces. This concept includes words, numeral characters, punctuation characters and their combinations. All the reflections concerning tagset proposal, meaning of individual marks and lemmatization was predetermined by the chosen means of tokenization. The result is that chains of alphanumeric characters (such as letters or numeral characters) between the two spaces are merged into one unit (token). Punctuation (colons, dots, question marks, marks of exclamation at the end of sentences, quotation marks, asterisks, mathematic symbols and others) are considered to be individual tokens in spite of the fact they are not separated from a preceding or following token by a space, e.g. in a sentence *„Win98 mi nefunguje!!!“* (lit. *„Win98 does not work!!!“*) there are 8 tokens (quotation marks “, *Win98, mi, nefunguje*, three marks of exclamation *!!!*, quotation marks “).

Tokenization is an important phase of automatic text processing because morphological analysis and disambiguation depend on it. Proposed principles of tokenization can raise questions concerning analytical forms such as *v rámci*, *na bielo*, *a teda*, word forms with hyphens or dashes (often used incorrectly) such as *kde-kto, čím – tým*, *8 – krát*, *Košice-Bratislava*, analytical forms *menej lukratívny*, collocations *Spišská Nová* Ves, numeral characters such as *1 984*, on the other hand there are agglutinated forms such as *oňho, akoby* and others. These words are divided into several tokens (despite they function as one language unit)[5] or integrated into one token (despite they function as two language units).

This kind of proposal of tokenization leads to such an interpretation of words and grammatical forms which does not always come to terms with our linguistic tradition. However, actual proposal does not exclude possibilities of implementation of logical module into text processing which would be employed afterwards as more appropriate basis for lemmatization and tagging.

3. Lemmatization of SNC

Lemma (l) is often defined as a “dictionary” form of a token. A set of language features which a lemma should include can be described as follows: the so-called basic values of morphological categories and a distinction between upper and lower case characters. But a concept of lemma is not applied absolutely only in its given semantic extension (that is analogous to the concept of token

[5] For instance, first parts of copmposite adjectives usually written with hyphen (or, incorrectly, with dash) are individual tokens. They are lemmatized on the basis of the word-form from the text: *česko – slovenský*, lemma (l) = *česko; bielo – červený,* l = *bielo*).
Composite pronoun forms such as *ten istý, tá istá, to isté, taký istý, tak isto, koľko ráz, -koľko ráz, koľký raz, -koľký raz, toľko ráz, toľký raz, tamto ten, tamto tá, tamto to*, and word forms which can be found in corpus with hyphen or with (incorrect) dash, such as *čo-to, koľký-toľký, koľko-toľko, aký-taký, ako-tak, kde-tu, kade-tade, kedy-tedy, kdesi-čosi, čosi-kdesi, čosi-kamsi, ten-ktorý, tá-ktorá, to-ktoré*, numerals such as *3-krát* etc., are indicated as two (or, three) tokens and each token is lemmatized, even though we respect the fact it is one lexeme.

and tokenization): Lemma is always indicated by a lowercase initial letter, e.g. *Alexander* = *alexander*. It's possible to argue that cancelling the distinction between upper and lower case characters may cause a loss of some semantic features of words, on the other hand the range of results based on small-lettered lemmas searching is considerably larger. Moreover, it can be assumed that occurrences of words primary written in capital letters can be easily found by means of precisely assigned queries in a corpus manager or by means of a negative filter. Information on proper names is also indicated on the tag level (see Garabík – Gianitsová – Horák – Šimková, 2004, chapter 3.1.).

Lemmas assigned to words belonging to inflected parts of speech can be of these values:

Substantives	**Adjectives**	**Verbs**[7]
Pronouns[6]	**words with ad-**	
Numerals	**jectival forms**	
particular gender	masculine	infinitive
singular (if exists)[8]	singular	
nominative (if exists)	nominative	
	base form	

Specific issues of lemmatization of some grammatical forms have been solved within the SNC tagset frame by detailed description (Garabík – Gianitsová – Horák – Šimková, 2004). Here we mention only some remarkable cases showing that homonymy (eventually homography) of some tokens is handled by manual annotation already on the lemma level:

od ***vedúcej*** *jedálne*	l = *vedúca*	***vedúcej*** *pretekárke*	l = *vedúci*
otcovi *priatelia*	l = *otcov*	*nepovedz* ***otcovi***	l = *otec*
jeho *nedobehneš*	l = *on*	***jeho*** *priatelia*	l = *jeho*
nechali ho ***samého***	l = *sám*	*šaty zo* ***samého*** *zlata*	l = *samý*
koľkí *žiaci prišli*	l = *koľko*	***koľkí*** *v poradí boli?*	l = *koľký*
dávať ***si*** *darčeky*	l = *si*	*kto* ***si*** *ty?*	l = *byť*
tuším *(asi)*	l = *tuším*	*niečo* ***tuším***	l = *tušiť*
začiatkom *júna*	l = *začiatkom*	*so* ***začiatkom*** *zimy*	l = *začiatok*

4. Morphological tagging of SNC

4.1. Means of assignment of attributes and their values. Several approaches to the

[6] A special case is represented by blended forms (*oňho, preňho, naňho, oň, preň, zaň*); these forms are considered to be forms of pronouns *on, ono*. Hence, lemma is composed of an independent preposition and a personal pronoun in nominative singular and respective gender: *oňho = o_on, preňho = pre_on, naňho = na_on*. These forms are considered to be agglutinated and this information is indicated on the tag level (see Garabík – Gianitsová – Horák – Šimková, 2004, chapter 3.3.13.).

[7] Negated forms are lemmatized by a negative infinitive form, e.g. there are lemmas *vidieť, mať, chcieť* as well as lemmas *nevidieť, nemať, nechcieť*. Special attention requires lemmatization of the negated verb form *byť*. Negative past and future tense forms are expressed by synthetic means (*nebol, nebude*), negation in the present tense is expressed by means of a particle *nie* (*nie je*). Presence of this particle often influences lemmatization of the verb (*nie je*; l = *nie, nebyť*) and indicates negation at the same time (see also Garabík – Gianitsová – Horák – Šimková, 2004, chapter 3.3.11.).

[8] Singular lemma can be found even in some pluralia tantum words, otherwise they are lemmatized as nominative plurals: e.g. *nohavice* (lit. N. pl. trousers), l = *nohavica* (lit. N. sg. trouser), but *Alpy,* l = *alpy* (N. pl., not N. sg. *alpa*).

morphological annotation (CNC, Multext-East Project, IPI PAN, etc.) represent several ways of notation (Foróová – Horák, 2003):

1. **Position** (Hajič): Every position is assigned one character, encoding one grammatical category. Values of irrelevant categories are indicated by dashes, e.g. *politikou,* t = NNFS7-----A---- (Noun, Noun-common, Feminine, Singular, 7th case, Affirmative).

2. **Abbreviated/attributive** (Multext-East): Only relevant categories for the given word form are assigned, e.g. *budeme*, t = Vcif1pan (Verb, copula, indicative, future, 1st person, plural, active voice, non-negative).

The advantage of the position notation is that it is more appropriate for computer proccessing, abbreviated notation is preferred because of better understanding by users. Taken into account of all aspects we have decided to make the best of both concepts.

Values of particular categories in SNC are encoded by one character taken from alphanumeric characters. A string of characters constitute one tag assigned to one token and lemma. A tag is then a set of characters that encode values of formal categories regarded relevant at the given word form. Number of characters varies but their order is obligatory.

Every tag is composed of two parts. First one defines morphological and grammatical properties of a token. It always begins with a character encoding part-of-speech, followed by characters for other categories, e.g. *Lingvista anotoval texty z korpusu.* (lit. A linguist annotated corpus texts.). There are 6 tokens (*lingvista, anotoval, texty, z, korpusu,* . (dot)), every token is assigned a lemma (l) and a tag (t):

Lingvista	l = *lingvista*	t = SSms1
anotoval	l = *anotovať*	t = VLescm+
texty	l = *text*	t = SSip4
z	l = *z*	t = Eu2
korpusu	l = *korpus*	t = SSis2
.	l = .	t = Z

The second (facultative) part specifies token as a part of specific word classes (proper names, defective forms). In most cases token does not belong to any of these specific classes, the second part of a tag is missing at that time. In cases of proper names after the first part we assign **:** (U+003A COLON) and a special character **r**. In cases of defective or wrong forms a colon is followed by **q:**

od Mináriká	l = *Minárik*	t = SSms2:r
Goldsteinovú tvár	l = *Goldsteinov*	t = AFfs4x:q

Detection of defective forms is instructed also by frequency of occurrences in corpus, eventually by a type of an "error". If a word form is not standard but adequate in the given cases (*neni, do Košicoch, za prvé, postavím sa do rady, prádlo*), it is not regarded wrong. As defective forms are viewed typos and obvious spelling mistakes.

4.2. General tagging principles. Category is indicated by a character if it is relevant for the given form. E.g. for the pronoun *ako* categories of gender, number and person are not relevant,

therefore we indicate only POS and a paradigm: l = ako, t = PD. Verbs in a base (infinitive) form cannot be assigned a category of number, person and gender congruency, therefore we indicate only POS (verb), verb form (infinitive), verb mode (completive), affirmation (affirmative): l = *vniknúť*, t = VId+.

Characters are assigned to values of (morphological) categories relevant for the given word form even in such cases when categories are not "visible" from a word form, hence, formally they are not transparent. In some cases these categories can be contextually determinable; since the context is unlimited for the purpose of manual tagging. First of all we take into account congruency within syntagmas or valency relations. Specificly we also indicate a category of person in -l-participle forms on the basis of presence or absence of grammatical morphemes *som, si, sme, ste*.

For example:

*Nechcem cestovať v tom **kupé**.*	(neuter, sg., L)	l = *kupé*	t = SUns6
*Dozvedel sa to od **päť** chlapov.*	(masc. anim., pl., G)	l = *päť*	t = NUmp2
*Pozdravil sa **jeho** sestre.*	(fem., sg., D)	l = *jeho*	t = PUfs3
***Nenašiel** som ani kúsok.*	(compl., sg., 1. pers., masc., neg.)	l = *nenájsť*	t = VLdsam-

In cases such as *kupé* (coupe) and *jeho* (his) essentially there's an absolute morphological homonymy, because these words have only one form through which they enter syntactic relations and thereby they are clearly defined only by the context.

Some forms usually referred to as inflexible, tend to be declined. Declined and non-declined forms of one lemma can occur in the same context and a user can find all possibilities and discover development of inflection. Frequency analysis can show their occurrences ratio. For example substantive *Philips* (l = *philips*) was in SNC observed in these forms of genitive singular:

*Cieľom transakcie je transformácia **Philips**...*

*...hovorí O. Š. z **Philips** Slovakia...*

*...s kapitálovou pomocou **Philipsu**...*

*...veľkých spotrebičov od **Philipsa**...*

Context is taken into account even when treating homonymy (eventually, homography) of some word forms in a paradigm of one lexeme:

e. g. a form *pekné* (beautiful) – possibilities:	N, A pl. masc. inanim.	t = AAmp1x	t = AAmp4x
	N, A pl. fem.	t = AAfp1x	t = AAfp4x
	N, A sg. neuter	t = AAns1x	t = AAns4x
	N, A pl. neuter	t = AAnp1x	t = AAnp4x

Context: *Dievčatá sú **pekné**.* (lit. Girls are beautiful.) The only one possibility is: t = AAnp1x.

Parts-of-speech homonymy is solved with the aid of codification books and dictionaries (e.g. MSJ, 1966; KSSJ, 2003); semantics is also taken into account:

***Tuším** vo vzduchu búrku.* (I feel)	l = *tušiť*	t = VKesa+
***Tuším** budú problémy.* (maybe)	l = *tuším*	t = T
***Prosím** si vodu.* (I beg for)	l = *prosiť*	t = VKesa+
*Poďte, **prosím**.* (please)	l = *prosím*	t = T

***Lepšie** to nebude.* (better, adj.) l = *dobrý* t = AAns1y
*Vieš to aj **lepšie**.* (better, adv**.**) l = *dobre* t = Dy
*Bolo **zima**.* (cold – adv.) l = *zima* t = Dx
*Prišla/Bola **zima**.* (winter – subst.) l = *zima* t = Ssfs1

Particular grammatical categories can be even assigned to such abbreviations, acronyms and units of measure coined from flexible parts of speech (*nám., ul., č., l, cm*), or to the forms functioning as declinable parts of speech (*do SND, v SR*).

4.3. Parts of speech. In SNC a set of word forms is divided into 19 classes, ten of them reflecting traditional word classes (parts of speech) – nouns (S), adjectives (A), pronouns (P), numerals (N), verbs (V), adverbs (D), prepositions (E), conjunctions (O), particles (T) and interjections (J) – nine of them representing various and specific language elements – formal participles (G), reflexive morphemes *sa/si* (R), conditional morpheme *by* (Y), numbers (0), abbreviations and symbols (W), unclassifiable parts of speech (Q), citation forms (%), punctuation (Z) and non-word elements (#). Traditional parts of speech basically reflect part-of-speech classification in Slovak codification books (KSSJ, 2003; PSP, 2000). Disputable issues concerning part-of-speech classification required some compromise solutions:

1. **Verbal nouns** (*písanie* – writing, *hovorenie* – speaking, *rešpektovanie* – respect, etc.) are treated as nouns.

2. **Agglutinated forms** (*oňho, preňho, naňho, oň, preň, zaň*) are labelled as pronouns; prepositional part is reflected as a part of a lemma (see Garabík – Gianitsová – Horák – Šimková, 2004, chapter 2.2.3.3.), their state of agglutination is conveyed by formal category and its value within a tag. E.g. *Starám sa **oňho*** (lit. I take care **about him**.) l = *o_on* t = PPms4g

3. **Secondary prepositions** such as *s ohľadom na* (regarding), *v závislosti od* (in dependence on), *na rozdiel od* (unlike), *v prípade* (in case) are tagged as junctions of a preposition (or, prepositions) and a substantive.

4. **Active and passive participles** and adjectives converted from verbs (*písaný, otvorený, obutý, píšuci, hrajúci, stojaci*) are regarded as transitional groups, therefore we decided to specify these adjectival forms as a group of formal participles. They are differentiated from adjectives on the ground of their form and origin. (Formal) passive participles are considered to be adjectival forms coined form infinitive stems of verbs by adding participle morphemes *-n+ý, -t+ý* (*sklad-a:ť* → *sklad-a:n-ý*). (Formal) active participles are considered to be adjectival forms coined from present tense stems (eventually from 3rd person plural forms) of verbs by adding participle morphemes *-úc+i/-uc+i*, *-iac+i/-ac+i* (*sklad-a:j-ú* → *sklad-a:j-úc-i*), e.g. *píšuci, písaný, žnúci, žatý, bijúci, bitý, spiaci, šijúci, šitý, sejúci, siaty*. Deverbative adjectives created as a result of word-formative process of derivation are not considered to be participles, e.g. *písací, skladací, žací, bicí, spací, šijací, sejací*.

5. At this level of annotation we do not differ between **reflexive pronouns *sa, si*** and *sa, si* as verbal components. Their distinction should be a subject of individual papers written on the basis of corpus database, this issue would eventually be treated on the level of syntactic annotation, too. In this case a possibility to specify morphological categories of *sa, si* as reflexive pronouns is excluded. On the other hand it's possible to disambiguate reflexive verb components (*pospať si, zaspievať si, zaspieval si si*) and a 2nd person, present tense, indicative form of a verb *byť – to be* (*ty si kľačal, zaspieval si si*).

6. **Morpheme *by***, a part of conditional verb form, is tagged as independent word class. Other forms with morpheme *by* (e.g. conjunctions *keby, aby, žeby, akoby, sťaby,* particles *aby, keby*) are tagged as conjunctions or particles (according to function they have) but we take into account agglutination of a morpheme *by*. This fact is reflected by adding (Y) character– conditionality.

7. **Numeral characters** (Roman and Arabic as well), eventually numeric symbols and combinations of numeral characters are assigned to independent class "numbers" (0).

8. **Symbols** such as *l, km, H_2O, X569847* and **abbreviations** such as *atď., tzv., t. j., pod., kt., i., XML, SND* fall into the class of "abbreviations and symbols" (W). On the other hand abbreviated words such as *Satur, Slovnaft, Rempo* on the basis of their function and meaning fall into the class of substantives.

9. **Multi-word lexemes** in SNC are composed of several tokens. First and non-individual parts (words) of a larger lexical unit are often impossible to define. That's why they fall into the class of the " unclassifiable part of speech": ***po*** *slovensky*, ***fast*** *food,* ***Los*** *Angeles*. Forms such as *slovensko-český*, *2-krát* are indicated as three tokens: *slovensko* (unclassifiable part of speech), - (punctuation mark), *český* (adjective); 2 (number), - (punctuation mark), *krát* (unclassifiable part of speech). Similar approach is applied when dealing with "juxtapositions" *až60* (unclassifiable part of speech).

10. **Citation forms** (%) include foreign multi-word phrases and sentences not adapted in a second language but functioning as parole units taken from a source language: *Take it easy!; Šaj pes dovakeras; „správně vychlazená dvanáctka"; Ta naše povaha česká; náměstí*. These tokens need not be indicated by quotation marks. Individual words of foreign origin such as *kuskus, ska, sitar, djembe, česnečka, květák* do not fall into this class because in Slovak sentences they function in accordance with Slovak grammatical rules.

4.4. Categories and their values. In the matter of morphological annotation our starting point was the theory of grammatical categories introduced by MSJ (1966), eventually by other works dealing with morphology. In a tagset there are indeed some categories with their values not explained and mentioned by traditional morphology (paradigm, verb form, agglutination, conditionality). There is a formal-morphological characteristics important in the process of token

disambiguation.

In nominal parts of speech the second position is occupied by a character indicating the type of **paradigm** with values: substantival, adjectival, pronominal, numeral, combined, uncompleted and adverbial.

Formal attribute "paradigm" is understood as specification of form of a particular word within a word class (e.g. *taký* (lit. such) is pronoun, but it has a form of an adjective, its tag is PA). Characters standing in for the substantival, adjectival, pronominal, numeral and adverbial paradigms are identical with the part-of-speech indicators (S, A, P, N, D).

Combined paradigm (F) is valid for words having a partial congruent paradigm. Development of their declension has undergone complicated processes and they're not unambiguously assignable to clearly-defined declension types. This class contains words such as *kuli, gazdiná* and nouns declined in the same way, *otcov, matkin* (all individual possessive adjectives), *on, ona, ono, kto, čo, nikto, nič...*, *môj (tvoj, náš, váš), ten (tá, to), sám, onen, žiaden, všetok, jeden.*

Uncompleted paradigm (U) is assigned to those substantives, adjectives, pronouns and numerals traditionally considered to be inflexible (*kupé, super, jeho*), or with a tendency to be declined (*kanoe* – only G pl. *od kanoí*), or usually inflexible (*pani, koľko, toľko, viacero, päť, sto, tisíc*). In these cases the distribution of declined and inflexible forms depends on several circumstances. As noun with uncompleted paradigm can be regarded also a form of such a substantive of foreign origin which is essentially declinable (*Phillips, Tesco*) but author of a text prefers inflexible form in the given case (genitive singular *od Phillips* – SU), even though prevalent is the declined form (genitive singular *od Phillipsu* – SS).

Adverbial paradigm (**D**) is assigned to inflexible pronouns and numerals, in KSSJ indicated by grammatical labels *neskl.* (inflexible) or *príslov*. (adverbial), eventually they function as adverbials (*koľkonásobne, koľkorako, tam, tu, vtedy, vždy, viacnásobne, dvojako*). In these cases a tag contains only this kind of information. Other categories such as gender, number or case are not indicated.

First two characters of a tag present the following combinations: **SS** – substantive with substantival paradigm (*mama*), **SA** – substantive with adjectival paradigm (*vedúci*), **SF** – substantive with combined paradigm (*gazdiná*), **SU** – substantive with uncompleted ("invisible") paradigm (*pani, kanoe*), **AA** – adjektive with adjektival paradigm (*pekný*), **AF** – adjektiv with combined paradigm (*otcova),* **AU** – adjektiv with uncompleted ("invisible") paradigm (*super*), **PS** – pronoun with substantival paradigm (*koľkátka*), **PA** – pronoun with adjektival paradigm (*taký*), **PP** – pronoun with pronominal paradigm (*ja*), **PF** – pronoun with combined paradigm (*on, sám, žiaden*), **PU** – pronoun with uncompleted ("invisible") paradigm (*jeho, jej, ich, koľko, toľko*), **PD** – pronoun with adverbial paradigm (*tam, niekedy*), **NS** – numeral with substantival paradigm (*milión, raz*), **NA**

– numeral with adjektival paradigm (*štvrtý*), **NN** – numeral with numeral paradigm (*tri*), **NF** – numeral with combined paradigm (*jeden*), **NU** – numeral with uncompleted ("invisible") paradigm (*päť, sto, tisíc, pätoro, veľa*), **ND** – numeral with adverbial paradigm (*dvakrát, mnohonásobne*).

As far as nominal parts of speech are concerned the indication of category is usually followed by elementary morphological characteristics:

- **Gender**: masculine animate (**m**); masculine inanimate (**i**); feminine (**f**); neuter (**n**); unspecified (**o**); general (**h**), the last holds true for pronouns and verbs;
- **Number**: singular (**s**), plural (**p**) and unspecified (**o**);
- **Case**: nominative (**1**)**,** genitive (**2**), dative (**3**), accusative (**4**), vocative (addressing) (**5**), locative (**6**), instrumental (**7**), unspecified (**o**);
- **Grade**: base form (or irrelevant grade) (**x**), comparative (**y**), superlative (**z**); holds true for adjectives, adverbs and formal participles.

Value "unspecified" (**o** character) in the position of gender, number or case is relevant for some morphologically non-transparent or homonymous forms if context indicates several conflicting values for one category (*Mužov, žien a detí je* ***päť.*** *Kúpili* ***kanoe. Mesto!***).

The second verb position **"verb form"** can be assigned to the following values: infinitive (**I**), formal present (indicative) (**K**), imperative (**M**), transgressive (**H**), -*l*-participle (**L**), future form (most often it is a form of verb *byť*, also synthetic future tense of uncompleted verbs – *poletím, ponesiem* etc.) (**B**). Establishment of the category of a verb form resulted from our attempt at description of analytic verb forms. Even though we regard this solution not ideal, for the time being it represents a systematic approach to this complicated issue. The categories of tense and modus are not indicated individually because they are included in particular definite verb forms being indicated as follows:

Indicative

	Verb form	Example	Lemma	Tag
present tense	formal present	*píšem*	*písať*	**VK...**
past tense	*-l*-participle + formal present	*písal som*	*písať* formal *byť* *	**VL...** **VK...**
future tense (uncompleted verbs)	future + infinitive	*budem písať*	formal *byť* *písať*	**VB...** **VI...**
	future	*ponesiem*	*niesť*	**VB...**
future tense (completed verbs)	formal present	*napíšem*	*písať*	**VK...**

Imperative

present tense	imperative	*píš!*	*písať*	**VM...**

Conditional

present tense	*-l*-participle + conditional morpheme + formal present	*písal by som*	*písať* *by* formal *byť**	**VL...** **Y** **VK...**
past tense	*-l*-participle + conditional morpheme + formal present + *-l*-participle	*bol by som písal*	formal *byť* *by* formal *byť** *písať*	**VL...** **Y** **VK...** **VL...**

* if it's present

Other verb values (if relevant) are as follows:

- **Aspect**: perfective (**d**), imperfective (**e**), with both aspects (**j**); its indication is done on the basis of dictionary qualifiers (in KSSJ, 2003);
- **Number**: singular (**s**), plural (**p**);
- **Person**: first (**a**), second (**b**), third (**c**); indicated also in -l-participle forms and grammatical morphemes (forms of a formal verb *byť*);
- **Gender congruency**: congruency of masculine animate (**m**), inanimate (**i**); feminine (**f**); neuter (**n**); undefined (**o**), general gender (**h**) – relevant only for -l-participle verb forms;
- **Negation**: affirmative (+) and negative (**-**); relevant only for verbs.

Referring to pronouns, numerals, formal participles, prepositions, conjunctions, particles and a morpheme *by* some other attributes are indicated. There are usually formal features we regarded important when doing a formal description of forms:

- **Agglutination:** a character for this attribute (**g**) is assigned to pronouns at the end of a tag, if it's relevant (forms such as *oňho, preňho, naňho, oň, preň, zaň* etc.).
- **Independent use of numerals** takes the second position in a tag (X), if it's an independent expression of quantity, e.g. mathematic operations (*2 – 2 = 0*), specific nominative (*Dráma 2000*) etc.
- **Type of participle**: active (**k**) and passive (**t**) takes the second position so far as formal

participles (**G**) are concerned (*píšuci, písaný*).

- **Form of preposition:** vocalized (**v**) or non-vocalized (**u**) takes the second position in a tag (*vo, v; ku, k; so, s*).
- **Conditionality:** holds true for conjunctions or particles *keby, aby, žeby, akoby, sťaby* (O**Y**, T**Y**).

Examples illustrating previous categories from the SNC tagset can be also found in current version on the website.[9] There is more detailed described the system of categories, their values and concrete solutions. On the basis of this tagset a manual tagging of the Slovak National Corpus is carried out.

5. Manual tagging of SNC – first results and possible perspectives

Since the morphological annotation of corpus in addition to tagset and computational tools (tagger) requires morphological dictionary and text data (namely training corpus and testing data) we consider manual tagging to be important step in obtaining material. By the first half of 2004 manual tagging of the novel 1984 by G. Orwell and texts from the internet magazine InZine have started. The annotation is carried out by three students of the Faculty of Philosophy of Comenius University in Bratislava and a workgroup of eleven students (Faculty of Arts, University of Prešov; cooperation with Prešov is held on the ground of the grant project *Morfosyntaktická analýza SNK* (VEGA 1/3149/04; *Morphosyntactic analysis of the SNC*).

Even though a manual annotation is time-consuming, by the end of May 2004 we managed to obtain a set of texts including about 19 000 tokens from the Orwell's novel 1984 and about 25 500 tokens from internet magazine InZine. It should be emphasised that these are manual anotations that have not been yet checked and unified in accordance with current annotation principles. We are going deal with this problem in the future. An average precision ratio of manual annotation is about 91,5 % and an average speed is about 80 tokens a minute. On the other hand we regard manual annotation important. Acquired material would be useful for further tagger training. However, we consider necessary to speed up the manual morphological annotation and make it more effective by means of automatically preprocessed annotated texts. Human anotator is subsequently given an automatically annotated text and he/she should decide whether the given tag is assigned correctly or not (the latter case requires correction).

The following period of time should be devoted to the manual annotation simultaneously with testing of appropriate tools and its applications for tagging of Slovak texts. On the basis of co-operation with Institute of Formal and Applied Linguistics (ÚFAL) at the Faculty of Mathematics and Physics, Charles University, Prague[10] we have at our disposal a morphological analyzer and

[9] http://korpus.juls.savba.sk/korpus/biblioteka/publikacie/Tagset-aktualny.pdf

[10] Grant project *Využitie spoločných vlastností češtiny a slovenčiny na budovanie anotovaných národných jazykových korpusov;* lit. *The Aplication of common properties of Czech and Slovak languages for the purpose of buiding of*

disambiguator developed by J. Hajič; also we are going to use a Slovak version of the morphological analyzer ajka proposed by R. Sedláček a M. Grác (Masaryk University, Brno). We believe this co-operation will be successful.

References

Dvonč, L. (1984): Dynamika slovenskej morfológie. Veda, Bratislava.

EAGLES (1996). Recommendations for the morphosyntactic annotation of corpora. EAG-CSG/IR-TR.1. ILC-CNR, Pisa. <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html/>

Erjavec, T. (2001): Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984. In: 6th Natural Language Processing Pacific Rim Symposium, NLPRS'01, Tokyo, pages 487-492. <http://nl.ijs.si/et/Bib/NLPRS01/mte-nlprs01.pdf>

Forróová, M., Horák, A.(2003, to be published): Morfologická anotácia korpusu. In: Proceedings of International Conference *Slovenčina na začiatku 21. storočia*. Prešov. <http://korpus.juls.savba.sk/korpus/biblioteka/publikacie/morfanot.pdf>

Forróová, M., Garabík, R., Gianitsová, L., Horák, A., Šimková, M. (2003, to be published): Návrh morfologického tagsetu SNK. In: Proceedings of *Slovko 2003 - Slovanské jazyky v počítačovom spracovaní,* Bratislava. <http://korpus.juls.savba.sk/aktivity/archiv/data/Tagset_oponentura.pdf>

Garabík, R., Gianitsová, L., Horák, A., Šimková, M. (2004): Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. (Current version of May 4, 2004) SNK JÚĽŠ, Bratislava. <http://korpus.juls.savba.sk/korpus/biblioteka/publikacie/Tagset-aktualny.pdf>

Hajič, J. (2000): Popis morfologických značek – poziční systém. ÚČNK – ÚFAL MFF UK, Praha. <http://ucnk.ff.cuni.cz/manual/znacky.html>

Hana, J., Hanová, H. (2002): Manual for morphological annotation. ÚFAL MFF UK, Praha.

Krátky slovník slovenského jazyka (2003). 4th edition. eds. J. Kačala, M. Pisárčiková, M. Považaj. Veda, Bratislava. (KSSJ)

Leech, G. (2000): Anotační systémy pro značkovaní korpusu. In: Studie z korpusové lingvistiky. Acta Universitatis Carolinae – Philologica 3 – 4. Karolinum, Praha, pages 185 – 197.

Morfológia slovenského jazyka (1966). ed. J. Ružička. Veda, Bratislava. (MSJ)

Oravec, J. – Bajzíková, E. – Furdík, J. (1984): Súčasný slovenský spisovný jazyk. Morfológia. SPN, Bratislava.

Pravidlá slovenského pravopisu (2000). ed. M. Považaj. Veda, Bratislava. (PSP)

Przepiórkowski, A. – Woliński, M. (2003): A Flexemic Tagset For Polish. In: Proceedings of Morphological Processing of Slavic Languages, EACL 2003, Budapest, pages 33 – 40. <http://dach.ipipan.waw.pl/~adamp/Papers/2003-eacl-ws12/ws12.pdf>

annotated national language corpora.

Resume

V príspevku sa zaoberáme problematikou morfologickej analýzy Slovenského národného korpusu. Automatická morfologická anotácia textov predstavuje aplikáciu, ktorá sa teší veľkej pozornosti najmä v súvislosti so spracovaním korpusových dát a vyžaduje si prípravu morfologického tagsetu pre ručné značkovanie SNK. V úvodnej časti predstavujeme úvahy a východiská, ktoré viedli k voľbe spôsobu spracovania. Otázkam morfologickej analýzy (vymedzeniu základných pojmov) a jej aplikácii pri značkovaní korpusu sa venujeme v teoretickej rovine. Zdôrazňujeme rozdiel medzi analýzou, ktorú vykonáva človek, ktorý je jazykovo kompetentný, a tou, ktorú vykonáva počítač. Stáli sme pred rozhodnutím, či navrhnutý súbor značiek bude výsledkom nového formálneho opisu jazyka alebo budeme vychádzať z lingvistických opisov, ktoré existujú a len sa ich pokúsime formalizovať. Za dôležitú považujeme najmä zásadu prístupnosti (koncovému používateľovi, ktorým býva aj nelingvista) a zásadu konsenzu vedeckých teórií, teoretickej „neutrality“. Preto sme sa rozhodli prihliadať najmä na systematický opis podaný v akademickej Morfológii slovenského jazyka (ďalej MSJ, 1966), prípadne na ďalšie morfologické práce (Oravec – Bajzíková – Furdík, 1984; Dvonč, 1984). Konflikt medzi reprezentáciou zaužívaných gramatických kategórií (často s nejasnými kritériami morfologickej klasifikácie) a možnosťou automatického spracovania jazyka sa odráža aj v koncepcii morfologického tagsetu pre SNK, ktorý vychádza aj zo skúseností zahraničných tvorcov tagsetov pre morfologickú analýzu. Pre potreby značkovania bol preto zvolený formálno-gramatický princíp, ktorý však s ohľadom na domácu lingvistickú, resp. gramatickú tradíciu má isté špecifiká.

V nasledujúcich častiach upozorňujeme na všeobecné črty a niektoré konkrétne riešenia problémov morfologickej anotácie textov SNK. Tie boli ovplyvnené i prístupom k segmentácii textu na tokeny, preto sa krátko najprv zmieňujeme o tokenizácii textov SNK a jej zásadách. Tokeny sa (s ohľadom na počítačový prístup) nemôžu kryť a ani sa nekryjú s pojmom slovo, či dokonca gramatický tvar. Tokenizácia je dôležitou etapou v automatickom spracovaní textu, pretože od jej výsledkov je priamo závislá morfologická analýza a dezambiguácia. Navrhnuté zásady tokenizácie môžu vyvolávať otázky pri zložených tvaroch, pri zápisoch so spojovníkom či pomlčkou (často aj chybne použitými), pri analytickom stupňovaní, združených pomenovaniach, zložených číslovkách a na druhej strane pri aglutinovaných podobách, keďže tieto lexikálne jednotky sú rozdelené na viac tokenov (napriek tomu, že tvoria jednu jazykovú jednotku) alebo zlúčené do jedného tokenu (napriek tomu, že ide o dve pôvodné jednotky). Tento návrh tokenizácie vedie k takej interpretácii slov a gramatických tvarov, ktorá nie vždy súhlasí s tradičnou lingvistickou tradíciou. Súčasné riešenie však nevylučuje možnosť zapojenia logického modulu do spracovania textu, ktorý by sa neskôr uplatnil ako vhodnejší základ lematizácie a morfologického značkovania. V časti o lematizácii textov SNK zdôrazňujeme, že aj pojem lemy neaplikujeme absolútne v jeho významovom rozsahu. Charakterizujeme niekoľko základných zásad a uvádzame niektoré zvláštne

prípady.

Predstavenie zásad morfologického značkovania a formy zápisu značiek tvorí strednú časť príspevku. Pri morfologickej notácii volíme kombinovaný pozično-atribútový spôsob. Zaujímavosťou je rozdelenie tagu na dve časti. Druhá (nepovinná) časť zaraďuje token do určitých špeciálnych skupín (ako sú vlastné mená alebo defektné zápisy). Ako všeobecnú zásadu sme prijali uvádzanie znakov pre jednotlivé atribúty aj vtedy, keď síce je hodnota pre daný tvar relevantná, ale nie je z formy slova „viditeľná“, teda nie je dostatočne formovo transparentná. V niektorých prípadoch ich môžeme určiť z kontextu, ktorý je pre potreby ručného značkovania neobmedzený. Prihliadame najmä na kongruenciu v rámci syntagmy alebo na valenčnú väzbu. Podobne sa rieši aj "tzv." nesklonnosť substantív, adjektív a i. Ich tvary totiž považujeme za absolútne morfologické homonymá, keďže tieto slová majú len jednu formu, ktorou sa zapájajú do syntaktických vzťahov a väzieb. Tým sú však jasne definovateľné z kontextu. Navyše, niektoré tvary, uvedené v odbornej literatúre ako nesklonné, sa v bežnej praxi začínajú skloňovať. Vyskloňované a nevyskloňované tvary jednej lexémy sa môžu v istom páde vyskytovať popri sebe a používateľ má možnosť nájsť všetky prípady a zistiť pokročilosť flektivizácie. Pri slovnodruhovej homonymii sa v zásade riadime kodifikačnými príručkami a sémantikou.

Nasleduje stručný opis súboru morfologických značiek. Množina slovných foriem používaných v slovenčine sa v morfologickom tagsete SNK rozdeľuje do 19 tried, z ktorých 10 v zásade zodpovedá tradične vydeľovaným slovným druhom a 9 obsahuje rôzne špecifické jazykové prvky. Pri klasických slovných druhoch sa v podstate rešpektuje slovnodruhové zaradenie podľa súčasných slovenských kodifikačných príručiek. V sporných otázkach určenia slovného druhu sme prijali niektoré kompromisné riešenia, ktoré v príspevku predstavujeme. Pri morfologickom značkovaní sme vychádzali z teórie gramatických kategórií, ako ich podáva akademická Morfológia slovenského jazyka (1966), príp. iné morfologické práce. V tagsete SNK sa však stretneme aj s kategóriami a ich hodnotami, o ktorých sa v tradičnej morfológii ako o kategóriách neuvažovalo (paradigma, slovesná forma, aglutinovanosť, kondicionálnosť). Ide o formálno-morfologické charakteristiky, ktoré sú dôležité na zjednoznačnenie tokenu. Z nich najviac miesta dostáva najmä objasnenie kategórie paradigmy, ktorú chápeme ako vymedzenie špecifickej formy konkrétneho člena slovného druhu. Kategória slovesnej formy je zas výsledkom pokusu o uchopenie analytických tvarov slovies. I keď si uvedomujeme, že toto riešenie nie je ideálne, na súčasnej úrovni znamená systematické uchopenie tejto náročnej problematiky.

Na základe tohto tagsetu v súčasnosti prebieha ručné značkovanie textov, ktoré obsahuje Slovenský národný korpus. Ide o román Georgea Orwella 1984 a texty z internetového časopisu InZine. Prvé výsledky a okolnosti ručného značkovania SNK a možné perspektívy ďalšieho rozvoja a zefektívnenia práce predstavuje záverečná kapitola príspevku.