Mária Pisárčiková, Slovak Academy of Sciences, L'. Štúr Linguistics Institute, Bratislava
Vladimír Benko, Comenius University, Faculty of Education, Computational Linguistics Laboratory, Bratislava

Slovak Synonym Dictionary

Abstract

This paper deals with the new *Slovak Synonym Dictionary*, a fundamental work describing Slovak lexis on some 1,000 pages containing more than 40,000 paragraphs. Some theoretical issues concerning lexical synonymy are discussed, initial lexicographic decisions are shown and an example entry is introduced. Lexical-computing questions addressed cover gathering of additional lexical evidence and lexical data representation and validation. A sample dictionary page is presented in the Appendix.

1. Introduction

The image of lexis is mirrored in synonym dictionaries by grouping words of the same part of speech into semantically close or equivalent microstructures joined by a common concept. These may be of different varieties: Most synonym dictionaries just present 'bare lists' of words with similar meanings (GLS 1974, DS 1990). There also exist, however, such lexicographic descriptions of lexical synonymy where the relationships between individual list members are analyzed, functional and stylistic characteristics are presented, and usage of words is shown in examples or citations from literature (SS 1975, WNDS 1984, MSS 1989, OT 1990). This type of complex description of lexical synonymy does not only make accessible one of the communicatively most important elements of lexis, but also comes with its own scholarly dimension, as it becomes a source of deeper knowledge about the language. The authors of the Slovak Synonym Dictionary (*Synonymický slovník slovenčiny* – SSS, or 3S 1995) have chosen as their objective to integrate these two (pragmatic and scholarly) ways of description in presenting the Slovak language in its full expressive richness and functional and stylistic differentiation, and to produce a practical and user-friendly dictionary.

2. Initial Considerations, Sources

The crucial question in compiling a synonym dictionary is how the concept of *lexical synonymy* is to be understood. More simply, we can speak about a narrower and broader scope of this concept. Our project has adopted the broadest definition of lexical synonymy with partial synonyms and quasi-synonyms also being taken into consideration in producing *synonymic chains*. The actual selection of synonyms has been governed by their real occurrence in the language, and it has been just a question of our lexicographic method, how to capture the image of lexical synonymy in its relative completeness.

While compiling the 3S, the authors considered descriptions found in explanatory (monolingual) and bilingual dictionaries, specialized dictionaries such as the Dictionary of Foreign Words, Dictionary of Slovak Slang, various sorts of terminological dictionaries, older synonym dictionaries, and also their own base of lexical evidence based on excerpts from works of fiction. This was the first, and most important, phase of data base collection.

3. Theoretical Issues

The core task in compiling this dictionary, however, was the creation of chains of synonyms which included a general analysis of mutual relationships between the individual members of these chains. In our approach the well-known paradox has been utilized that in treating synonyms (words of identical or similar meaning, but of different acoustic and/or graphic shape) the attention must be paid, not to *identity*, but to *differences* between the individual chain members.

The dictionary entry is headed by the root member of the synonymic chain or *a dominant*. The dominant expresses in a most general way the meaning common to all members of the chain. It mostly belongs to the core of lexis and is usually stylistically neutral (i.e. not labelled as *colloquial*, *expressive*, *bookish*, etc.). If the headword is polysemous, it typically becomes dominant in its basic (non-derived) meaning, while the other meanings are usually members of other synonymic chains. The dominants are typically native lexical units, though, in some cases, also the borrowed word can be dominant if it is widely known and more frequently used than the native synonymous (but marginal) expression.

Identification of a dominant does not usually present a great theoretical or practical problem. The synonymic chains are not formed by groups of words that would share a totally identical meaning. In fact, quite the opposite is almost invariably true: the individual chain members differ by their frequency, meaning shades, register or other attributes, and the process of determining a dominant is a relatively straightforward one. If, nevertheless, two semantically and functionally identical candidate headwords should appear in the chain, then either two different chains might be created that would likely contain identical chain members (though a complete similarity is not expected), or a secondary formal criterion (e.g. alphabetical) could be applied to determine the dominant.

The next step in compiling an entry is the description of the dominant's meaning: it must be general enough to cover the essential meaning of all members of the chain. Differential semantic components (*sememes*) of the other chain members are always explained with respect to their relationships to the root chain member or the dominant. This is why our dictionary is also explanatory, which makes it different from most of the other synonym dictionaries. In our dictionary, the meaning explanations may also include an antonym (if any exists).

To examine and compare meanings of individual members of the chain, a method of complex semantic analysis has been applied. Semantic components of higher orders of generalization (so called integration or identification sememes) are common to all chain members, while the individual synonyms differ mostly by lower-order sememes (specification sememes) and, finally, by their so-called 'subsememes', which represent what we usually call meaning (or other) shades. This was where the attention was paid in creating comments to individual synonyms. The explanatory comments on words with meanings that gradually diverge from that of the main headword can guarantee proper understanding of synonymic relationships between the dominant and the members of the chain. These relationships can be mainly found between the chain members and the dominant (they are mutually interchangeable), while between the individual chain members themselves such relationships need not necessarily exist and they also need not be mutually interchangeable. If, however, a word has a synonymic relationship to some members of a chain but not to the dominant, it *does not* belong to this chain.

The position in the chain assigned to the individual synonyms is mainly governed by proximity of their meanings to the dominant and their stylistic and functional labels. The word with the meaning closest to the dominant is placed in the first position, immediately after the dominant, usually regardless of its stylistic label, and, likewise, the word that is most distant by its meaning and, possibly, also by its stylistic label, is placed at the end of the chain. The chain, however, is not 'completed' by this word – synonymic chains are open systems with the potential for additional members to be appended.

Since synonymy is closely related to stylistics, it has been the dictionary authors' ambition to present the chains in their full stylistic and functional differentiation. The words are labelled as *colloquial, bookish, newspaper-style, special purpose* (scholarly or scientific style), *administrative, poetic, biblical* or *religious*. From the point of view of emotional assessment, synonyms can be marked as *expressive*, or more precisely as *pejorative, euphemistic, hypocoristic* or *family-use*, *ironic* or *jocular*, and also as *rude* or *vulgar*. From the temporal aspect, the synonyms are qualified as *old-fashioned* or *obsolete*, *archaic* or *historical*. Regarding the frequency, some words are labelled as *rare*, and from the codification aspect, the synonyms are marked as *dialectal, slang* or *substandard*, with special graphic marking of *'incorrect'* and *'uncodified'* lexical units.

A very important component of the dictionary entries are examples, that demonstrate the use of synonyms in contexts. As examples, typical collocations or broader contexts are chosen. In the case of very 'exclusive', marginal, or in other way 'very marked' synonyms, the authors' names are also indicated.

A separate theoretical problem in this kind of a dictionary is the treatment of semi-synonymy. Semi-synonyms are words with family (generic) relationships, or with relationships of different levels of intensity or specificity. With words having a very general meaning and a certain level of meaning diffusion, partial chains (subchains) are often developed, e.g. with verbs like *ísť* ('to go'), *hovoriť* ('to speak'), or adjectives like *veľký* ('big', 'great', or 'large'), *ostrý* ('sharp'), etc. Having the user in mind, our dictionary generously presents many of these subchains. It is usually fairly easy to recall the general concept in one's mind – it is, however, much more difficult to recall a word that would reflect some specific features of a given reality. As an example, refer to dictionary entries for verbs like *zjesť* ('to eat up'), *ísť* ('to go'), etc.

Other types of complex lexicographic problems can be found in treating polysemy, reflexiveness and aspect of verbs (typical features of Slavonic languages), as well as verbs with productive prefixes. The experience gathered during the compilation of the dictionary has shown that, in many cases, no direct 'lexicographic templates' could be used, but a set of elaborate methodical procedures needed to be applied to allow for identification of the real position of any given word within the microsystem of the synonymic chain. The lexicographic solution adopted need not be the only possible one. It must, however, reflect the language reality in a truthful way. Authors of this dictionary are aware of many theoretical and practical lexicographic problems having alternative solutions.

An important part of the dictionary is presented by cross-reference entries. They represent some 3/4 of the total 40,000 main entries. In many cases the headword refers to more than one main headword. The network of cross-references has not been built in an 'exhaustive' manner, as the reference lists might grow too long.

4. Example entry

The 3S entry contains: header (it contains the dominant member of the synonymic chain), general explanation of the dominant, antonyms (if any exist), members of synonymic chains, explanatory notes to define the relationships of the individual chain members toward the dominant, examples of synonym usage in a context, qualification labels (from the point of view of style, frequency, etc.) and references. Polysemous entries may contain several synonymic chains and/or several reference lists. An example entry is shown in figure 1.

štebotat’ vydávat’ jemný, tichý, príjemný zvuk (o vtákoch; pren. expr. i hovorit’) • **švitorit’:** *v kroví štebocú, švitoria drobné vtáčiky; deti štebocú, švitoria* • **šveholit’** (spevavo): *škovránok šveholí rannú pieseň* • expr. **ševelit’:** *vtáča ševelí* • expr.: **čvikotat’** • **čikotat’:** *lastovičky č(v)ikocú* • expr. **čipčat’** (vydávat’ piskľavý zvuk; aj o drobnej hydine): *kurence čipčia; mladé v hniezde čipčia* • **čvirikat’** • **čvrlikat’** (o vrabcoch) • **džavotat’** (aj o ľuďoch) • **trilkovat’** • **tidlikat’** • poet. **klokotat’** (vydávat’ trilky): *sláviky trilkujú, klokotajú*

Fig. 1: Example entry *štebotat’* ('to chirp')

The members of synonymic chains are indicated by **boldface**, examples are in *italics* and labels are set in smaller typeface. All other elements of an entry are in plain Roman.

5. Lexical Computing

The 3S dictionary belongs to those lexicographic projects where computer technology has been introduced in a relatively late phase of the project's life cycle. Basically, most of the draft text of the dictionary had already been prepared in a traditional 'paper and pencil' way, when the decision was made to speed up the dictionary-making process by using PC(s) in the final stages of the publication preparation.

A so-called 'late computational support' approach has been adopted. This methodology has been developed to cope with several on-going lexicographic projects at the Linguistics Institute. Two rather serious constraints had to be taken into consideration in designing this methodology. First, the Institute was (and, to a certain level still is) under-equipped with hardware and software resources, available computers tend to be of low computational power. Second, the level of 'computer literacy' among individual authors was rather low. On the other hand, the 'goodwill' of some authors was an important positive factor that helped greatly to computerize this project.

The key issues to be addressed from the computer scientist's point of view were as follows: (1) providing additional lexical evidence, (2) designing a scheme to represent the dictionary data, (3) validating the dictionary data, (4) creating the procedure for merging and alphabetization of the text, (5) preparing the final layout of the publication.

The main source of lexical evidence after the introduction of technology into the Project has been the machine-readable version of KSSJ *(Krátky slovník slovenského jazyka).* This is a concise explanatory dictionary that has been indexed by the WordCruncher corpus-processing package. The procedure of processing and indexing has been repeated iteratively (three times) to obtain the most suitable access mode for the dictionary compilers.

A simple markup language has been designed (Benko 1991) to represent the dictionary text and additional information needed in further data processing and validation. The markup language (MOM – 'my own markup') uses four types of objects that denote structure and/or typeface tags, special characters, extra accented characters and dic-tionary entry identifiers. Most of these objects are represented by one- or two-character sequences, as shown in figure 2.

!v183
"štebotat'" vydávat' jemný, tichý, príjemný zvuk (o vtákoch; pren. expr. i hovorit'), "švitorit'": 'v kroví štebocú, švitoria drobné vtáčiky; deti štebocú, švitoria', "šveholit'" (spevavo): 'škovránok šveholí rannú pieseň', |expr.| "ševelit'": 'vtáča ševelí', |expr.:| "čvikotat', čikotat'": 'lastovičky č(v)ikocú', |expr.| "čipčat'" (vydávat' pisk l'avý zvuk; aj o drobnej hydine): 'kurence čipčia; mladé v hniezde čipčia', "čvirikat', čvrlikat'" (o vrabcoch), "džavotat'" (aj o l'uďoch), "trilkovat', tidlikat'", |poet.| "klokotat'" (vydávat' trilky): 'sláviky trilkujú, klokotajú'

Fig. 2: Example entry in MOM notation

The data validating procedures included the use of a validating parser (based mostly on regular-grammar descriptions) to check common errors like misplaced punctuation and delimiters, unbalanced paired elements (brackets, start/close tags), incorrect sequence of meaning numbers, etc. The ad-hoc 'batch' validation procedures were designed to check errors like duplicate chain members, microstructure syntax violations, missing or incorrect qualifiers and erroneously 'merged' or 'split' entries. The most important validation procedure has been designed to check the completeness and correctness of the cross-reference network. Both 'straight' and reference entries were transformed into a uniform '*synonym* see *dominant'* representation that was loaded into a relational database. After comparison of the two database files, the matching pairs of references were marked as 'correct'. The superfluous entities on the 'straight' side were considered as new candidates for reference entries. The unmarked 'reference side' entities were (manually) checked to find the cause of error. This procedure was iteratively performed until all references could be marked as correct.

6. References

Benko, V. 1991. "Slovak Language Lexical Database (SLLD)", in: *Computational Lexicography (proceedings), Balatonfüred, 8–11 September 1990.* Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences.

Blanár, V. 1984. *Lexikálnosémantická rekonštrukcia.* Bratislava: VEDA.

Garaj, J. K. 1937. *Príručný slovník diferenciálny a synonymický.* Turčiansky sv. Martin: MS.

Pisárčiková, M. and Michalus, Š. 1973. *Malý synonymický slovník.* Bratislava: VEDA.

Pisárčiková, M. 1989. "Synonymia slovies so všeobecným významom", in: *Kultúra slova,* Vol. 23. Bratislava: VEDA.

DS 1991. *Dictionaire des Synonymes.* Paris: Roberts.

GLS 1974. *Das große Lexikon der Synonyme.* München: Wilhelm Heyne Verlag.

KSSJ 1989. *Krátky slovník slovenského jazyka.* Bratislava: VEDA.

MSS 1989. *Magyar Szinonimaszótár.* Budapest: Akadémiai Kiadó.

OT 1990. *The Oxford Thesaurus.* Oxford: OUP.

SS 1975. *Slovar' sinonimov, Spravočnoe posobie.* Moskva: Nauka.

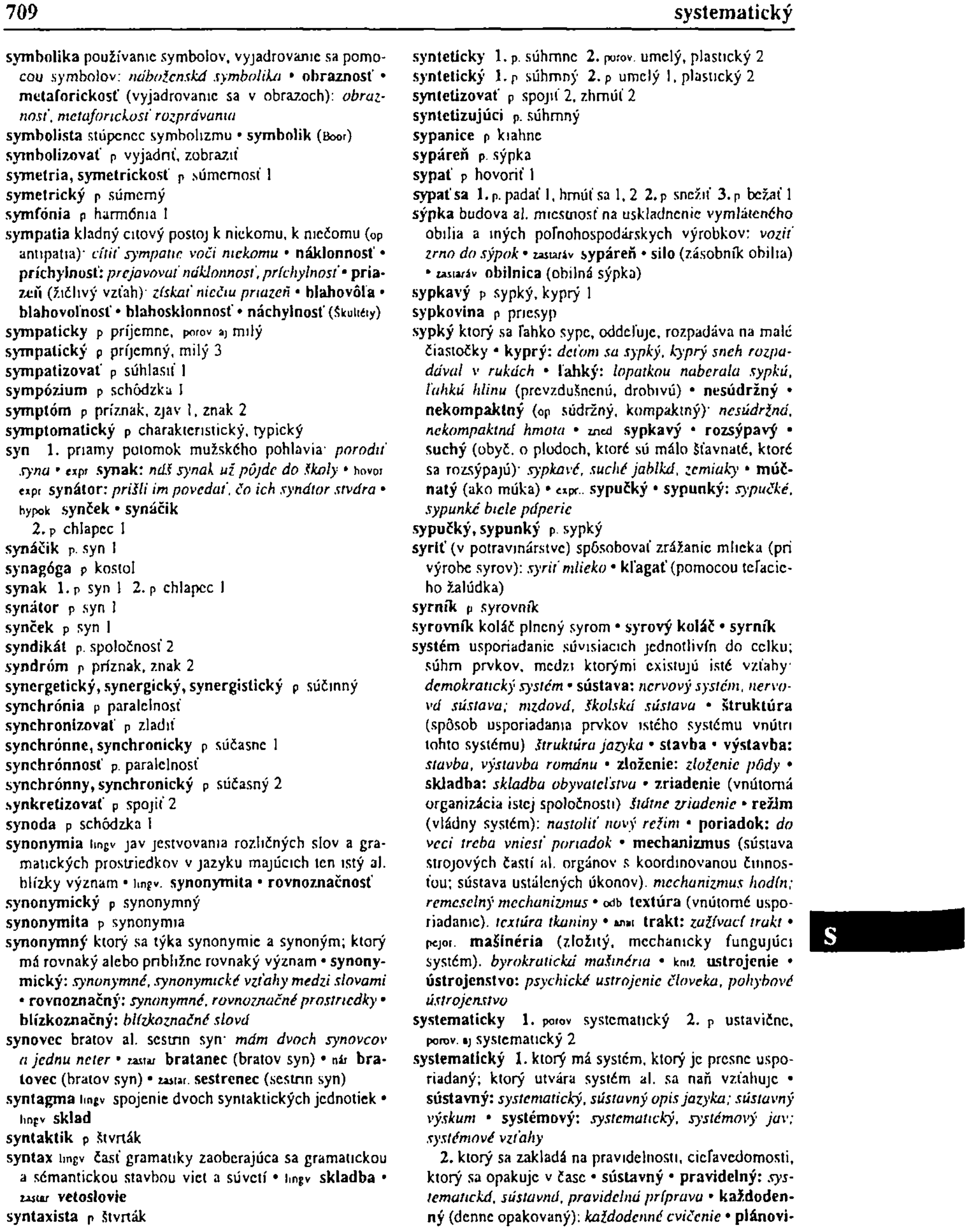
SSS 1995. *Synonymický slovník slovenčiny.* Bratislava: VEDA.

WNDS 1984. *Webster's New Dictionary of Synonyms.* Springfield: Merriam-Webster.

Acknowledgement

This work was supported, in part, by the Slovak Grant Agency for Science (grant No. 1274/94).

Appendix: Sample 3S page ('synonymy' and 'synonymous')



709 **systematický**

symbolika používanie symbolov, vyjadrovanie sa pomocou symbolov: *náboženská symbolika* • **obraznosť** • **metaforickosť** (vyjadrovanie sa v obrazoch): *obraznosť, metaforickosť rozprávania*

symbolista stúpenec symbolizmu • **symbolik** (Boor)

symbolizovať p vyjadriť, zobraziť

symetria, symetrickosť p súmernosť 1

symetrický p súmerný

symfónia p harmónia 1

sympatia kladný citový postoj k niekomu, k niečomu (op antipatia): *cítiť sympatie voči niekomu* • **náklonnosť** • **príchylnosť**: *prejavovať náklonnosť, príchylnosť* • **priazeň** (žičlivý vzťah): *získať niečiu priazeň* • **blahovôľa** • **blahovoľnosť** • **blahosklonnosť** • **náchylnosť** (Škultéty)

sympaticky p príjemne, porov aj milý

sympatický p príjemný, milý 3

sympatizovať p súhlasiť 1

sympózium p schôdzka 1

symptóm p príznak, zjav 1, znak 2

symptomatický p charakteristický, typický

syn 1. priamy potomok mužského pohlavia: *porodiť syna* • expr **synak**: *náš synak už pôjde do školy* • hovor expr **synátor**: *prišli im povedať, čo ich synátor stvára* • hypok **synček** • **synáčik**

2. p chlapec 1

synáčik p. syn 1

synagóga p kostol

synak 1. p syn 1 2. p chlapec 1

synátor p syn 1

synček p syn 1

syndikát p. spoločnosť 2

syndróm p príznak, znak 2

synergetický, synergický, synergistický p súčinný

synchrónia p paralelnosť

synchronizovať p zladiť

synchrónne, synchronicky p súčasne 1

synchrónnosť p. paralelnosť

synchrónny, synchronický p súčasný 2

synkretizovať p spojiť 2

synoda p schôdzka 1

synonymia lingv jav jestvovania rozličných slov a gramatických prostriedkov v jazyku majúcich ten istý al. blízky význam • lingv. **synonymita** • **rovnoznačnosť**

synonymický p synonymný

synonymita p synonymia

synonymný ktorý sa týka synonymie a synoným; ktorý má rovnaký alebo približne rovnaký význam • **synonymický**: *synonymné, synonymické vzťahy medzi slovami* • **rovnoznačný**: *synonymné, rovnoznačné prostriedky* • **blízkoznačný**: *blízkoznačné slová*

synovec bratov al. sestrin syn: *mám dvoch synovcov a jednu neter* • zastar **bratanec** (bratov syn) • nář **bratovec** (bratov syn) • zastar. **sestrenec** (sestrin syn)

syntagma lingv spojenie dvoch syntaktických jednotiek • lingv **sklad**

syntaktik p štvrták

syntax lingv časť gramatiky zaoberajúca sa gramatickou a sémantickou stavbou viet a súvetí • lingv **skladba** • zastar **vetoslovie**

syntaxista p štvrták

synteticky 1. p. súhrnne 2. porov. umelý, plastický 2

syntetický 1. p súhrnný 2. p umelý 1, plastický 2

syntetizovať p spojiť 2, zhrnúť 2

syntetizujúci p. súhrnný

sypanice p kiahne

sypáreň p. sýpka

sypať p hovoriť 1

sypať sa 1. p. padať 1, hrnúť sa 1, 2 2. p snežiť 3. p bežať 1

sýpka budova al. miestnosť na uskladnenie vymláteného obilia a iných poľnohospodárskych výrobkov: *voziť zrno do sýpok* • zastaráv **sypáreň** • **silo** (zásobník obilia) • zastaráv **obilnica** (obilná sýpka)

sypkavý p sypký, kyprý 1

sypkovina p priesyp

sypký ktorý sa ľahko sype, oddeľuje, rozpadáva na malé čiastočky • **kyprý**: *deťom sa sypký, kyprý sneh rozpadával v rukách* • **ľahký**: *lopatkou naberala sypkú, ľahkú hlinu* (prevzdušnenú, drobivú) • **nesúdržný** • **nekompaktný** (op súdržný, kompaktný): *nesúdržná, nekompaktná hmota* • zried **sypkavý** • **rozsýpavý** • **suchý** (obyč. o plodoch, ktoré sú málo šťavnaté, ktoré sa rozsýpajú): *sypkavé, suché jablká, zemiaky* • **múčnatý** (ako múka) • expr.. **sypučký** • **sypunký**: *sypučké, sypunké biele páperie*

sypučký, sypunký p. sypký

syriť (v potravinárstve) spôsobovať zrážanie mlieka (pri výrobe syrov): *syriť mlieko* • **kľagať** (pomocou teľacieho žalúdka)

syrník p syrovník

syrovník koláč plnený syrom • **syrový koláč** • **syrník**

systém usporiadanie súvisiacich jednotlivín do celku; súhrn prvkov, medzi ktorými existujú isté vzťahy: *demokratický systém* • **sústava**: *nervový systém, nervová sústava; mzdová, školská sústava* • **štruktúra** (spôsob usporiadania prvkov istého systému vnútri tohto systému) *štruktúra jazyka* • **stavba** • **výstavba**: *stavba, výstavba románu* • **zloženie**: *zloženie pôdy* • **skladba**: *skladba obyvateľstva* • **zriadenie** (vnútorná organizácia istej spoločnosti) *štátne zriadenie* • **režim** (vládny systém): *nastoliť nový režim* • **poriadok**: *do veci treba vniesť poriadok* • **mechanizmus** (sústava strojových častí al. orgánov s koordinovanou činnosťou; sústava ustálených úkonov). *mechanizmus hodín; remeselný mechanizmus* • odb **textúra** (vnútorné usporiadanie). *textúra tkaniny* • anat **trakt**: *zažívací trakt* • pejor. **mašinéria** (zložitý, mechanicky fungujúci systém). *byrokratická mašinéria* • kniž **ustrojenie** • **ústrojenstvo**: *psychické ustrojenie človeka, pohybové ústrojenstvo*

systematicky 1. porov systematický 2. p ustavične, porov. aj systematický 2

systematický 1. ktorý má systém, ktorý je presne usporiadaný; ktorý utvára systém al. sa naň vzťahuje • **sústavný**: *systematický, sústavný opis jazyka; sústavný výskum* • **systémový**: *systematický, systémový jav; systémové vzťahy*

2. ktorý sa zakladá na pravidelnosti, cieľavedomosti, ktorý sa opakuje v čase • **sústavný** • **pravidelný**: *systematická, sústavná, pravidelná príprava* • **každodenný** (denne opakovaný): *každodenné cvičenie* • **plánovi-**

S