# A New Grammatical Tagset for the Slovak-Chinese Parallel Corpus

Ľuboš Gajdoš

*Nový tagset pre Slovensko-čínsky paralelný korpus*

*Resumé* Cieľom príspevku je kritická analýza existujúceho tagsetu čínskej strany Slovensko-čínskeho paralelného korpusu (*Si-Hanku* 斯漢庫) a návrh nového. Predkladaný návrh vychádza z existujúceho tagsetu *LCMC* a prispôsobuje sa pravidlám a súboru značiek v tagsete Slovenského národného korpusu. V novom návrhu tagsetu sa kladie dôraz na klasifikáciu slovies a niektorých synsémantických slov.

*Abstract* The aim of this article is to critically analyse the existing tagset in the Slovak-Chinese parallel corpus (*Si-Hanku* 斯漢庫) and propose a new one. The present proposal is based on the existing tagset of the Lancaster Corpus of Mandarin Chinese, and adapts the rules and tags of the Slovak National Corpus. In developing the new tagset, more emphasis is placed on the classification of verbs and function words.

*Key words* Slovak, Chinese, Language · Corpus Linguistics, Parallel Corpus, Grammatical Tagset, Chinese POS-tagging

## 1 *Introduction*

Corpora have been used for several decades as a source of language data in linguistic research and second language acquisition. Requirements for linguistic research versus language teaching may differ, sometimes to the extent that these two approaches end up on opposite sides. One such approach might be using a

tagset—a compromise solution between a purely linguistic approach and language teaching.[1]

## 2 *The Slovak-Chinese Parallel Corpus*

Our department has been developing the Slovak-Chinese Parallel Corpus (*Si-Hanku* 斯漢庫, hereafter SHK)[2] since 2014[3] with the original intention of forming a balanced corpus. Currently (April 2015) the corpus is dominated by legal texts from the PRC.[4] However, in the future this disparity will be gradually eliminated in favour of other genres (registers), and the corpus is updated regularly. SHK is annotated with bibliographic, style-genre annotation and morphological tags. Currently the following genres (style-genre annotation) can be found and used as a single subcorpus:

- journalistic texts (tag: »baokan«)
- literary texts (tag: »wenxue«)
- legal texts (tag: »falv«)
- legal texts in *fanti zi* 繁體字 (tag: »ffalv«)
- professional texts (tag: »zhuanye«)

1 It might be, for example, a question of morphological annotation in terms of linguistics, which reflects the current state of research on language. On the other hand, it may be a requirement for simplification which cannot entirely conform with the purely linguistic point of view and which approximates the needs of the target language, in this case Slovak. For more information on annotations, see e.g. Cui Gang and Sheng Yongmei, »Yuliaoku zhong yuliao de biaozhu 語料庫中語料的標註« [Corpus Annotation], *Journal of Tsinghua University* 15,1 (2000), 89–94.

2 See more at: <fphil.uniba.sk/katedry-a-odborne-pracoviska/katedra-vychodoazijskych-studii/slovensko-cinsky-paralelny-korpus/>.

3 For more details, see Ľuboš Gajdoš, »Slovak-Chinese parallel corpus«, *Studia Orientalia Slovaca* 12,2 (2013), 313–317.

4 The current predominance of legal texts in the corpus is related to issues of quantitative research on this register.

SHK has been built with the help of open-source and free software tools[5] which are available for users through a web interface.[6]

## 3 *The Original Tagset*

At present the Chinese part of the SHK corpus uses the tagset of the Lancaster Corpus of Mandarin Chinese,[7] which is also used as a basis for the new tagset. By designing the new tagset (POS-tagging), we have followed the conventional division of the lexicon between functional (*xūcí* 虛詞) and lexical words (*shící* 實詞). Special attention has been paid to the category of verbs and functional words, which is not large but is very frequent.

## 4 *The New Version of the Tagset—General Principles*

The process of designing the new tagset may be divided into two steps: the first step is renaming the original tagset so it is in line with the SNC tagset[8] used in the Slovak part, where capital letters are obligatory and the second letter is optional. One positive aspect of searching the corpus is that it is possible to search only for a subset of verbs (e.g. Vl, Vc and Vd),[9] or, if necessary, for the whole category (e.g. V).[10] Essentially this first step is usually only the formal redesignation of the original tagset to the SNC tagset rules.

5 The Nosketch Engine is used. See <https://www.sketchengine.co.uk/>.

6 The corpus is only available from the local network (Eduroam) at <sihanku.fphil.uniba.sk /run_guest.cgi/first_form?>.

7 See *the Lancaster Corpus of Mandarin Chinese*: <www.lancaster.ac.uk/staff/xiaoz/lcmc/lcmc_tagset.htm> (last retrieval 12 April 2015).

8 See the Slovak National Corpus: <http://korpus.juls.savba.sk/morpho_en.html> (last retrieval 10 April 2015).

9 CQL query: [tag="Vl"]. When writing a query, CQL signs and symbols are used (Corpus Query Language), e.g. Boolean operators conjunction »&«, disjunction »|«, negation »!«. See at: <www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying> (last retrieval April 10, 2015).

10 CQL query: [tag = "V.*"]. Subset of verbs (e.g. l, c, d) may be replaced by the symbol »*«.

In the second step, we propose the following modifications:

- a separate annotation for verbs with the marker $LE_1$
- a separate annotation for modal verbs
- a separate annotation for resultative verbs
- a separate annotation for directional verbs
- a separate annotation for $LE_1$ and $LE_2$[11] from the original group of auxiliary words[12] (the new group contains just ZHE, GUO, ZAI and ZHENGZAI)
- a separate annotation for negative adverbs
- a subdivision of pronouns

### 4.1 *Pronouns*

The word class of pronouns is divided into three subclasses:

| Pronouns P | Token |
|---|---|
| interrogative (Pi) | ***shénme, zěnme, nà*** 什麼，怎麼，哪[13] |
| personal (Pp) | ***wǒ, nǐ, tāmén*** 我，你，他們 |
| demonstrative (Pd) | ***zhè, nà, gāi*** 這，那，該 |

### 4.2 *Verbs*

Verbs in general belong to the category of lexical words which displays the most grammatical categories, e.g. tense, aspect and mood (often as a part of verbal morphology). The annotation (classification) of Chinese verbs is based upon the traditional division of verbs into main and auxiliary verbs, though without any sign of semantic (hierarchic) criteria. The category of verbs varies widely in terms of syntactic properties. There are differences between main verbs and auxiliary verbs in terms of collocational preferences, different degrees of grammaticalization and the verb's relationship to dependent elements of structure (the transitive ***jíwù dòngcí*** 及物動詞 versus the intransitive ***bùjíwù dòngcí***

11 In the case of the enclitic LE, the division into $LE_1$ (as a temporal marker) versus $LE_2$ (as a modal particle), which is often the language teaching approach, is used.

12 CQL query: [tag = "u"].

13 The complete list of subclasses is prepared.

不及物動詞). Therefore, we consider a more detailed subdivision of verbs as one of the leading aspects of the new tagset. Based upon our previous linguistic research into some verbal categories and from the perspective of language teaching, it is useful to divide verbs (V) into:

- resultative verbs (Vr)[14]
- monosyllabic directional verbs (Vd)[15]
- modal verbs (Vm)
- verbs + the marker $LE_1$ (Vl)
- adverbs of degree + verbs (Vp)

Some parts of the mentioned subdivision can then be done with the corpus-driven approach (e.g. monotransitive or ditransitive).

### 4.3 *Modal Verbs*

As already mentioned, it is practical to set the category of modal verbs apart because of some special grammatical properties. This can be done mostly on formal principles.

| Modal Verbs (Vm) | Token |
|---|---|
| [can, able to] | ***néng, nénggòu*** 能, 能够 |
| [should, ought to] | ***yīnggāi, gāi, yīngdāng, yīng*** 應該, 該,[16] 應當, 應 |
| [may, be possible] | ***kěyǐ*** 可以 |
| [may, can] | ***huì*** 會 |
| [want] | ***yào*** 要[17] |

14 Included here are the verbs ***wán*** 完, ***chéng*** 成, ***zháo*** 著.

15 Because the position and the bonding between directional verbs and predicative verbs are different, the group only contains monosyllabic verbs of the second category, e.g. ***jìn*** , ***chū*** 进, 出, etc. For more information on the verbal aspect, see Ľuboš Gajdoš, »Verbálny aspect v čínštine« (Doctoral dissertation) [Verbal Aspect in Chinese] (Bratislava: Comenius University, 2010).

16 It may also be a demonstrative pronoun.

17 This category also includes the words ***xiǎng*** 想 and ***děi*** 得, but taking into consideration the ambiguity of these words (as a verb or modal verb and as a verb or a complement marker respectively) we excluded them.

| | |
|---|---|
| [need, have to] | *bìxū, xūyào* 必須, 需要 |
| [be willing to, would like] | *yuànyì, kěn* 願意, 肯 |
| [dare] | *gǎn* 敢 |

Marking modal verbs with an automatic tagger can lead to ambiguity. The solution in this case is a manual annotation.

### 4.4 *Verbs with LE-Marker* V1 *(Corpus-Driven)*

The collocability of verbs+grammatical markers (LE, GUO) is very interesting since the acquisition of LE-usage causes some trouble, and LE is one of the most frequent »words« in Chinese.[18] For that reason, it is desirable to set $LE_1$ apart from the category of modal particles and mark the verbs (adjectives) with the ability to co-occur with $LE_1$.

Marking the collocability of a verb/adjective can be done as set out below. In the existing corpus with an original tagset we can look for the collocation/concordance ($V+LE_1$) with the following query:
[word="了"] ([tag="r|q"]|[tag="r|m|q"]|[tag="a|u"]) [tag="n|ns|nt"] within <s/>.[19]
The result of the query (ordered by frequency occurrence—»node one left« in Nosketch engine) is the frequency list of $V+LE_1$.

### 4.5 *Resultative Verbs* (V1)

As shown by our previous research into the verbal aspect in Chinese and Slovak, resultative verbs have various collocational preferences.[20] That is to say, co-occurrence with main verbs is sometimes the result of semantic limitation caused by their original (lexical) meaning.

18 In most corpora, the enclitic LE is separately tokenized.

19 There are several solutions to this problem. Thus we present only one possibility.

20 For example, the verb *jiàn* 見 is only associated with a certain number of »collocation partners« (only with the verb type *verba percipiendi*), so it is appropriate to consider it as a part of a verb (suffix) and to tokenize it together with a verb. Similarly, it is also possible to characterize the verbs *zháo* 著, *huì* 會 and *dǒng* 懂 in the same manner.

The table shows the list of resultative verbs:

| Resultative Verbs (Vr) | Token |
|---|---|
| [arrive] | *dào* 到 |
| [live] | *zhù* 住 |
| [complete] | *wán* 完 |
| [fall] | *diào* 掉 |
| [good] | *hǎo* 好[21] |
| [become] | *chéng* 成 |

### 4.6 *Directional Verbs* (Vd)

Directional verbs (often called »complements« *qūxiàng bǔyǔ* 趨向補語) are subdivided into three groups:

- monosyllabic: *lái* 來, *qù* 去
- monosyllabic: *shàng* 上, *xià* 下, *jìn* 進, *chū* 出, *huí* 回, *guò* 過, *kāi* 開, *qǐ* 起
- disyllabic (a combination of the previous groups)[22]

The »stiffness« of the V+Vd[23] connection in terms of syntactic properties is not comparable to V+Vr.[24] Nevertheless, we assume that it is appropriate to separate the group of monosyllabic directional verbs because of their temporal-aspectual

21 Although the word *hǎo* 好 is an adjective, we believe that it shares inherent qualities with this group due to the partial grammaticalization. Therefore, it can be included.

22 Chinese linguists do not agree on the number of directional complements; some sources suggest 26—Meng Cong. 孟琮 & al., *Hanyu dongci yongfa cidian* 漢語動詞用法詞典 [Dictionary of Verb Usage in Chinese], (Beijing: Shangwu yinshuguan, 2003), and others 28—Liu Yuehua 刘月华 & al., *Shiyong xiandai Hanyu yufa* 實用現代漢語語法 [Practical Grammar of Chinese Language], (Beijing: Waiyu jiaoxue yu yanjia chubanshe, 2004). We have added the verbs *dào* 到 and *zǒu* 走 to the group of resultative complements.

23 A predicate verb may be often separated from the directional verb (directional complement) with the marker LE or a syntactic object.

24 The collocation V+Vr cannot be separated with the marker LE$_1$, ZHE, GUO or verbs. The collocation V+Vr may only be separated with the atonic marker DE. Nevertheless, this then becomes the potential mood (complement of potentiality).

influence on predicate verbs. There are three scenarios of a predicate verb and directional verb co-occurrence:

- V+Vd+O
- V+Vd1+O+$Vd_2$
- V+O+Vd

The list of monosyllabic directional verbs:

| Directional Verbs (Vd) | Token |
|---|---|
| [come] | *lái* 來 |
| [go] | *qù* 去 |
| [go up] | *shàng* 上 |
| [rise] | *qǐ* 起 |
| [open] | *kāi* 開 |
| [pass] | *guò* 過 |
| [go down] | *xià* 下 |
| [enter] | *jìn* 進 |
| [come out] | *chū* 出 |
| [return] | *huí* 回 |

### 4.7 Verba Percipiendi *(Verbs/Adjectives with Adverbs of Degree)* (Vp)

In the case of consistent tokenization and annotations, *verba percipiendi* are all those verbs that can co-occur with adverbs of degree. The subgroup of these verbs may be found with the following query:
[word="很|非常|挺|比較|最"][25] [tag= "v"]
and the manual control of results.

25 Evidently, there are other adverbs of degree in Chinese. This query does not include all of them.

### 4.8 *Grammatical Markers*

As already indicated in the previous section (verbs with LE-marker VI), it is desirable, and not only from a language-teaching point of view, to further separate the following »words« from the group of auxiliary words[26]:

- $LE_1$ as a separate tag
- $LE_2$ as a separate tag[27]

### 4.9 *Negative Adverbs*

From the category of adverbs [tag="d|o"] in the original tagset, we suggest detaching the following negative adverbs:

| Negative adverbs | Token |
|---|---|
| [not, do not] | ***bù*** 不 |
| [not, have not] | ***méi / méiyǒu*** 沒/沒有 |
| | ***bié*** 別 |

26 In the original tagset tag »u« [tag = "u"].

27 The dividing of $LE_1$ and $LE_2$ is done in the original tagset, although not always consistently (e.g. $LE_2$ before punctuation is tagged as a marker and not as a modal particle). Formally, it is possible to solve this problem with the following query:
[word = "了"] ([? tag = "ew"] | [! word = '. |, |: |! |? "]).

5 *A proposed New Tagset*

| Part of Speech (Word-Class) | | New Tag | Chinese Term | Original Tag (If Exists) |
|---|---|---|---|---|
| SUBSTANTIVE | | S | 名詞 | n |
| | temporal N. | St | 時間詞 | ng |
| | location N. | Sl | 處所詞 | s |
| | N. of locality | Sp | 方位詞 | |
| ADJECTIVE | | A | 形容詞 | a |
| | predicative[28] | Av | | |
| | non-predicative | An | | ag |
| PRONOUN | | P | 代詞 | r |
| | interrogative | Pi | 疑問代詞 | |
| | personal | Pp | 人稱代詞 | |
| | demonstrative | Pd | 指示代詞 | |
| VERB | | V | | v |
| | with the marker $LE_1$ | Vl | | |
| | copula | Vc | 關係動詞 | |
| | modal | Vm | 能源動詞 | |
| | directional | Vd | 趨向動詞 | |
| | Verba percipiendi | Vp | 狀態動詞 | |
| ADVERB | | D | 副詞 | d |
| PREPOSITION | | E | 介詞 | p |
| CONJUNCTION | | O | | c |
| PARTICLE | | T | 語氣助詞 | u |
| INTERJECTION | | J | | e |
| PUNCTUATION | | Z | | ew |
| NEGATIVE ADVERBS | | N | 否定副詞 | |

28 Adjectives with an adverb (for example *hěn* 很) as a predicate.

## 6 *Conclusion*

The proposed tagset (*LuSi* 魯斯) is the synthesis of our experience with using the corpus in language teaching (second-language acquisition) and linguistic research. We have not just followed a purely morphological approach. As the new tagset shows, it combines morphological and syntactic principles. We hope that after the new tagset is applied, we will be able to evaluate the benefits as well as the shortcomings of the new tagset. Equally interesting is the perspective of using the corpus in language teaching.

The presented tagset can be described as a corpus-based approach, although some steps (e.g. collocation of V+$LE_I$) are the results of a corpus-driven approach. In the next stage, we would like to adjust the new version of the *LuSi* tagset according to the results of the corpus-driven approach.

Despite the fact that similar results can be achieved via the original tagset, when considering the measure of effort and difficulty we modestly think that the proposed tagset brings some improvements compared to the original tagset. In addition, for students to have similar tagsets on both sides of the corpus seems to be advantageous too. In the future, we plan to further modify the morphological annotation, namely the more detailed classification of function words, adjectives and so on.

*Comenius University in Bratislava, Department of East Asian Studies*