

MAXIMILIANO E. ORLANDO

(Independent Researcher)

Language Models Plus Specialized Documentation: Building Regular Expressions that Match Chinese Grammar Patterns

Abstract

Previous research had suggested using authentic teaching materials, corpora and regular expressions in Chinese as a Foreign Language (CFL) courses. Where regular expressions are concerned, however, it can be argued that language teachers are not expected to be experts on language programming or on authoring digital applications when it comes to build them. Given the importance of digital text in the study of additional languages nowadays, this case study discussed the use of language models in combination with specialized documentation to learn about the construction of regular expressions. It is hoped this discussion will be useful for both CFL teachers and learners who are interested in learning to use these expressions. The target language is the *shì...de* (*shì...de* (是...的)) grammar patterns. The existence of different *shì...de* (*shì...de* (是...的)) grammar patterns made it possible to illustrate the benefits of using language models to build regular expressions that match Chinese grammar patterns. Indeed, in enhancing the accuracy of the technical aspects of regular expressions, language models can also depict nuances in Chinese language text. Nevertheless, these benefits come with challenges, which together with the limitations of this study are tackled in this paper.

Keywords: Chinese language, regular expressions, grammar patterns, *Python*, digital text

奥兰多·马西米利亚诺·埃德华多

结合语言模型和专业文档：构建匹配汉语语法模式的正则表达式

摘要

在汉语作为第二语言（CFL）教学中，既往研究建议使用真实的教学材料、语料库以及正则表达式。然而，就正则表达式的使用，语言教师无需成为编程或数字应用开发方面的专家。鉴于数字文本在当前的语言教学中的重要性，因此本文探讨了如何结合语言模型和专业文档以生成正则表达式的方法。希望本文的研究能对有意学习使用正则表达式的汉语教师和学习者提供参考。本文以“是...的”语法格式为研究对象。由于“是...的”存在不同格式，这为展示利用语言模型构建匹配汉语语法模式的正则表达式的优势提供了可能。实际上，语言模型不仅能提高正则表达式在技术层面的准确性，还能捕捉汉语文本中的细微差别。然而，这些优势也伴随着挑战。本文讨论了上述挑战与优势的同时，亦指出了本研究的局限性。

关键词： 中文，正则表达式，语法句式，*Python*，数字化课文

Introduction

In a discussion on the benefits of the application of task-based language teaching in the teaching of grammar in Chinese as a foreign language (CFL) contexts, Yuan (2025) argued that in many CFL courses taught in China there is an approach from which the following scenario is expected to result: Learners understand grammar rules, but they find it difficult to use them in real-world situations.¹ Therefore, Yuan (2025) highlighted the benefits of the use of authentic situations in foreign/second language teaching. These benefits have not only been emphasized in task-based language teaching, but also in the lexical approach.

In that approach, “noticing collocations and language patterns” and the use of language corpora² to that end play a fundamental role. Where electronic lan-

¹ Yuan 2025: 14.

² Torres Ramírez 2012: 249.

guage corpora are concerned, however, their use may require skills in handling concordancers. Indeed, the complexities of the query interface of an electronic corpus may be suspected to deter learners of an additional language who have little or no knowledge of corpus linguistics methodology from using it. In addition, the use of concordancers requires skills in the analyses of their outputs. Despite not making reference to these two situations specifically, Carter and Absalom (2023), for example, reported that some of the members of a group of tertiary students of Italian as an additional language participating in a study on the use of online resources did not employ corpora owing to the complexity of their use.³ Nonetheless, Carter and Absalom (2023) mentioned other sources of naturally occurring language learners of additional languages can make use of. Two cases in point are online blogs and websites.⁴

Finding a specific character such as *shì* (是) in a corpus of digital text can be easily accomplished by using a find bar. However, a more sophisticated tool may come into play when looking for grammar patterns such as the *shì...de* (是...的) ones. In this paper, a *shì...de* (是...的) grammar pattern is defined as the order in which groups of words are used in conjunction with *shì* (是) and with *de* (的) to perform one of the six functions discussed below, and the tool in question is regular expressions. Nonetheless, teachers of additional languages in general and teachers of Chinese as an additional language in particular are not expected to be experts on language programming or on authoring electronic applications that can later be used by students in the courses they teach. For this reason, it was decided to share a report on an exploratory case study in which the combination of specialized documentation and artificial intelligence facilitated the creation of regular expressions. Indeed, the objective of this paper is to report on the use of artificial intelligence in combination with specialized documentation with a view to constructing regular expressions that match the *shì...de* (是...的) grammar patterns and that can be used in simplified present day Chinese text. Nonetheless, there is a disclaimer to be made. Because the purpose of the present study is to report on the process of designing the regular expressions in question, it is the responsibility of the users of the end result of this study, i.e. of these regular expressions, to verify their reliability and effectiveness and to improve them to cater for these users' specific needs.

³ Carter-Absalom 2023: 59.

⁴ Carter-Absalom 2023: 53.

Theoretical background

In a review of pre-electronic and computerized corpora, Xu (2019) observed that the quantitative study of sentence patterns was rarely tackled in pre-computer time, whereas it was underdeveloped in the latter.⁵ Nonetheless, Xu (2019) underlined the importance of Shuhua Zhao's *Basic Sentence Patterns of Modern Chinese* project, where sentence types and grammatical categories were dealt with. A case in point was the “focus constructions 是...的 *shì...de*”.⁶ Xu (2019) also concluded that where the pedagogical implementation of corpora findings is concerned, there is still a lot of work to be done to adapt those findings “to classroom friendly teaching materials”.⁷ For example, in a discussion of the benefits of direct use and indirect use of Chinese corpora, Ištvanová (2021) noted “a limited number of teaching materials in the Slovak – Chinese language combination”⁸, and went on to observe that direct and indirect use of corpora may come in useful to enhance language instruction. Nonetheless, where the direct use of a given corpus is concerned, Ištvanová (2021) explained that this use entails some training in the handling of the interface of the corpus with a view to facilitating the interpretation of queries.⁹

There is no denying that the use of electronic corpora and concomitantly of the interfaces that accompany them should be considered while teaching an additional language in contexts where the target learners can take advantage of this use.¹⁰ Nevertheless, a traditional way of learning an additional language has been through printed materials, such as textbooks, literary works, and reference works among others. It is well known that nowadays some of these materials also come in electronic format. A case in point is electronic textbooks.

The search for grammar patterns in digital text can be achieved in different ways. One can read digital text with the express purpose of noticing the use of these patterns. A search bar can also come in useful in this case. For example, when studying the use of the *shì...de* (是...的) grammar patterns, a search bar can be employed to look for the occurrences of *shì* (是) or *de* (的). However, the problem that this tool may pose is that since *shì* (是) and *de* (的) are likely to appear in other grammar patterns, the user may be confronted by the time-consuming task of having to discard irrelevant ones. If the goal of a classroom

⁵ Xu 2019: 41.

⁶ Xu 2019: 41.

⁷ Xu 2019: 49.

⁸ Ištvanová 2021: 140.

⁹ Ištvanová 2021: 132.

¹⁰ Chen 2012.

task is strictly limited to the observation of the *shì...de* (是...的) grammar patterns in digital text, then the use of regular expressions is to come in useful.

A regular expression “is a sequence of characters that forms a search pattern”¹¹, and it allows the user to “specify the rules for the set of possible strings that [the user wants] to match”.¹² As regards studies that relate to Chinese, regular expressions have proved to be useful in various contexts. Yuan et al. (2021), for example, used regular expressions to remove private information from the dataset used to build *WuDaoCorpora*. Regular expressions were used to match, for instance, phone numbers and identity numbers. Lau et al. (2024) used a “rule-based approach utilizing simple string matching”¹³ instead of machine learning in a “tool for classifying Written Cantonese, Standard Written Chinese, and the intermediate varieties used by Cantonese-speaking users from Hong Kong”.¹⁴ In this approach, Cantonese and Standard Written Chinese features were defined in regular expressions, i.e. Cantonese and Standard Written Chinese features were identified by regular expression matches.¹⁵ Fang (2017) employed regular expressions to normalize text in the “prototype of a text analyzer for a rule-based Mandarin Chinese [text-to-speech] voice”.¹⁶

In a more pedagogical context, Gajdoš (2017) discussed the implications of using the Chinese monolingual corpus *Hanku*, and more specifically, the sub-corpus of legal text *zh-law* for language pedagogy. In that discussion, Gajdoš (2017) suggested that students of Chinese practise writing regular expressions to look for patterns of sentences.¹⁷ Owing to this suggestion, and to the notion that teachers of CFL can also take advantage of the use of regular expressions, the present study explores and reports on the use of artificial intelligence and specialized documentation to build regular expressions that match the Chinese *shì...de* (是...的) grammar patterns. The *shì...de* (是...的) grammar patterns have been chosen for three reasons.

The first reason is that these grammar patterns perform different grammatical functions. *Chinese Grammar Wiki* presents six uses. These will be introduced below in the order in which they are presented by *Chinese Grammar Wiki*. Examples that illustrate each use are present on *Chinese Grammar Wiki*'s site and are not repeated in this paper owing to the length of the present study.¹⁸

¹¹ W3Schools n.d.: *Python RegEx* section.

¹² Kuchling n.d.: *Introduction* section.

¹³ Lau et al. 2024: 28.

¹⁴ Lau et al. 2024: 24.

¹⁵ Lau et al. 2024: 27.

¹⁶ Fang 2017: 2.

¹⁷ Gajdoš 2017: 86.

¹⁸ Chinese Grammar Wiki n.d.: *The “shì... de” patterns: an overview* section.

1. The first use consists in omitting a noun and in using *de* (的) to represent it.
2. In the second use, the *shì...de* (是...的) grammar pattern contains a distinguishing word.
3. The third case is the classic pattern. It is explained that it “is used to draw attention to certain information in a sentence”, that it “is frequently used when asking or telling details about the past”, and that “most commonly it emphasizes time, manner, or place”.¹⁹
4. *Chinese Grammar Wiki* also describes the *shì...de* (是...的) pattern when it is used to explain “why you are doing something or what an object is used for”.²⁰
5. The fifth use is to talk about people’s occupation.
6. The last use is to emphasize affirmations.

Nonetheless, for methodological purposes, in this paper the author has taken the view that there exists a generic, high-level, parent *shì...de* (是...的) pattern. This pattern can be filled up with different components, which gives birth to the six patterns described above. As a result, within the context of this paper each of the patterns presented above will be regarded as a child pattern of the generic *shì...de* (是...的) pattern. The author has adopted this view to use a generic term, namely, the pattern *shì...de* (是...的) or the *shì...de* (是...的) pattern, when prompting the models. This was done with a view to prompting the models from the general to the specific. Indeed, it will be seen how the author has moved from the general in Exact Prompt 1, where the models are asked to create an HSK 3 text containing about 100 words and the pattern *shì...de* (是...的), and in Exact Prompt 7, where the models are asked to try this pattern, to the specific in Exact Prompt 9, where it is requested that the models find the emphatic structure. It should be noted that in this instance the term *structure* was used because even though the author had used the term *pattern* in his conversation with the models, *Gemini* came up with the term “classic ‘emphatic’ structure”, and *Qwen3.5-Plus* used the term “The *shì...de* (是...的) structure”.

The second reason why the *shì...de* (是...的) grammar patterns have been selected for the purpose of this paper is that they are expected to occur in a wide range of learning materials. This is because they can be used to perform several communicative functions. Finally, it can be claimed that these grammar patterns are also expected to be tackled at more than one level of a CFL program. This is because the complexity of the elements that are within these patterns may vary.

The methodology used to conduct the present exploratory study will be described in the next section.

¹⁹ Chinese Grammar Wiki n.d.: *The “shì... de” construction for emphasizing details* section.

²⁰ Chinese Grammar Wiki n.d.: *The “shì... de” construction for indicating purpose* section.

Methodology

This exploratory study began with a CFL learner's, namely, the author's, basic knowledge of regular expressions, and therefore with an elementary regular expression containing three groups and the non-greedy quantifier +?. A non-greedy quantifier matches "as little text as possible".²¹ For example, in

Wǒ shì zuò wǒ de fēijī lái de. (我是坐我的飞机来的。)

Translation: I came on my plane.

(shì)(\w+)(de), namely, the exact expression (是)(\w+)(的) matches *shì zuò wǒ de fēijī lái de* (是坐我的飞机来的, to be precise), whereas *(shì)(\w+?)(de)*, namely, the exact expression (是)(\w+?)(的), matches *shì zuò wǒ de* (是坐我的, to be exact), i.e. it matches as little text as possible. More information on the latter expression will be given below.

Afterwards, artificial intelligence models were prompted through a series of conversation strings with a view to examining the possibility of enhancing the original expression and acquiring new knowledge on the part of the user. These models initially were a free online version of the *Gemini 3* family and the free online *Qwen3.5-Plus*. Since artificial intelligence can make mistakes, the information supplied by these models was contrasted with that provided by specialist online sites.

Even though it was a free online model of the *Gemini 3* family that was used, a description of *Gemini 3 Flash Preview (Reasoning)* will be provided. Two salient features of this model are that it is a leading model in intelligence and notably fast, but somewhat verbose.²² It has a one million tokens context window.²³ It scores 51% on the *SciCode (Coding)*, whereas *Gemini 3.1 Pro Preview*, which leads the field, scores 59%.²⁴ The *Gemini 3* family was chosen owing to the following combination: It has a free online version; *Gemini 3 Flash Preview (Reasoning)* has a high score on the *Artificial Analysis Intelligence Index*. This score is 46.²⁵

Since the online *Qwen3.5-Plus* was prompted, a description of *Qwen3.5-397B-A17B (Reasoning)* and of *Qwen3.5-Plus* will be presented. *Qwen3.5-397B-A17B (Reasoning)* scores high on the *Artificial Analysis Intelligence Index* (its score is 45), but it is somewhat verbose.²⁶ It is an open-weight model,

²¹ Kuchling n.d.: *Greedy versus Non-Greedy* section.

²² Artificial Analysis n.d.: *Gemini 3 Flash Preview (Reasoning) Intelligence, performance & price analysis* section.

²³ *Ibid.*

²⁴ *Ibid.*

²⁵ *Ibid.*

²⁶ Artificial Analysis n.d.: *Qwen3.5 397B A17B (Reasoning) Intelligence, performance & price analysis* section.

and it has been reported to score more highly than *Gemini-3 Pro* on three coding-related benchmarks.²⁷ *Qwen3.5-Plus* is the hosted model, has a one million tokens context window and features “official built-in tools and adaptive tool use”.²⁸ *Qwen3.5-Plus* was selected because of the following features: It is free; it is online; *Qwen3.5-397B-A17B* has been reported to score well on coding-related benchmarks. However, when this paper was reviewed by the author, *Qwen3.6-Plus* had been launched and was used as well. Each model will be cited in the analyses.

Some of those regular expressions suggested by the above-mentioned models were tested on the online regular expression tester *regex101*. This tester has several interesting features, such as a section that explains each component of the target expression, a match information section that highlights each component of this expression as included in the target text, a quick reference section, and the possibility of selecting a programming language. As regards this language, *Python 3* was used. Indeed, the *Python* flavor was turned on on the *regex101* tester to validate the regular expressions. Chinese text produced by *Gemini* (*Appendix A*) and by *Qwen3.5-Plus* (*Appendix B*) was used during this process. The prompt used to generate this text was the following.

Exact Prompt 1:

“please give me an hsk 3 text containing about 100 words and the pattern 是...的”²⁹

The prompt referred to the parent grammar pattern *shì...de* (是...的) to give the models ample opportunity to generate text concerning any of the six grammar patterns described above.

Four limitations of this study will be discussed.

First, this paper reports on a single case study. Indeed, its findings apply to the language models the researcher employed. This limitation has been minimized by bringing up concepts that it is hoped will come in useful in future research that examines learners’ use of regular expressions as well as the application of corpora and regular expressions to teach and to learn Chinese. This leads to the next limitation.

In this study text generated by artificial intelligence has been used to ensure a high density of occurrence of *shì...de* (是...的) grammar patterns. It was thought that this density would facilitate testing regular expressions – in other words, that it would enable the realization of the exploratory nature of this

²⁷ Qwen Team 2026: *Qwen3.5: Towards native multimodal agents* section.

²⁸ *Ibid.*

²⁹ The author is citing the exact prompt used during the research. The pinyin of the characters used by the author is *shì...de* (是...的).

paper. However, the author is of the opinion that testing regular expressions on learning materials (either digitalized materials such as electronic textbooks or digital materials such as online articles) can be an interesting future endeavor.

A more technical limitation is that this study discusses the raw regular expressions that were pasted on *regex101's Regular Expression* bar. As a result, none of the following content will be included in the presentation of the target expressions: methods such as *match()*, module-level functions such as in *re.match()*, the use of the *r* prefix to avoid a *SyntaxError*³⁰ or the flags used such as the *global flag* and the *single line flag*. As a result, it is recommended that the content of this paper should also be checked on other platforms or on code editors.

Finally, *Chinese Grammar Wiki* describes situations in which *shì* (是) is optional. This is in the classic pattern and when *shì...de* (是...的) is used to emphasize affirmations. These cases have not been considered in this study since it has focused on the use of the full *shì...de* (是...的) grammar patterns.

Report and Discussion

The point of departure of this research is *(shì)(\w+?)(de)*, namely, the following exact regular expression: *(是)(\w+?)(的)*. It is known that the first group matches the character *shì* (是), whereas *\w* in the second group “matches any alphanumeric character; this is equivalent to the class *[a-zA-Z0-9]*”.³¹ Nonetheless, after testing the above-mentioned regular expression on the text from *Appendix A* and *Appendix B* on *regex101*, this tester explained that the second group matches “any word character in any script” and this is “equivalent to *[\p{L}\p{N}]*”. Not being familiar with the syntax of this character class, both models were consulted using the following prompt in *conversation string A*.

Exact Prompt 2:

“I am using your text to test regular expressions on regex 101. I use this regex: *(是)(\w+?)(的)*. Regex 101 says that *\w* match any word character in any script (equivalent to *[\p{L}\p{N}]*) What is this?”³²

As for *(shì)(\w+?)(de)*, namely, the author’s exact regular expression *(是)(\w+?)(的)*, *Qwen3.5-Plus* took a sentence from *Appendix B*, i.e. from the text it

³⁰ Kuchling n.d.

³¹ Kuchling n.d.: *Matching Characters* section.

³² The author is citing the exact prompt used during the research. The pinyin of the characters used by the author is *shì* (是) and *de* (的). These characters are employed in the expression *(是)(\w+?)(的)*.

had already generated, and explained why this expression should work. This is shown in Table One. Please note that the model did not use pinyin, and that its exact output was cited in this table.³³

Aspects of regular expression	Model’s exact output except for the translation
Model’s sentence	我们是星期六早上出发的。
Translation of the model’s sentence	We set off on Saturday morning.
Explanation of the first group	(是) matches: 是
Explanation of the second group	(\w+?) matches: 星期六早上出发 (Because these are all Unicode “letters”)
Explanation of the third group	(的) matches: 的
Model’s warning	While this works on regex101, be careful if you use this regex in actual programming code: Python 3: \w matches Unicode characters by default. (Your regex will work). JavaScript: \w does not match Chinese characters unless you use the Unicode flag (/u).

Table 1. Qwen3.5-Plus’s explanation of the initial regular expression

Where *regex101*’s feedback regarding `[p{L}p{N}]` is concerned, *Gemini* explained that `p{L}` “matches any character in any language that is classified as a ‘letter’”, and that `p{N}` “matches any kind of numeric character in any script”. *Qwen3.5-Plus* explained that when *Unicode* is enabled, `p{L}` “stands for Unicode Letter”, and `p{N}` “stands for Unicode Number”. This content was validated on *The Modern JavaScript Tutorial* site, where it is explained that *Unicode* properties written as `p{...}` can describe a character’s category, that `p{L}` “denotes a letter in any language”, and that `N` is an alias for number.³⁴

It is also interesting to point out that *Qwen3.5-Plus* issued a warning which concerns the use of flags: in *JavaScript* you need to use the `/u` flag to match Chinese characters, whereas in *Python 3* `\w` matches *Unicode* characters by default.

³³ The pinyin of the characters used by the model in Table One is *women shì xīngqīliù zǎoshang chūfā de* (我们是星期六早上出发的), *shì* (是), *xīngqīliù zǎoshang chūfā* (星期六早上出发), and *de* (的).

³⁴ Kantor 2022: *Unicode: flag “u” and class p{...}* section.

However, in the original setting of the *Python* flavor on *regex101* that the author had used, the *Unicode* flag, namely, *u*, was used. Nonetheless, *regex101* informed that this was redundant in *Python 3*. Given the warning issued by *Qwen3.5-Plus* and the message given by *regex101*, the author still wanted to query *Gemini* on this issue through a prompt that belonged in *conversation string A1*. This prompt is cited below.

Exact Prompt 3:

“I am using *Python* on *regex101* (“flavor is set to *Python*”). I turned on the “*u*” (*Unicode*) flag, but *regex101* says: *u* modifier: *unicode*. Make *\w*, *\W*, *\b*, *\B*, *\d*, *\D*, *\s* and *\S* perform matching with *Unicode* semantic (redundant in *Python 3*). Do I really need this flag to match Chinese characters?”

When asked if the *Unicode u* flag was needed to match Chinese characters while using *Python 3* on *regex101*, *Gemini* explained that this was not so because in *Python 3* “all strings are native *Unicode* by default”. This statement was validated by the *Python 3.14.3 documentation*, where it says that “in *Python 3*, *Unicode* characters are matched by default for *str* patterns”, and that the *re.UNICODE* “flag is therefore redundant with no effect and is only kept for backward compatibility”.³⁵ Where backward compatibility is concerned, Davies (2014), for example, tackled the use of the *u* prefix in *Python 2* while building a site for learning Chinese characters.³⁶

Another topic was brought up by the models in *conversation string B*. This was matching only Chinese text. This happened when they were asked to build regular expressions for Chinese grammar patterns through the prompts that are shown below.

Exact Prompt 4:

“Do you know about regular expressions?”

Gemini answered “I certainly do!”, provided a cheat sheet with some of the most common characters used in these expressions, supplied a real-world example, and offered help. *Qwen3.5-Plus* answered “Yes, I’m familiar with regular expressions (*regex*)!”, provided three samples, and offered help. This was followed by the next prompt.

Exact Prompt 5:

“Ok. Can you help me build *regex* expressions to build Chinese patterns?”

At that point, the use of the script `\p{Script=Han}` was suggested by *Gemini*. However, this script failed when the *Python* flavor was selected on *regex101*.

³⁵ Python Software Foundation n.d.: *Module Contents* section.

³⁶ Giles 2014.

As a result, *Gemini* and *Qwen3.6-Plus* were prompted in *conversation string C*. This prompt is cited below.

Exact Prompt 6:

“Do we use this in Python 3: `\p{Script=Han}`?”

Both models agreed that using the third party *regex* module makes it possible to use this script in *Python 3*. When browsing the Web to get more details about this information, it was observed that several regular expressions had also been suggested to match Chinese characters.³⁷ Where the suggested script is concerned, Ayaka (n.d.) pointed out that `\p{sc=Han}` does not only match Chinese characters. Indeed, Goyvaerts (2025) explains that “Unicode offers the Hiragana, Katakana, Han, and Latin scripts that Japanese documents are usually composed of”³⁸, i.e. Han scripts that are used in Japanese. In addition, the *Unicode Consortium* (2024) explain that “the Unicode Standard leaves the issues of language tagging and word recognition up to a higher level of software and does not attempt to encode the language of the Han characters”.³⁹

Nonetheless, the objective of the present paper is not to delve into the limitations of the regular expressions or scripts that match Chinese characters. Nor is it to look into the historical analysis of rare or historic characters, or to discriminate between Chinese characters from the characters of other languages, but to report on the use of artificial intelligence in combination with specialized documentation with a view to constructing regular expressions that match the *shì...de* (是...的) grammar patterns and that can be used in simplified present day Chinese text.

Even though the script `\p{Script=Han}` failed when the *Python* flavor was selected on *regex101*, the author observed that this script worked on *regex101* when the *JavaScript* flavor was selected and when the *Unicode* flag was on. However, it will just be pointed out that a problem that was encountered while using `\p{Script=Han}` on *regex101* (more specifically, on *Java*, *PCRE2 (PHP)* or *JavaScript*) was naturally that it did not match numbers. However, the presence of numbers between *shì* (是) and *de* (的) in online Chinese newspaper articles should come as no surprise, and its study can be very interesting to further the knowledge of Chinese grammar. For example, the grammar pattern containing *shì* (是) and *de* (的) in the following sentence can be used to contrast this sentence’s grammar pattern with the classic *shì...de* (是...的) grammar pattern.

³⁷ Ayaka, n.d.; Fu, 2021.

³⁸ Goyvaerts 2025: *Unicode Scripts and Script Extensions* section.

³⁹ Unicode Consortium 2024: Chapter 18, *Distinguishing Han Character Usage Between Languages* section.

Zhè shì 2025 nián 1 yuè 2 rì pāishè de mǎ lì zhèn “yún zhōng qiúchǎng” (wú rén jī zhàopiàn).

(这是2025年1月2日拍摄的马利镇“云中球场”(无人机照片)。)⁴⁰

Translation: This is a January 2nd 2025 picture of Mali Town’s “Field in the Clouds” (a drone photo).

Owing to the situation described above, it was decided to continue to use the \w character. Nevertheless, after *Exact Prompt 5* in *conversation string B*, the author sent the following prompt. The slip of the pen in the spelling of *Let* was kept for transparency.

Exact Prompt 7:

“LEt us try this pattern: 是...的”⁴¹

When asked to use regular expressions to match the *shì...de* (是...的) pattern, both models proposed *shì.+?de*, namely, the exact expression 是.+?的, where “.” is often used where you want to match ‘any character’”, and “+ requires at least *one* occurrence”⁴². The non-greedy quantifier *+?* matches “as little text as possible”.⁴³ *Qwen3.5-Plus* described the components of the suggested expression succinctly. This is shown in Table Two. Please note that the model did not use pinyin, and that the exact output was cited in this table.⁴⁴

Aspects of regular expression	Model’s exact output except for the translation:
Explanation of the first item: <i>shì</i> (是)	Literal character “是”
Explanation of the second item: <i>.+?</i>	Match any character, non-greedy (stops at the first 的)
Explanation of the third item: <i>de</i> (的)	Literal character “的”
Model’s sentence	我是昨天来的，他是坐飞机来的
Translation of the model’s sentence	I arrived yesterday, he came by plane.

⁴⁰ Xinhuanet 2026: 2025年新华社年度照片·体育新闻 (The author added the pinyin version).
⁴¹ The author is citing the exact prompt used during the research. The pinyin of the characters used by the author is *shì...de* (是...的).
⁴² Kuchling n.d.: *Matching characters* section and *Repeating things* section.
⁴³ Kuchling n.d: *Greedy versus Non-Greedy* section.
⁴⁴ The pinyin of the characters used by the model in Table Two is *shì* (是), *de* (的), *wǒ shì zuótiān lái de*, *tā shì zuó fēijī lái de* (我是昨天来的，他是坐飞机来的), *shì zuótiān lái de*, *tā shì zuó fēijī lái de* (是昨天来的，他是坐飞机来的), *shì zuótiān lái de* (是昨天来的), and *shì zuó fēijī lái de* (是坐飞机来的).

Explanation of the greedy component: <i>shì.+de</i> , namely, 是.+的	matches: 是昨天来的，他是坐飞机来的 ✗ (too long!)
Explanation of the non-greedy component: <i>shì.+?de</i> , namely, 是.+?的	matches: 是昨天来的 ✓ and 是坐飞机来的 ✓

Table 2. Qwen3.5-Plus’s explanation of *shì.+?de*, namely, the exact expression 是.+?的

As a result, *shìde* (是的) cannot be matched because at least one occurrence is required in between. The models also issued a series of warnings. *Qwen3.5-Plus* suggested adding the *re.DOTALL* flag in case content spanned lines, which made the author activate the *single line* (*Dot matches new-line*) flag on *regex101*. In addition, if using *shì.+de*, namely, the exact expression 是.+(without the non-greedy ?), *Gemini* recommended establishing the number of characters to be matched by replacing + with the {m,n} quantifier⁴⁵ “to limit the number of characters allowed in the middle”. For example, *shì.{1,5}de*, namely, 是.{1,5}的 to be exact, matches *shì qí zìxíngchē qù de* (是骑自行车去的 to be precise (went by bike)) because there is at least one repetition and at most five.⁴⁶ However, it will not match *shì zuò dìtiě lái de* (是坐地铁来公司的 to be exact (came to the company by subway)) because there are six repetitions in this chunk.

Furthermore, it was observed that in the sentence

shì de, tā shì zuò dìtiě lái de (是的，他是坐地铁来的 to be exact),

Translation: Yes, he came by subway.

shì.+?de, namely, the exact expression 是.+?的, would match the whole chunk because at least one occurrence is required. Nevertheless, *Qwen3.5-Plus* suggested different regular expressions in addition to *shì.+?de*, namely, the exact expression 是.+?的. One of them was *shì.*?de*, namely, the exact expression 是.*?的. This expression matched *shìde* (是的 to be precise), and *shì zuò dìtiě lái de* (是坐地铁来的 to be exact) separately because, as *Qwen3.5-Plus* put it, it allows “empty content”. Indeed, “* allows zero or more characters” (*Qwen3.5-Plus*) as against +, where at least one occurrence is required. In other words, * “specifies that the previous character can be matched zero or more times, instead of exactly once”.⁴⁷

Another suggestion was made by *Gemini* in *conversation string D*, where the author prompted the following.

⁴⁵ Kuchling n.d.: *Repeating things* section.

⁴⁶ Kuchling n.d.: *Repeating things* section.

⁴⁷ Kuchling n.d.: *Repeating things* section.

Exact Prompt 8:

“I have the following sentence: 是的，他是坐地铁来的。The following regular expression 是.+?的 matched the segment that goes from the first 是 to the end. I want to match 是的 and 是坐地铁来的 separately in Python. Can you please give me a suitable regular expression?”⁴⁸

The use of a character class starting with ^ was suggested in *shide|shì[^de]+de*, namely, in the exact expression 是的|是[^de]+的. This matches “the characters not listed within the class by *complementing the set*”.⁴⁹ Gemini’s explanation of this regular expression is provided in Table Three. Please note that the model did not use pinyin, and that the exact output was cited in this table.⁵⁰

Aspects of regular expression	Model’s exact output
Explanation of the first item: <i>shide</i> (是的)	It looks for the exact character 是 immediately followed by the exact character 的
Metacharacter ⁵¹	acts like an or operator in logic
Explanation of the second item: <i>shì</i> (是)	The engine looks for the character 是
Explanation of the third item: <i>[^de]</i> , namely, <i>[^de]</i>	Any character that is not 的
Explanation of the fourth item: +	One or more times
Explanation of the third and the fourth items combined: <i>[^de]+</i> , namely, <i>[^de]+</i>	Tells the engine to match a continuous stream of text, but it must stop the moment it hits the character 的.
Explanation of the last item: <i>de</i> , namely, 的	Finally, the engine looks for the closing 的 character right after that stream.

Table 3. Gemini’s explanation of *shide|shì[^de]+de*, namely, the exact expression 是的|是[^de]+的

Afterwards, the author decided to move on to a specific *shì...de* (是...的) grammar pattern: the classic one or that which emphasizes details.⁵² This was in *conversation string B*, i.e. after the models proposed *shì.+?de*, namely, the exact

⁴⁸ The author is citing the exact prompt used during the research. The pinyin of the characters used by the author is *shì* (是), *shì de*, *tā shì zuò dìtiě lái de* (是的，他是坐地铁来的), *shide* (是的), *shì zuò dìtiě lái de* (是坐地铁来的), and *shì* (是) and *de* (的) in the expression 是.+?的.

⁴⁹ Kuchling n.d.: *Matching Characters* section.

⁵⁰ The pinyin of the characters used by the model in Table Three is *shì* (是) and *de* (的).

⁵¹ Kuchling n.d.: *More metacharacters* section.

⁵² Chinese Grammar Wiki n.d.: *The “shì... de” construction for emphasizing details* section.

expression是+?的. Therein, *Qwen3.5-Plus* offered to “extract the emphasized part (e.g., get ‘昨天来’ from ‘我是昨天来的’)”⁵³, and *Gemini* asked what the goal for the *shì...de* (是...的) pattern was. The answer is given below.

Exact Prompt 9:

“to find the emphatic structure.”

Interesting warnings were made. Perhaps the most relevant one was that which pointed out that *shì...de* (是...的) also has other uses. *Gemini* expressed “to find the emphatic structure (是...的) specifically, we need to be careful. In Chinese, 是 (is) and 的 (possessive/adjective marker) appear constantly in non-emphatic contexts”.⁵⁴ *Qwen3.5-Plus*, for example, pointed out the importance of distinguishing the use of *shì* (是) and *de* (的) for possession and for emphasis.

Gemini offered several suggestions to match the classic grammar pattern specifically. It highlighted that “the emphatic structure usually highlights Time, Location, or Manner”, and suggested using *shìzài* (是在 to be precise), at the beginning when the target is location, adding a verb such as *zuò* (做 to be precise) in *zuòde* (做的 to be exact) at the end when the target is location, and incorporating a verb such as *lái* (来 to be precise) in *láide* (来的 to be exact) at the end when the target is manner. This is expressed in Table Four as quoted from *Gemini*. Indeed, please note that the model did not use pinyin, and that its exact output was cited in this table.⁵⁵ However, a column for translations was added by the author.

Target Emphasis	Regex Pattern	Example Match	Translation
Time (Years/Hours)	是(\d+年 \d+点)的	是2022年的	Belongs to the year 2022
Location (At/In)	是在([\u4e00-\u9fa5]+)做的	是在办公室做的	Was done in the office
Manner (How)	是([\u4e00-\u9fa5]{1,3})来的	是开车来的	It was by car that [someone] came

Table 4. Gemini’s description of the emphatic structure

⁵³ The author is citing the exact model’s output. The pinyin of the characters used by the model is *zuótiān lái* (昨天来) and *wǒ shì zuótiān lái de* (我是昨天来的).

⁵⁴ The author is citing the exact model’s output. The pinyin of the characters used by the model is *shì* (是), *de* (的), and *shì...de* (是...的).

⁵⁵ The pinyin of the characters used by the model in Table Four is *shì 2022 nián de* (是2022年的), *shì zài bàngōngshì zuò de* (是在办公室做的), and *shì kāichē lái de* (是开车来的). In addition, *shì* (是), *de* (的), *diǎn* (点) and *nián* (年) in是(\d+年|\d+点)的, and *shì* (是), *zài* (在), *zuò* (做), and *de* (的) in是在([\u4e00-\u9fa5]+)做的. Finally, *shì* (是), *lái* (来) and *de* (的) in是([\u4e00-\u9fa5]{1,3})来的.

Gemini showed an inaccuracy: It is not clear whether *shì*(*d*+*nián*|*d*+*diǎn*)*de*, namely, 是(*d*+年|*d*+点)的 to be exact, where one or more digits (*d*+) is followed by *nián* (年 to be exact) or *diǎn* (点 to be precise) is an emphatic pattern. In *conversation string D*, however, *Qwen3.6-Plus* expressed a few limitations concerning the manner pattern. This was done after exact prompt 10.

Exact prompt 10:

“please explain 是([\u4e00-\u9fa5]{1,3})来的”.⁵⁶

Qwen3.6-Plus explained that the manner pattern accepts up to three characters only ({1,3}), and [\u4e00-\u9fa5] excludes rare archaic characters amongst other items. In addition, it made a smart comment where it argued that the manner pattern “may match semantically odd phrases like 是苹果来的 if not filtered by a downstream NLP step (e.g., named entity recognition)”.⁵⁷ Indeed, named-entity-recognition technology has the power of looking into semantic aspects of language.

Gemini also highlighted a *punctuation* problem. It explained that “in the emphatic structure, the 的 is often at the very end of the sentence” and suggested a strategy “if you want to be 100% sure you are catching the Grammatical Structure and not just a random ‘是...的’ phrase”.⁵⁸ This was *shì*(.+?)*de*(?=[° ! ?]|\s|\$), namely, the exact expression 是(+?)的(?=[° ! ?]|\s|\$), and it is described in Table Five.

Aspects of regular expression	Model’s exact output except where otherwise indicated
Explanation of the first item: <i>shì</i> (.+?) <i>de</i> , namely, the exact expression是(+?)的	Please, refer to Table Two (the author’s note).
Positive lookahead assertion: (?=...) ⁵⁹	Only match this if it’s followed by
Explanation of the second item: [° ! ?]	a Chinese period, an exclamation/question mark
Metacharacter ⁶⁰	Please, refer to Table Three (the author’s note).

⁵⁶ The author is citing the exact prompt used during the research. The pinyin of the characters used by the author is *shì* (是), *lái* (来) and *de* (的) in是([\u4e00-\u9fa5]{1,3})来的.

⁵⁷ The author is citing the exact model’s output. The model used *shì píngguǒ lái de* (是苹果来的 to be precise).

⁵⁸ The author is citing the exact model’s output. The pinyin of the characters used by the model is *de* (的) and *shì...de* (是...的).

⁵⁹ Kuchling n.d.: *Lookahead assertions* section.

⁶⁰ Kuchling n.d.: *More metacharacters* section.

Explanation of the third item: \s	a space
Metacharacter ⁶¹	Please, refer to Table Three (the author's note).
Explanation of the fourth item: \$	the end of the string.

Table 5. Gemini's explanation of `shì(+?)de(?=[ ° ! ? ]\s|$)`, namely, the exact expression 是(+?)的(?=[° ! ?]\s|\$)

One may have the impression that these analyses could go on ad infinitum. Indeed, making up regular expressions seems to be an endless task owing to the fact that catches may constantly appear. Some of them are spotted immediately; others are identified later during the testing process. As already explained, the purpose of this paper is not to present a perfect regular expression or systems of expressions to match the *shì...de* (是...的) grammar patterns. On the contrary, it is to report on the use of artificial intelligence and specialized documentation to construct these expressions.

In using artificial intelligence and specialized documentation to construct regular expressions, it can be observed that language models such as the ones used in this paper can perform several interesting functions where the study of the *shì...de* (是...的) grammar patterns is concerned.

1. The most obvious function is the presentation of regular expressions for the user to test them for their validity. Two cases in point were *shì.+?de*, namely, the exact expression 是.+?的, and *shì(+?)de(?=[° ! ?]\s|\$)*, namely, the exact expression 是(+?)的(?=[° ! ?]\s|\$).
2. More interestingly, another function is the presentation of nuances not only of the components of regular expressions (*Qwen3.6-Plus*, for example, warned about the limitations of {1,3}), but also of the language these expressions match (*Gemini*, for example, suggested avoiding mixing up the emphatic and the possessive patterns containing *shì* (是) and *de* (的)). However, the following must be borne in mind. Even though language models should be expected to supply accurate information about regular expressions and about the language they match in response to users' prompts, they may make mistakes. In this regard documentation plays an essential role. This is to validate the content produced by language models. Indeed, the range of regular expressions that language models propose can be compared with information as presented by existing technical or linguistic specialized documentation. That is, specialized documentation

⁶¹ Kuchling n.d.: *More metacharacters* section.

is expected to present accurate information about the components of regular expressions or about the language they match. Specialized documentation is also to illustrate this information with examples.

3. Language models can also help users advance their knowledge of regular expressions by presenting expressions that are more complex than the ones they are already familiar with. This can encourage users to move beyond their comfort zone. For example, *Gemini* introduced the concept of *lookahead*, which was later validated with specialized documentation.
4. In dealing with regular expressions, language models can also help users advance their knowledge of the Chinese language by means of the presentation of the nuances alluded to above. An example of these nuances is the different communicative functions that the *shì...de* (是...的) grammar patterns can perform. For example, in Table Four, *Gemini* explained the focus on manner, place or time where the emphatic pattern is concerned.
5. Language models can also make users aware of catches inside regular expressions. For example, *Qwen3.6-Plus* explained that *shì*(`[u4e00-\u9fa5]{1,3}`)*láide*, namely,是(`[u4e00-\u9fa5]{1,3}`)来的 to be precise, match odd phrases such as *shì píngguǒ lái de* (是苹果来的 to be exact (the author leaves the translation of this text to the readers' imagination)).

Conclusion

It can be concluded that the most complete function language models such as the ones discussed in this paper can perform while being used in conjunction with specialized documentation to construct regular expressions that match Chinese grammar patterns is the following: To enhance regular expressions while delving into the nuances that Chinese grammar patterns may convey. However, it must be said that the use of language models to make up regular expressions that match Chinese grammar patterns comes with challenges and requirements.

One of these challenges is the readiness to spot inaccurate information. For example, in Table Four, *Gemini* presented an example, *shì 2022 niánde* (是2022年的 to be exact), which was not clearly emphatic. Indeed, in the present context, knowledge of the nuances of the target grammar patterns or the ability to learn about these nuances from the text matched by the regular expressions provided by the models is required to analyze these models' information critically. For example, if the target grammar pattern is the classic one, the user can explore whether the text matched by the regular expressions provided by the models actually agrees with this pattern. As a result, in dealing with the nuances of the target grammar patterns, documentation that concerns Chinese grammar comes

into play. To illustrate this, one can cite *Chinese Grammar Wiki* to show that the use of the *shì...de* (是...的) grammar patterns transcends that of expressing emphasis on details concerning past events.

Another challenge is the complexity that is inherent in regular expressions. Indeed, one may argue that the use of regular expressions requires at least a basic understanding of the programming language that is selected to implement them. Information about this language can be provided by specialized documentation and should come in useful when regular expressions are tested. Specialized documentation also serves as a source that can be consulted to confirm the accuracy of the information provided by language models. As a result, reviewing specialized documentation may also be required with a view to succeeding in understanding regular expressions. These considerations lead to several pedagogical implications where the use of regular expressions to teach and learn Chinese is concerned.

The language models proved to be skillful in their presentation of regular expressions in spite of the above-mentioned challenges. They expressed nuances regarding not only the functions Chinese grammar patterns may perform, but also the text those expressions match. In so doing, they provided options, illustrated them, and explained them. As suggested above, users that employ these models to build regular expressions therefore should have at least some basic knowledge of the programming language selected to implement those expressions. This translates into learning about the components of these expressions, namely, of their code, and, as a result, into analyzing the range of data provided by models critically. Training that fosters this knowledge can make use of specialized documentation. Future studies could delve into the implications of providing this training. Indeed, reports on teachers’ and learners’ opinions of traversing Diagram A should provide interesting input in regard to the convenience and challenges of using language models and specialized documentation to build regular expressions.

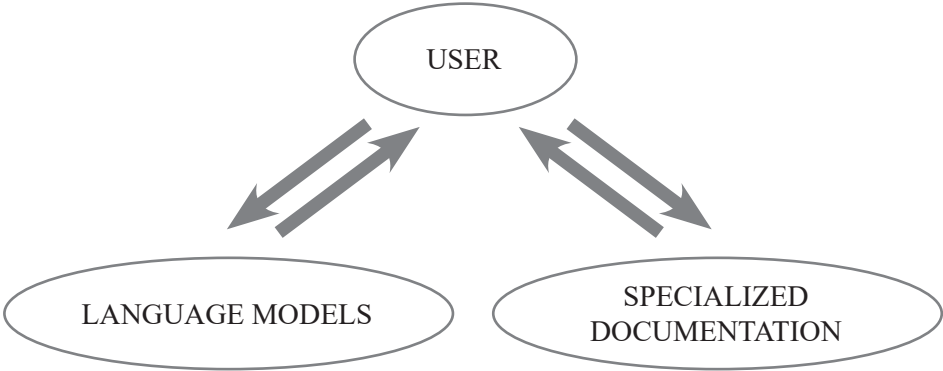


Diagram A. Regular expression learning graph

In Diagram A, the point of departure of the learning process can be a regular expression provided by a language model. The user consults specialized documentation that both explains and validates the features of this expression. Alternatively, a regular expression is suggested by specialized documentation. The user consults a language model to obtain further details about that expression, and then validates these details with specialized documentation. Another point of departure is the user, who comes up with a regular expression (possibly after consulting a language model or specialized documentation), and validates it by consulting a language model and specialized documentation. It must be highlighted that throughout these interactions the user tests regular expression candidates. Indeed, this paper closes by considering further research on the feasibility of using regular expressions to study Chinese grammar patterns in the classroom while navigating Diagram A, where the user reads specialized documentation, prompts language models, writes code and tests these expressions.

Bibliography

- Artificial Analysis. (n.d.). *Gemini 3 Flash Preview (Reasoning) Intelligence, performance & price analysis*. <https://artificialanalysis.ai/models/gemini-3-flash-reasoning> (last accessed: 24.08.2026).
- Artificial Analysis. (n.d.). *Qwen3.5 397B A17B (Reasoning) Intelligence, performance & price analysis*. <https://artificialanalysis.ai/models/qwen3-5-397b-a17b> (last accessed: 24.08.2026).
- Ayaka (n.d.). *How to properly match Chinese characters with regular expression*. <https://ayaka.shn.hk/hanregex/> (last accessed: 24.08.2026).
- Carter, A. – Absalom, M. 2023. ‘Giving Students the Tools: Looking at Teaching and Learning using Corpora’, *The EuroCALL Review*, 30(1): 52–62. <https://doi.org/10.4995/eurocall.2023.14950> (last accessed: 24.08.2026).
- Chen, H. 2012. *Using language corpora to enhance grammatical proficiency in Chinese* (Publication No. 3342). [Master’s thesis, Brigham Young University]. <https://scholarsarchive.byu.edu/etd/3342> (last accessed: 24.08.2026).
- Chinese Grammar Wiki (n.d.). *The “shi... de” construction for emphasizing details*. https://resources.allsetlearning.com/chinese/grammar/The_%22shi..._de%22_construction_for_emphasizing_details (last accessed: 24.08.2026).
- Chinese Grammar Wiki (n.d.). *The “shi... de” construction for indicating purpose*. https://resources.allsetlearning.com/chinese/grammar/The_%22shi..._de%22_construction_for_indicating_purpose (last accessed: 24.08.2026).
- Chinese Grammar Wiki (n.d.). *The “shi... de” patterns: an overview*. https://resources.allsetlearning.com/chinese/grammar/The_%22shi..._de%22_patterns_an_overview (last accessed: 24.08.2026).
- Dib, F. (n.d.). *Regex101*. Regular Expressions 101. <https://regex101.com/> (last accessed: 24.08.2026).
- Fang, C. 2017. *A prototype text analyzer for Mandarin Chinese TTS system*. [Master’s thesis in language technology, Uppsala University]. <https://www.diva-portal.org/smash/get/diva2:1111826/FULLTEXT01.pdf> (last accessed: 24.08.2026).

- Fu, A. 2021. *Match Chinese characters*. Anthony Fu. <https://antfu.me/posts/match-chinese-characters> (last accessed: 24.08.2026).
- Gajdoš, Luboš. 2017. ‘Chinese legal texts – Quantitative Description’, *Acta Linguistica Asiatica*, 7(1): 77–87. <https://doi.org/10.4312/ala.7.1.77-87> (last accessed: 24.08.2026).
- Giles 2014. ‘HSK东西 Scripts: A site for learning Chinese characters – or, “Handling Chinese characters with Python Unicode strings is less hassle than I thought it would be”’, *PythonAnywhere*. <https://blog.pythonanywhere.com/79/> (last accessed: 24.08.2026).
- Google 2026. *Gemini* [Large language model]. <https://gemini.google.com> (last accessed: 24.08.2026).
- Goyvaerts, J. 2025. *Unicode Scripts*. Regular-Expressions.info. <https://www.regular-expressions.info/unicodescript.html> (last accessed: 24.08.2026).
- Ištvánová, M. 2021. ‘On the Use of Corpora in Second Language Acquisition – Chinese as an Example’, *Acta Linguistica Asiatica*, 11(2): 131–142. <https://doi.org/10.4312/ala.11.2.131-142> (last accessed: 24.08.2026).
- Kantor, I. 2022. ‘Unicode: flag “u” and class \p{...}’, *The Modern JavaScript Tutorial*. <https://javascript.info/regexp-unicode> (last accessed: 24.08.2026).
- Kuchling, A. M. (n.d.). ‘Regular Expression HOWTO’, *Python 3.14.3 documentation*. <https://docs.python.org/3.14/howto/regex.html#regex-howto> (last accessed: 24.08.2026).
- Lau, C. – Lau, M. – To, A. 2024. ‘The extraction and fine-grained classification of written Cantonese materials through linguistic feature detection’, *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)*: 24–29. <https://aclanthology.org/2024.eurali-1.4/> <https://doi.org/10.63317/3obi2uptjo3x> (last accessed: 24.08.2026).
- Python Software Foundation (n.d.). ‘re – Regular expression operations’, *Python 3.14.3 Documentation*. <https://docs.python.org/3/library/re.html#id1> (last accessed: 24.08.2026).
- Qwen Team 2026. ‘Qwen3.5: Towards native multimodal agents’, *Alibaba Cloud*. <https://qwen.ai/blog?id=qwen3.5> (last accessed: 24.08.2026).
- Torres Ramírez, A. 2012. ‘The lexical approach: collocability, fluency and implications for teaching’, *Revista De Linguas Para Fines Especificos*, 18: 237–254. <https://ojsspd.ulpge.es/ojs/index.php/LFE/article/view/44> (last accessed: 24.08.2026).
- Unicode Consortium 2024. *The Unicode® Standard Version 16.0 – Core Specification*. South San Francisco. <https://www.unicode.org/versions/Unicode16.0.0/core-spec/> (last accessed: 24.08.2026).
- Xinhuanet 2026. 2025年新华社年度照片, 体育新闻. Xinhua Agency’s photos of the year 2025: Sports News. [Author’s translation]. <https://www.news.cn/photo/20260104/08216945fbbb4b10b38ec3f6f56a3623/c.html> (last accessed: 24.08.2026).
- Xu, J. 2019. ‘The Corpus Approach to the Teaching and Learning of Chinese as an L1 and an L2 in Retrospect’, In: Lu, X. – Chen, B. (eds). *Computational and Corpus Approaches to Chinese Language Learning. Chinese Language Learning Sciences*. Springer Nature Singapore: 33–53. https://corpus.bfsu.edu.cn/style/Xu2019_Chapter_TheCorpusApproachToTheTeaching.pdf https://doi.org/10.1007/978-981-13-3570-9_3 (last accessed: 24.08.2026).
- W3Schools (n.d.). *Python RegEx*. https://www.w3schools.com/python/python_regex.asp (last accessed: 24.08.2026).
- Yuan, S. – Zhao, H. – Du, Z. – Ding, M. – Liu, X. – Cen, Y. – Zou, X. – Yang, Z. – Tang, J. 2021. ‘WuDaoCorpora: A super large-scale Chinese Corpora for pre-training language models’, *AI Open*, 2: 65–68. <https://doi.org/10.1016/j.aiopen.2021.06.001> (last accessed: 24.08.2026).
- Yuan, Y. 2025. ‘Reflections of Task-Based Language Teaching on Grammar Teaching of Teaching Chinese as a Foreign Language’, *GBP Proceedings Series*, 5: 7–19. <https://www.gbspress.com/index.php/GBPPS/article/view/326> <https://doi.org/10.71222/tgvm3515> (last accessed: 24.08.2026).

Appendix A: Text generated by *Gemini*

我上个星期换了一份新工作。我的新公司就在火车站附近，环境非常好。我是在网上看到招聘信息的，然后发了我的简历。

周三下午，我去参加了面试。经理人很好，我们聊得很开心。他是坐地铁来公司的，而我是骑自行车去的。他问我什么时候可以开始上班，我说随时都可以。

我很喜欢这个新工作，因为同事们都很热情。我的这份工作是我努力争取来的，我一定会好好表现。

Translation:

Last week I changed jobs. My new company is right near the railway station, the atmosphere is very good. I saw the job advertisement online, I submitted my CV afterwards.

On Wednesday afternoon, I attended the interview. The manager was very nice, and the conversation went very well. He came to the company by subway, whereas I went by bike. He asked me when I could start work, and I said any time.

I really like this new job, because my colleagues are very welcoming. I earned this job of mine through hard work. I will certainly perform well.

Appendix B: Text generated by *Qwen3.5-Plus*

上个周末，我和朋友一起去北京旅游了。我们是星期六早上出发的。因为家离火车站很近，所以我们是走路去车站的。到了北京，我们是坐地铁去长城的。虽然有点累，但是我们玩得很开心。这张照片是我们在长城上拍的。这次旅游是我今年最高兴的事。

Translation:

Last weekend, my friends and I travelled to Beijing. We left on Saturday morning. We walked to the railway station because it is near our place. When we arrived in Beijing, we went to the Great Wall by subway. Even though we were a bit tired, we really had a lot of fun. We took this photo on the Great Wall. This trip is the happiest thing that has happened to me this year.

Statement about the use of artificial intelligence: Artificial intelligence tools were used to refine the Chinese text and the translated text.