Grammatical error detection using HPSG grammars: Diagnosing common Mandarin Chinese grammatical errors

Luis Morgado da Costa
Palacký University Olomouc

Francis Bond
Palacký University Olomouc

Proceedings of the 29th International Conference on
Head-Driven Phrase Structure Grammar

Online (Nagoya/Tokyo)

Stefan Müller, Elodie Winckel (Editors)

2022

Frankfurt/Main: University Library

pages 157–176

Keywords: HPSG, Mal Rules, Grammatical Error Detection, Mandarin Chinese

Morgado da Costa, Luis & Francis Bond. 2022. Grammatical error detection using HPSG grammars: Diagnosing common Mandarin Chinese grammatical errors. In Stefan Müller & Elodie Winckel (eds.), *Proceedings of the 29th International Conference on Head-Driven Phrase Structure Grammar, Online (Nagoya/Tokyo)*, 157–176. Frankfurt/Main: University Library. DOI: 10.21248/hpsg.2022.9.

Abstract

Computational Grammars can be adapted to detect ungrammatical sentences, effectively transforming them into error detection (or correction) systems. In this paper we provide a theoretical account of how to adapt implemented HPSG grammars for grammatical error detection. We discuss how a single ungrammatical input can be reconstructed in multiple ways and, in turn, be used to provide specific, high-quality feedback to language learners. We then move on to exemplify this with a few of the most common error classes made by learners of Mandarin Chinese. We conclude with some notes concerning the adaptation and implementation of the methods described here in ZHONG, an open-source HPSG grammar for Mandarin Chinese.

1 Introduction

In recent years, the fields of automated Grammar Error Detection (GED) and Correction (GEC) have gained popularity. English has, no doubt, attracted the most attention. This is shown by the number of shared-tasks made available in the recent years (Dale & Kilgarriff, 2011; Dale et al., 2012; Ng et al., 2013, 2014; Daudaravicius et al., 2016; Bryant et al., 2019).

Similar efforts have started for Mandarin Chinese Grammar Error Detection (CGED) and Correction (CGEC). Most of such efforts revolve around the shared-task organized by the NLP-TEA held from 2014–2018 (Yu et al., 2014; Lee et al., 2015, 2016; Gaoqi et al., 2017; Rao et al., 2018). Rao & Lee (2018) provide an overview of all previous CGED tasks, drawing attention to the intrinsic difficulty of this task, and the long road ahead.

Constraint-based grammars are ideal for GED/GEC because they model grammaticality directly. In this paper we will first introduce the concepts of *mal-rules* (Schneider & McCoy, 1998), and show how multiple different meanings can be reconstructed from a single ungrammatical input using *mal-rules* modeled using Head Driven Phrase Structure Grammars (Pollard & Sag, 1994; Sag et al., 1999, HPSG). We will then introduce work on *mal-rules* applied to Mandarin Chinese Grammatical Error Detection based on first-hand data collected from learners of Mandarin Chinese. Finally, we will end with some notes on the actual implementation of mal-rules in ZHONG (Fan et al., 2015) – an open source Mandarin Chinese HPSG grammar, currently being transformed into a GED system.

To the best of our knowledge, even though there is previous work dealing with mal-rules in HPSG, there have been no papers attempting to discuss *mal-rules* from a more theoretical perspective – providing full examples of different ways to correct similar errors or discussing how it is possible and often important to ambiguate an ungrammatical input into multiple possible corrections. In addition, there are no previous reports of *mal-rule* enhanced HPSG grammars for Mandarin Chinese. This paper will address these gaps.

The rest of this paper is structured as follows: Section 2 introduces the concept of *mal-rules*, and provides examples of how they are implemented in HPSG. Some examples of *mal-rules* targeting common errors among learners of Mandarin Chinese are shown in Section 3, followed by a few notes on their implementation in a working parser in Section 4. Finally, we conclude.

2 Mal-Rules in HPSG

In constraint-based linguistic language models, such as HPSG grammars, robustness is an early and ever present concern. When compared with shallow parsing methods, the explicit nature of constraint-based models makes them less robust. Forms of input that were not explicitly accounted for in grammar are simply rejected. This is by design: constraint-based models make an explicit grammaticality judgment when they parse or reject an input – which is usually not true for statistical-based parsers. This rigidity (i.e., the lack of robustness for ill-formed or unknown input) that could be considered a problem for some NLP applications, becomes a valuable trait when we need to deal with problems concerning grammaticality.

Mal-rules (Schneider & McCoy, 1998) extend computational grammars in order to analyze ungrammatical phenomena. *Mal-rules* can be used to identify and correct specific grammatical errors, and to trigger corrective feedback messages to help language learners. Depending on the type of parser they are implemented in, *mal-rules* can be designed to reconstruct the semantics of ungrammatical sentences, and can be selectively available for parsing but not for generation (Bender et al., 2004). Consider (1), below:

(1) * *This students sleep.*

Any English grammar should reject (1) as a proper sentence. This is enough to identify something is wrong with the sentence. However, if the intention were to diagnose what is wrong with it, then the problem gains a new layer of complexity. We would argue that, without context, it would be impossible to choose a single correction to (1). Two possible corrections are shown in (2) and (3), but a few more most certainly exist.

(2) *These students sleep.*

(3) *This student sleeps.*

In order to correct (1), we first need to guess what was the intended meaning behind the ungrammatical sentence. And to make this decision, need to be able to generate a set of candidate intended meanings.

Mal-rules are able to do exactly this. *Mal-rules* reconstruct ungrammatical input in meaningful ways – enabling both error detection and correction. There are, potentially, two sources of ungrammaticality in (1): the first is concerned with the problem of agreement between the determiner *this* and the noun *students*; and the second is concerned with the problem of subject-verb agreement, but is dependent on how the first is corrected. Different sets of mal-rules are needed to allow reconstructing the meaning of (2) and of (3). This will be discussed in great detail, step-by-step, in Section 2.1.

Adding *mal-rules* to a grammar is sufficient not only to detect multiple possible corrections of a sentence like (1), but would also be sufficient to explain how the sentence needed to change (i.e. in linguistic terms). This makes *mal-rules* specially interesting in the field of education. *Mal-rules* can be used to trigger corrective feedback messages to help language learners understand why a sentence is ungrammatical.

2.1 Mal-Rules in HPSG

Using *Mal-Rules* in HPSG grammars has a long history. There have been efforts for English (Bender et al., 2004; Flickinger & Yu, 2013), Norwegian (Hellan et al., 2013), German (Heift, 1998), Spanish (Costa et al., 2006) and French (Hagen, 1994). From these, only English and Norwegian are still in active development.

As discussed in Bender et al. (2004), the implementation of *mal-rules* in HPSG grammars can be done through three major classes of linguistic objects: syntactic rules, lexical rules, and lexical items. And even though each method has some degree of specificity, making them useful in detecting different kinds of errors, there is also overlap in their explanative power (i.e. similar errors could be captured in more than one way). These degrees of specificity, and how they interact, have not been fully discussed prior to this paper. In this paper, we will explore these different levels of specificity, as well as how multiple mal-rules can be used together to predict multiple plausible corrections for a single ungrammatical sentence.

2.1.1 Syntactic Mal-Rules in HPSG

The use of syntactic *mal-rules* in HPSG is both powerful and flexible. Consider the ungrammatical noun phrase (NP) *this students*. Under normal circumstances, this phrase is not grammatical. In HPSG, this is ensured by the Specifier Head Agreement Constraint (SHAC) present in the Head-Specifier Rule (4), as proposed in Sag et al. (1999). According to the SHAC, phrases taking a specifier are required to unify their agreement features with those of their specifier – this is shown by $\boxed{2}$ in (4). The specifier of a NP is its determiner, so this is what establishes the required agreement between the noun and the determiner.

(4)
$$\begin{bmatrix} \textit{head-specifier-rule} \\ \text{SYN} \begin{bmatrix} \text{VAL} \begin{bmatrix} \text{SPR} \; \langle\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \rightarrow \boxed{1} \;\; \mathbf{H} \begin{bmatrix} \text{SYN} \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{AGR} \; \boxed{2} \end{bmatrix} \\ \text{VAL} & \begin{bmatrix} \text{SPR} & \left\langle \boxed{1} \begin{bmatrix} \text{AGR} \; \boxed{2} \end{bmatrix} \right\rangle \\ \text{COMPS} & \langle\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

(5)
$$\begin{bmatrix} \textit{mal-head-specifier-rule} \\ \textsc{syn} \begin{bmatrix} \textsc{val} \begin{bmatrix} \textsc{spr} \; \langle\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \longrightarrow \boxed{1} \; \mathbf{H} \begin{bmatrix} \textsc{syn} \begin{bmatrix} \textsc{head} \begin{bmatrix} \textsc{agr} \; X \end{bmatrix} \\ \textsc{val} \begin{bmatrix} \textsc{spr} & \left\langle \boxed{1} \begin{bmatrix} \textsc{agr} \; Y \end{bmatrix} \right\rangle \\ \textsc{comps} & \langle\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

One possible way to build the NP *this students* would be to relax the constraint imposed by the SHAC. Creating a new rule where this constraint is not enforced would qualify it as a *mal-rule* – since such rule would allow ungrammatical phrases to be licensed by the grammar. This *mal-rule* can be found in (5). Note that where $\boxed{2}$ in (4) made sure both the head-daughter (i.e. noun) and its specifier (i.e. determiner) agreed, in (5) this is not true. (5) would allow the grammar to build *this students* as a valid NP.

The Head-Specifier Rule as described in Sag et al. (1999), is used to build many kinds of phrases, including full sentences (i.e. linking NP subjects and their VP predicates). This means that the *mal-rule* shown in (5) would also license sentences such as '*Students sleeps.*' or ' *I sleeps.*' – where the subject does not agree with the main verb. This accounts for the flexible power of syntactic *mal-rules*, but also shows that even though (5) could be used to detect ungrammatical sentences, it has a fairly low precision with regard to what kind of error it licences – i.e., an unspecified problem in agreement.

2.2 Lexical Mal-Rules in HPSG

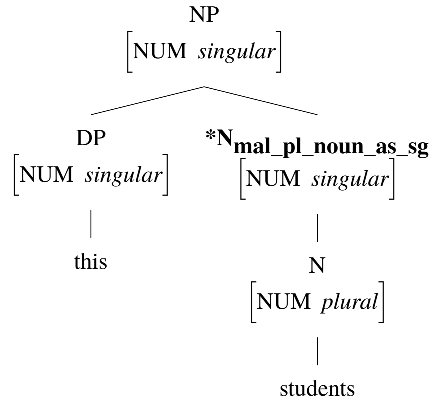
HPSG grammars often have a rich hierarchy of lexical rules. An alternative way to build the NP *this students* would be through lexical *mal-rules*. This could be done with a lexical rule that allows, for example, a plural noun to be used as a singular noun. An example of this rule is shown in (6).

(6)
$$\begin{bmatrix} \textit{mal\_pl\_noun\_as\_sg\_lrule} \\ \textsc{input} \left\langle \boxed{1}, \begin{bmatrix} \textit{word} \\ \textsc{syn} \begin{bmatrix} \textsc{head} \begin{bmatrix} \textit{noun} \\ \textsc{agr} \begin{bmatrix} \textsc{num} \; \boldsymbol{pl} \end{bmatrix} \end{bmatrix} \\ \textsc{val} \begin{bmatrix} \textsc{spr} & \left\langle \boxed{2}\,\text{DP} \right\rangle \\ \textsc{comps} & \left\langle (\boxed{3} \ldots \boxed{n}) \right\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \right\rangle \\ \textsc{output} \left\langle \boxed{1}, \begin{bmatrix} \textbf{\textit{mal\_pl\_noun\_as\_sg}} \\ \textsc{syn} \begin{bmatrix} \textsc{head} \begin{bmatrix} \textit{noun} \\ \textsc{agr} \begin{bmatrix} \textsc{num} \; \boldsymbol{sg} \end{bmatrix} \end{bmatrix} \\ \textsc{val} \begin{bmatrix} \textsc{spr} & \left\langle \boxed{2} \right\rangle \\ \textsc{comps} & \left\langle (\boxed{3} \ldots \boxed{n}) \right\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}$$

This lexical *mal-rule* can only be applied to plural nouns, and produces a copy of the input noun, changing only the number feature (i.e. from *plural* to *singular*).

Using the lexical *mal-rule* shown in (6), our English grammar would be able to build the NP *this students* by first changing the number feature of the word *students* to singular, and then using the normal rule that joins nouns and determiners – as is shown in (7).

(7)



2.3 Mal Lexical Entries in HPSG

Finally, a third way to build the NP *this students* is to use a *mal* lexical entry. This is similar, in spirit, to lexical *mal-rules*, but instead of generalizing across word classes, it provides an alternative *mal* lexical entry for specific words that are known to be source of errors. One such example would be the correct and *mal* lexical entries for *this*, shown as (8) and (9), respectively.

Entries (8) and (9) differ only slightly. The first of these differences is the value for the number feature. For the *mal* lexical entry, shown in (9), it is set to *plural*. Additionally, the semantic relation it introduces is similar to what would be expected of an entry for the determiner *these*. In short, (9) behaves like the word *these* but carries the form *this*. This *mal* lexical entry would license the NP *this students* following the tree shown in (10).

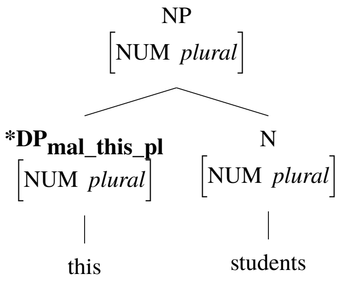
(8)

$$\left\langle \textit{this}, \begin{bmatrix} \textit{determiner} & \\ \text{SYN} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{DET} & \\ \text{AGR}|\text{NUM} & \boldsymbol{\textit{singular}} \end{bmatrix} \end{bmatrix} \\ \text{SEM} & \begin{bmatrix} \text{INDEX} & i \\ \text{RESTR} & \left\langle \begin{bmatrix} \text{RELN} & \boldsymbol{\textit{this}} \\ \text{BV} & i \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \right\rangle$$

(9)

$$\left\langle \mathit{this}, \begin{bmatrix} \textit{mal\_this\_pl} & \\ \textsc{syn} & \begin{bmatrix} \textsc{head} & \begin{bmatrix} \textsc{det} & \\ \textsc{agr}|\textsc{num} & \boldsymbol{\mathit{plural}} \end{bmatrix} \end{bmatrix} \\ \textsc{sem} & \begin{bmatrix} \textsc{index} & i \\ \textsc{restr} & \left\langle \begin{bmatrix} \textsc{reln} & \boldsymbol{\mathit{these}} \\ \textsc{bv} & i \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \right\rangle$$

(10)



2.4 Combining Approaches

Although they might seem to provide similar results, the trees shown in (7) and (10) differ in one key aspect – the value for the number feature of the produced NP. In HPSG, the syntactic number of a phrase is determined by the head of that phrase – in a NP, this would be the noun. This is a good example of how *mal-rules* can be used to reconstruct different possible meanings from a single ungrammatical input.

To be able to evaluate the full reach of meaning reconstruction, let us consider a variation of the *mal-rule* introduced in (5). The rule shown in (11) changes the general Head-Specifier rule into a Head-Subject rule (by selecting verb has the head type for the daughter), but agreement is not enforced. In short, (11) selectively allows sentences where the subject and the main verb of a sentence do not agree.

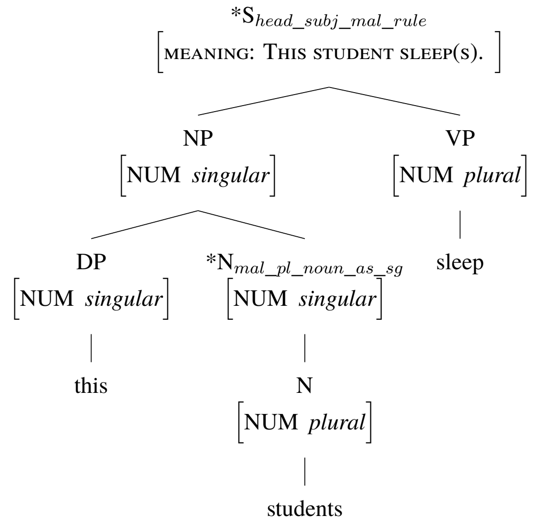
(11)

$$\begin{bmatrix} \textit{head-subj-mal-rule} \\ \textsc{syn} \begin{bmatrix} \textsc{val} \begin{bmatrix} \textsc{spr} \ \langle\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \longrightarrow \boxed{1}\ \mathbf{H} \begin{bmatrix} \textsc{syn} \begin{bmatrix} \textsc{head} & \begin{bmatrix} \textbf{verb} & \\ \textsc{agr} & \boldsymbol{X} \end{bmatrix} \\ \textsc{val} & \begin{bmatrix} \textsc{spr} & \left\langle \boxed{1} \begin{bmatrix} \textsc{agr} & \boldsymbol{Y} \end{bmatrix} \right\rangle \\ \textsc{comps} & \langle\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

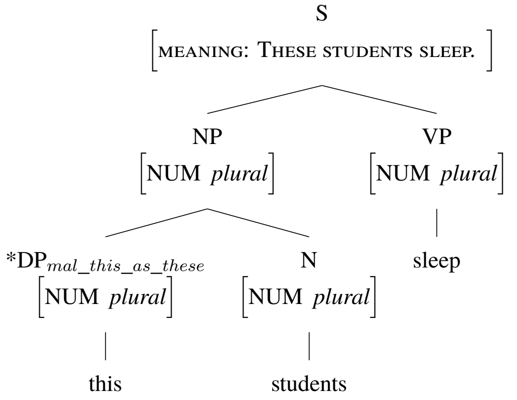
Using only the three *mal-rules* shown in (6), (9) and (11), we can get the the two reconstructions discussed for the ungrammatical sentence in (1). These reconstructions were introduced as (2) and (3), above, and are shown in (12) and (13) in the form of syntactic trees. The main difference between these two trees is the reconstructed meaning. In (12), the grammar reconstructed a sentence where only a single student is sleeping. And in (13), the reconstructed meaning assumes more than one student is sleeping.

For systems where the goal is simply grammatical error detection (i.e. without correction), traversing the parsing tree and looking for nodes where *mal-rules* were used is enough to diagnose the ways in which a sentence is ungrammatical. However, if a grammar has generation capabilities, reconstructing different meanings also allows the generation of the corrected counterparts. For this reason, most implemented HPSG grammars can be used to produce fully capable error detection and correction systems.

(12)



(13)



3 Detection of Common Mandarin Chinese Errors

In this section we focus on the design of rules that detect common errors among learners of Mandarin Chinese as a second language.

3.1 A New Mandarin Learner Corpus

GED is usually done against labeled learner data, known as Learner Corpora. Before one can hope to design GED or GEC systems, it is first necessary to know what errors learners of a given language actually make (Granger, 2003). These kind of corpora are also useful to measure the performance of error detection or correction systems (Schulze, 2008). And when semantically annotated, Learner Corpora are useful resources to help predict the intended meaning behind students' input (Hellan et al., 2013).

Since building learner corpora is extremely time consuming, there are few freely available learner corpora. There are some resources for English, but there are no freely available learner corpora made from learners of Mandarin Chinese that focus on written language. The Jinan Learner Corpus (Wang et al., 2015) seems to no longer be accessible online, and the iCALL Corpus (Chen et al., 2015) is a speech corpus mostly concerned with errors in pronunciation.

The TOCFL Learner Corpus (Lee et al., 2018) and the Lang-8 corpus (Mizumoto et al., 2011) are the only known learner corpora with a focus on written Mandarin Chinese. Unfortunately, both of them are released under restrictive non-comercial non-redistribution licenses. In addition, both corpora have been created with specific tasks in mind. The TOCFL, for example, includes only four very broad error types: 'redundant words', 'missing words', 'word selection errors', and 'word ordering errors'. And the Lang-8 corpus is an automatically collected corpus providing only pairs of sentences and their respective corrections. While both data sources would be extremely valuable if open, the restrictive licenses constrain their use.

Therefore we decided to collect our own data. This data comes from Mandarin Chinese learners at Nanyang Technological University, Singapore. We collected 5,513 sentences from student exams, which, after removing duplicates, corresponded to 2,300 unique sentences. After a thorough annotation process, we identified 544 errors divided among 490 problematic sentences (i.e., around 21.3% of the sentences had at least one error tag assigned to them). A summary of results is shown in Table 1.

A full description of this corpus is beyond the scope of this paper. Nevertheless, we will be using four of the most frequent classes of errors to further explore *mal-rules* and see how these can be used to catch (and correct) common errors made by Mandarin Chinese learners. Each of the four classes of errors will be discussed separately.

3.2 Question Particle Redundancy (ID-1)

The most frequent grammatical error in our corpus is the misuse of the question particle 吗 *ma*. The proper use of 吗 *ma* transforms propositions into polar (i.e. yes-no) questions. This particle often confuses learners into assuming that it is similar to a question mark (i.e. simply marking the existence of a question: which is the behaviour of the the Japanese question marker か *ka*). However, as can be seen in (15) and (17), this is not the case in Chinese. In sentences where other interrogative words or structures are used, such as (14), the question particle 吗 *ma* should not be added. In (16), the usage of a special syntactic construction (Verb-NOT-Verb) already implies a polar question. It is then ungrammatical to redundantly add *ma*,

ID	Description	Total
1	吗 (*ma*, question particle) redundancy	26
2	Usage of 和 (*hé*, and) vs. 也 (*yě*, also)	25
3	Position of adverbial clauses	25
4	Usage of 是 (*shì*, to be) with adjectival predicates	23
5	Usage of 中国 (*zhōngguó*, China) vs. 中文 (*zhōngwén*, Chinese language)	18
6	Position of 也 (*yě*, also)	14
7	Usage of 有点儿 (*yǒudiǎnr*, somewhat) vs. 一点儿 (*yīdiǎnr*, a bit)	14
8	Bare adjectival predicates	9
9	Usage of 是... 的 (*shì...de*, focus cleft) constructions	8
10	Usage of 不 (*bù*, no) with specified adjectival predicates	6
11	Incorrect measure word	6
12	Missing measure word	5
13	Attributive 多 (*duō*, many) and 少 (*shǎo*, few) without degree specifiers	5
14	Usage of 二 (*èr*, two) vs. 两 (*liǎng*, two)	4
15	Usage of 不 (*bù*, no) vs. 没有 (*méiyǒu*, no)	3
16	Syntactic order of 也 (*yě*, also), 都 (*dōu*, all), 不 (*bù*, no)	3
17	Syntactic order of nominal 的 (*de*, possessive marker) modification	2
18	Other Errors	348
	Total	544
	Sentences w/errors	490

Table 1: Distribution of Mandarin Chinese Error Tags by Frequency

as seen in (17). More generally, *ma* should never be used in sentences that are, by themselves, already questions.

(14) 你 要 什么 ？
nǐ yào shénme ?
2SG want Q.what ?
'What do you want?'

(15) *你 要 什么 吗 ？
nǐ yào shénme ma ?
2SG want Q.what Q.polar ?
(intended) 'What do you want?'

(16) 你 有没有 中文 书 ？
nǐ yǒu-méi-yǒu zhōngwén shū ?
2SG have-not-have Chinese book ?
'Do you have a Chinese textbook?'

(17) *你 有没有 中文 书 吗 ？
nǐ yǒu-méi-yǒu zhōngwén shū ma ?
2SG have-not-have Chinese book Q.polar ?
(intended) 'Do you have a Chinese textbook?'

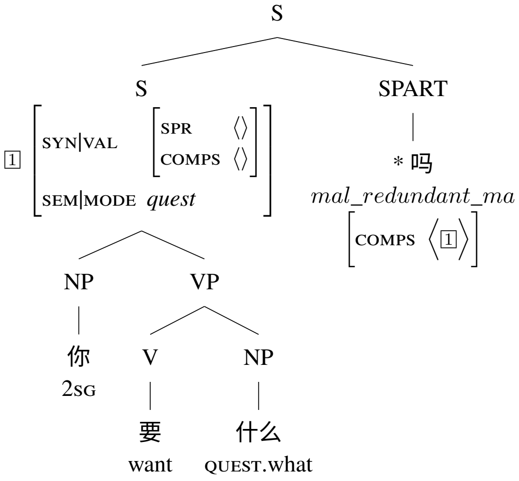
We deal with this error by adding to the grammar an extra *mal* lexical entry for 吗 *ma*, shown in (18). This *mal* lexical entry – which is identified as a mal-rule by the type's name – provides a second entry for *ma* as a sentence final particle (i.e. *spart*). This sentence particle expects a single VP complement, that is defined to have empty values for SPR (specifier) and COMPS (complements). This guarantees that it modifies only complete sentences. It is also marked as $\begin{bmatrix}\textsc{posthead} & +\end{bmatrix}$, restricting its use to post-head position (i.e., a sentence final particle). Finally, and most importantly, its complement has a SEM|MODE value equal to *quest* – meaning that the sentence it selects must already be identified as a question.

In other words, the lexical entry for 吗 *ma* shown as (18) attaches only to full sentences that are already questions. Using this *mal* lexical entry in an existing grammar of Mandarin Chinese would allow it to parse ungrammatical sentences like the one shown in (19). All similar ungrammatical sentences, where a well formed question is followed by a redundant *ma*, can be detected using this same *mal* lexical entry.

(18)

$$\left\langle \text{吗}, \begin{bmatrix} \textbf{\textit{mal\_redundant\_ma}} & \\ \textsc{syn} & \begin{bmatrix} \textsc{head} & \textbf{\textit{spart}} \\ \textsc{posthead} & + \\ \textsc{val} & \begin{bmatrix} \textsc{comps} & \left\langle \textsc{vp}\begin{bmatrix} \textsc{syn}|\textsc{val} & \begin{bmatrix}\textsc{spr} & \langle\rangle \\ \textsc{comps} & \langle\rangle\end{bmatrix} \\ \textsc{sem}|\textsc{mode} & \textbf{\textit{quest}}\end{bmatrix}\right\rangle \end{bmatrix}\end{bmatrix}\end{bmatrix}\right\rangle$$

(19)



3.3 Use of Copula with Adjectival Predicates (ID-4)

We now look at the use of copula 是 *shì* with adjectival predicates. Examples (20) through (23) exemplify the simplest minimal pairs illustrating the usage of Mandarin adjectival predication. Even though these restrictions may differ in informal speech or contrastive constructions, in a prescriptive environment, adjectival predicates need to be modified by an adverbial phrase. In addition, adjectival predicates should not use the copula verb (regardless of having been modified, or not, by an adverbial phrase). Because of this, in a beginner's classroom, examples (21) through (23) are problematic.

(20) 她 很 美 。
tā hěn měi .
3SG.FEM very beautiful .
'She is beautiful.' (lit. 'She is very beautiful.')

(21) * 她 美 。
tā měi .
3SG.FEM beautiful .
(intended) 'She is beautiful.'

(22) * 她 是 美 。
tā shì měi .
3SG.FEM COP.be beautiful .
(intended) 'She is beautiful.'

(23) * 她 是 很 美 。
tā shì hěn měi .
3SG.FEM COP.be very beautiful .
(intended) 'She is very beautiful.'

We will focus on detecting the use of 是 *shì* with adjectival predicates. In the interest of space, however, we will not delve in the related error concerning (21), dealing with the further requirement that adjectival predicates must be generally preceded by an adverbial intensifier. These are two different errors, and we will only discuss the first.

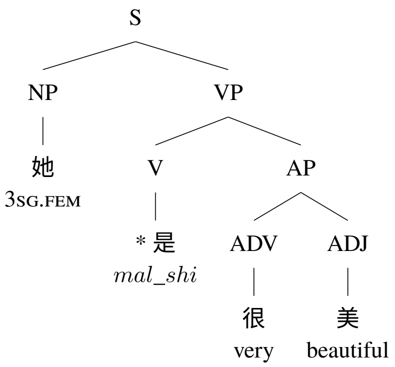
We address this error by creating a *mal* lexical entry for a dummy copula 是 *shì* that behaves like a transitive verb, but that selects only adjective phrases (AP) complements – shown as (24). This entry adds nothing to the semantics, just linking its own subject with the subject of the adjective.

Using this *mal* lexical entry, our grammar would be able to license sentences such as the one shown in (22)/(25), giving it the same semantics as the sentence without *shì*. Once again, this analysis generalises for other sentences where adjectival predicates are preceded by *shì*.

(24)

$$\left\langle \text{是}, \begin{bmatrix} \textit{mal\_shi} & \\ \text{SYN} & \begin{bmatrix} \text{HEAD} & \textit{verb} \\ \text{VAL} & \begin{bmatrix} \text{SPR} & \langle \text{NP} \rangle \\ \text{COMP} & \langle \text{AP} \rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \right\rangle$$

(25)



3.4 Bare Nominal Predicates (ID-18)

The third error class we will discuss concerns bare nominal predicates. In Mandarin Chinese, although adjectival predication happens without the use of a copula verb, nominal predication requires the use of a copular verb (是, *shì*) – rendering sentences like (27) ungrammatical.

(26) 我 是 大学生 。
wǒ shì dàxuéshēng .
1SG COP.be university.student .
'I am a university student.'

(27) * 我 大学生 。
wǒ dàxuéshēng .
1SG university.student .
(intended) 'I am a university student.'

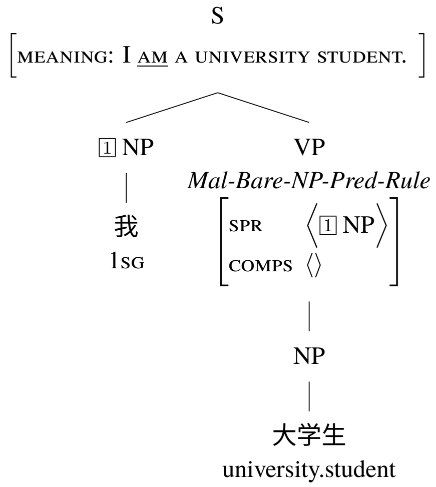
The contrastive behavior of adjectival predication is likely that the source of this error. Learners generalize this behavior and assume that nominal predication behaves similarly – and thus produce ungrammatical sentences.

We currently address this problem through the use of a *mal* 'pumping' rule, shown in (28). This pumping rule transforms any fully specified NP into something akin to an intransitive verb – i.e., it behaves like a VP in the sense that it expects an NP as specifier (i.e., a subject).

Making use of (28) allows a grammar to parse sentence (27) and other similar sentences. The tree for this analysis is shown in (29). In it, we can see that 大学生 (*dà xué shēng* "university student") is pumped from an NP into a VP, capable of taking 我 (*wǒ*, "I") as its subject. In order to reconstruct the meaning, this rule also adds a copula predicate.

(28)
$$\begin{bmatrix} \textit{mal-bare-nppred-rule} \\ \text{SYN} \begin{bmatrix} \text{HEAD } \textit{verb} \\ \text{VAL} \begin{bmatrix} \text{SPR} & \langle \text{NP} \rangle \\ \text{COMPS} & \langle\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix} \longrightarrow \begin{bmatrix} \text{SYN} \begin{bmatrix} \text{HEAD } \textit{noun} \\ \text{VAL} \begin{bmatrix} \text{SPR} & \langle\rangle \\ \text{COMPS} & \langle\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

(29)



3.5 Non-Prototypical Complements (ID-5)

Our final set of examples are drawn from a lexical conflation between *China* (中国, *zhōngguó*) and *Chinese Language* (中文, *zhōngwén*). Although sentences such as (32) are not strictly ungrammatical, as shown by (31), learners often use (32) when they intend to say *I speak Chinese*.

(30) 我 说 中文 。
wǒ shuō zhōngwén .
1SG speak Chinese.lang .
'I speak Chinese.'

(31) 我 说 中国 。
wǒ shuō zhōngguó .
1SG speak China .
'I say China.'

(32) # 我 说 中国 。
wǒ shuō zhōngguó .
1SG speak China .
(intended: 'I speak Chinese.')

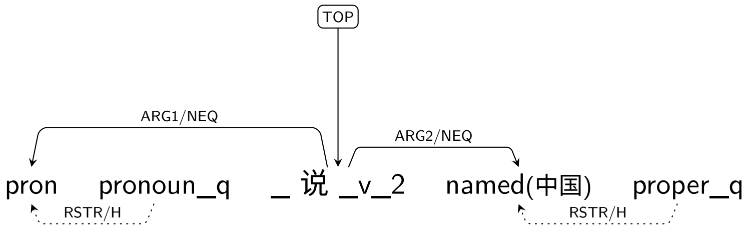
More generally, this class of errors addresses the use of unlikely (i.e. non-prototypical) complements. These are not stricly syntactic errors: the sentence is grammatical, but the meaning is unexpected.

According to our learner corpus, the conflation between *China* (中国, *zhōng guó*) and *Chinese Language* (中文, *zhōng wén*) happens most frequently as the complement of the verb 说 (*shuō*, "to speak, to say"). Learners often want to express the meaning of (30), but use 中国 (*zhōng guó*, "China") instead of 中文 (*zhōng wén*, "Chinese Language").

Even though it would be possible to detect the use of non-prototypical complements using *mal* lexical entries, it is much easier to dealt with it at the semantic level.

One advantage of working with grammars able to produce semantics is the fact that the semantic output can also be used to identify certain kinds of problems in language usage. This is especially relevant for non-syntactic issues such as non-prototypical complements. Lets consider the simplified semantic representation for (32) "I say China" (intended: "I speak Chinese") as a Dependency MRS (Copestake, 2009) shown as (33).

(33)



This semantic representation shows that *China* (中国, *zhōng guó*) is the ARG2 of 说 (shuō, *to speak, to say*) – i.e. *what is said*. So instead of creating a special *mal* lexical entry for 说 (shuō) – which would be a possible solution, a simple semantic check can be done to see if (中国, *zhōng guó*) is used as the ARG2 of the verb 说 (shuō). Given the deep semantic analysis performed by these kind of grammars, the semantic arguments are also easily detectable in the presence of discontinuous arguments (e.g. topicalization, etc.) – which can be a problem when using shallow text based methods.

This kind of semantic analysis is also our preferred method to deal with similar problems, such as the use of inappropriate classifiers in NP quantification.

4 Implementation in a Grammar

The errors described above, as well as many others omitted in the interest of space, have been implemented in ZHONG – a Mandarin Chinese HPSG grammar (Fan et al., 2015). ZHONG is a medium-sized HPSG grammar able to produce Minimal Recursion Semantics (Copestake et al., 2005; Copestake, 2007, MRS). Both ZHONG and the mal-rule extensions discussed in this paper are fully open-source.[1]

ZHONG currently contains more than 60 *mal rules* (including lexical entries) – which covers about half of the types of errors we were able to find in our learner corpus. As such, describing each individual rule would not be possible nor desirable, as many *mal rules* share design principles.

The process of transferring the *mal-rules* described in this paper, which are fairly theoretical, into an implemented grammar such as ZHONG is not always simple. Each individual grammar has its own idiosyncrasies, and the final form

[1] https://github.com/delph-in/zhong

of some of the *mal-rules* described above had to be adapted to match ZHONG's type hierarchy.

Our current method is to implement *mal-rules* in a graded fashion – i.e., starting with errors made by beginners before moving on to higher levels of proficiency. This mainly helps the applicability of our efforts (i.e., the grammar can immediately be used in learning systems targeting low proficiency learners). It is also important to note that *mal-rules* are not constrained by the current level of complexity of a grammar. A well designed *mal-rule* will always accompany the complexity of a grammar as it grows. For example, a subject-verb disagreement error will always be relevant, regardless of the complexity of the subject or of the verb phrase in question.

More importantly, the design of our *mal-rules* is targeted specifically at a level of granularity that would be adequate to use for student feedback.

4.1 Learner Treebanks

Following what was discussed above, the number of generated corrections for an ungrammatical sentence is often greater than what we would expect. Despite employing logical constraint-based approaches to generate parses, normal/prescriptive HPSG grammars often make use of treebanks to produce parse ranking models and order the available parses by likelihood. This is usually seen as a necessary step for implemented grammars, since without it a grammar's analysis is usually quite useless.

This is also an issue when we use *mal-rules*. With the addition of *mal-rules*, grammars become increasingly more ambiguous. This is not necessarily bad in the sense that this ambiguity is reflected on the ability to predict multiple different corrections for the same ungrammatical input, but it becomes a problem when parsing grammatical sentences, because *mal-rules* will be competing with descriptive rules.

Using a parse-ranking model that has been trained in the absence of *mal-rules* will inevitably produce cases where very unlikely parses are ranked higher than very likely errors. This is why it is important to invest early in treebanks that contain learner data – which we have named Learner Treebanks (Morgado da Costa et al., 2022).

Morgado da Costa et al. (2022) provide a full account of the design and impact of using Learner Treebanks alongside *mal-rule* enhanced grammars. These treebanks enable the creation of *mal-rule* enhanced parse ranking models Toutanova et al. (2005), which help rank multiple corrections in order of likelihood, while avoiding having to resort to creative ways to be able to perform well (e.g., the use of very restrictive vocabulary, the use of other methods to filter the results, or the of sub-optimal heuristics to select the best parse – e.g., select the parse with fewest number of mal-rules). For these reasons, we have also stared working on a new Learner Treebank for Mandarin Chinese.

5 Conclusion

Scholars are trying to elaborate on the role of formal linguistics in the wider field of Computational Linguistics[2] (currently dominated by statistical/neural-based methods). This paper discusses an excellent example of the continued relevance of computational grammars. Working with computational grammars to perform error detection alongside language teachers has also proved to be productive in managing their expectations over the balance between quality and performance – something 'black-box' statistical systems have a hard time doing.

This paper describes, in some detail, how to perform grammatical error detection using HPSG grammars. It shows that *mal-rules* in HPSG enable the prediction of multiple corrected forms for a single ungrammatical sentence – which is arguably an extremely important feature in language education contexts. Most of the current work in GED and GEC uses optimization-based statistical models that are designed to provide a single 'best' result. The use of *mal-rules* can free systems from this restriction, and open new ways of looking at how the problems of Grammar Error Detection and Correction could be redefined for the future.

Finally, this paper also makes contributions to the specific field of Mandarin Chinese Grammatical Error Detection. We analyze and design *mal-rules* to detect some of the most common errors made by second language learners of Mandarin Chinese, based on empirical data collected for our new learner corpus for Mandarin Chinese. More than 60 *mal-rules* have been implemented in ZHONG. The work that will be presented in this paper is being conducted as part of a larger project looking into building a Computer Assisted Language Learning system to help learners of Mandarin Chinese improve their language proficiency. In the near future, we will integrate this grammar in an online language tutoring system, where learners can test their knowledge of Mandarin Chinese and where each *mal-rule* (and semantic check) will be linked to corrective feedback messages describing the errors and how best to correct them.

6 Acknowledgements

This research received support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-IF-2020 CHILL – No.101028782, as well as the European Regional Development Fund Project "Sinophone Borderlands –Interaction at the Edges", CZ.02.1.01/0.0/0.0/16_019/0000791.

References

Bender, Emily M, Dan Flickinger, Stephan Oepen, Annemarie Walsh & Timothy Baldwin. 2004. Arboretum: Using a precision grammar for grammar checking

[2] See, for example: https://gdr-lift.loria.fr/bridges-and-gaps-workshop/

in CALL. In *Proceedings of the InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, Italy.

Bryant, Christopher, Mariano Felice, Øistein E Andersen & Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 52–75.

Chen, Nancy F, Rong Tong, Darren Wee, Peixuan Lee, Bin Ma & Haizhou Li. 2015. iCALL corpus: Mandarin Chinese spoken by non-native speakers of European descent. In *Sixteenth Annual Conference of the International Speech Communication Association*, 324–238.

Copestake, Ann. 2007. Semantic composition with (robust) minimal recursion semantics. In *Proceedings of the Workshop on Deep Linguistic Processing*, 73–80. Association for Computational Linguistics.

Copestake, Ann. 2009. Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 1–9. Athens.

Copestake, Ann, Dan Flickinger, Carl Pollard & Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation* 3(2–3). 281–332.

Costa, Flávio M, José Carlos L Ralha & Célia G Ralha. 2006. Aprendizagem de língua assistida por computador: Uma abordagem baseada em HPSG. *Revista Brasileira de Informática na Educação* 14(1).

Morgado da Costa, Luis, Francis Bond & Roger V. P. Winder. 2022. The Tembusu Treebank: An English learner treebank. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, Marseille, France: European Language Resources Association (ELRA).

Dale, Robert, Ilya Anisimoff & George Narroway. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 54–62. Association for Computational Linguistics.

Dale, Robert & Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, 242–249. Association for Computational Linguistics.

Daudaravicius, Vidas, Rafael E Banchs, Elena Volodina & Courtney Napoles. 2016. A report on the automatic evaluation of scientific writing shared task. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 53–62.

Fan, Zhenzhen, Sanghoun Song & Francis Bond. 2015. An HPSG-based shared-grammar for the Chinese languages: ZHONG [|]. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) Workshop*, 17–24.

Flickinger, Dan & Jiye Yu. 2013. Toward more precision in correction of grammatical errors. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, CoNLL 2013*, 68–73.

Gaoqi, RAO, Baolin Zhang, XUN Endong & Lung-Hao Lee. 2017. IJCNLP-2017 task 1: Chinese grammatical error diagnosis. In *Proceedings of the IJCNLP 2017, Shared Tasks*, 1–8.

Granger, Sylviane. 2003. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly* 37(3). 538–546.

Hagen, L Kirk. 1994. Unification-based parsing applications for intelligent foreign language tutoring systems. *Calico Journal* 12(2&3). 5–31.

Heift, Trude. 1998. An interactive intelligent language tutor over the internet. In *Proceedings of ED-MEDIA, ED-TELECOM 98, World Conference on Education Multimedia and Educational Telecommunications*, vol. 2, 508–512.

Hellan, Lars, Tore Bruland, Elias Aamot & Mads H Sandoy. 2013. A Grammar Sparrer for Norwegian. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, 435–439.

Lee, Lung-Hao, Gaoqi RAO, Liang-Chih Yu, Endong XUN, Baolin Zhang & Li-Ping Chang. 2016. Overview of NLP-TEA 2016 shared task for Chinese grammatical error diagnosis. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications*, Osaka, Japan: The COLING 2016 Organizing Committee.

Lee, Lung-Hao, Yuen-Hsien Tseng & Liping Chang. 2018. Building a TOCFL learner corpus for Chinese grammatical error diagnosis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2298–2304.

Lee, Lung-Hao, Liang-Chih Yu & Li-Ping Chang. 2015. Overview of the NLP-TEA 2015 shared task for Chinese grammatical error diagnosis. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, Beijing, China: Association for Computational Linguistics.

Mizumoto, Tomoya, Mamoru Komachi, Masaaki Nagata & Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, 147–155.

Ng, Hwee Tou, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto & Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *CoNLL Shared Task*, 1–14.

Ng, Hwee Tou, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto & Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, Sofia, Bulgaria: Association for Computational Linguistics.

Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press.

Rao, Gaoqi, Qi Gong, Baolin Zhang & Endong Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, Melbourne, Australia: Association for Computational Linguistics.

Rao, Gaoqi & Lung-Hao Lee. 2018. NLP for Chinese L2 Writing: Evaluation of Chinese Grammatical Error Diagnosis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association.

Sag, Ivan A, Thomas Wasow, Emily M Bender & Ivan A Sag. 1999. *Syntactic theory: A formal introduction*, vol. 2. CSLI Stanford.

Schneider, David & Kathleen F. McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 2* COLING '98, 1198–1204. Stroudsburg, PA, USA: Association for Computational Linguistics.

Schulze, Mathias. 2008. AI in CALL – artificially inflated or almost imminent? *Calico Journal* 25(3). 510–527.

Toutanova, Kristina, Christopher D. Manning, Dan Flickinger & Stephan Oepen. 2005. Stochastic HPSG parse disambiguation using the Redwoods corpus. *Research on Language & Computation* 3(1). 83–105.

Wang, Maolin, Shervin Malmasi & Mingxuan Huang. 2015. The Jinan Chinese Learner Corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 118–123.

Yu, Liang-Chih, Lung-Hao Lee & Li-Ping Chang. 2014. Overview of grammatical error diagnosis for learning Chinese as a foreign language. In *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications*, 42–47.