Grammatical Error Detection Using HPSG Grammars: Diagnosing Common Mandarin Chinese Grammatical Errors

Luis Morgado da Costa, Francis Bond
Department of Asian Studies, Palacký University Olomouc, The Czech Republic
lmorgado.dacosta@gmail.com, bond@ieee.org

1. Introduction

In recent years, the fields of automated Grammar Error Detection (GED) and Correction (GEC) have gained increased popularity. English has, no doubt, attracted the most attention. However similar efforts have also started for other languages like Mandarin Chinese (Lee et al., 2016; Gaoqi et al., 2017; Rao et al., 2018).

Constraint-based grammars such as HPSG are ideal for GED/GEC as they directly model grammaticality. In this paper we first discuss previous work on *mal-rules* (Schneider and McCoy, 1998) and show that they naturally handle ambiguity even in erroneous utterances. This allows us to reconstruct multiple possible readings from a single ungrammatical input. We then introduce work on *mal-rules* applied to Mandarin Chinese Grammatical Error Detection, based on first-hand data collected from learners of Mandarin Chinese.

To the best of our knowledge, even though there have been projects concerned with implementing 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.

2. Mal-Rules in HPSG

Mal-rules (Schneider and McCoy, 1998) are rules that 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).

In constraint-based linguistic language models, such as those based on HPSG, robustness is an early and ever present concern. When compared with shallow parsing methods, the explicit nature of constraint-based linguistic language models makes these models much less robust. In other words, 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 not true for statistical-based parsers. As such, this rigidity (i.e., the lack of inherit 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.

Using *Mal-Rules* in HPSG grammars has a long history. There have been efforts for English (Bender et al., 2004; Flickinger, 2010; Flickinger and 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. Ambiguating Ungrammaticality

One of the main benefits of using different kinds of *mal-rules* is the ability to carefully ambiguate the task of error correction. To illustrate this, we will take (1) as an example, while arguing that both (2) and (3) are equally plausible corrections.

(1) * *This students sleep.*

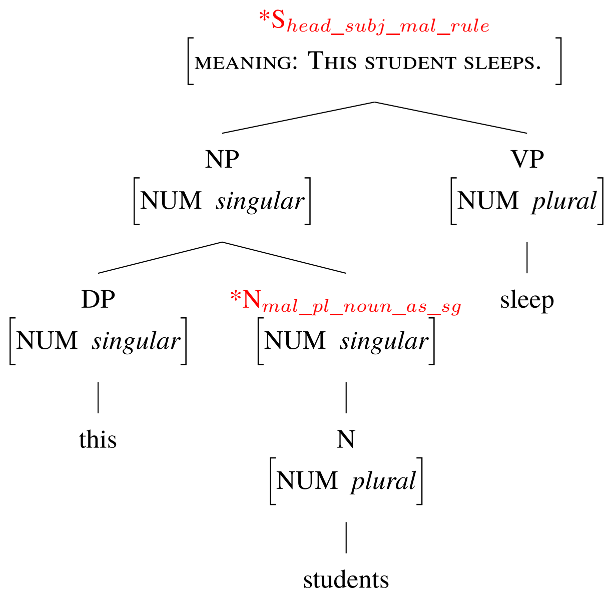
(2) *These students sleep.*

(3) *This student sleeps.*

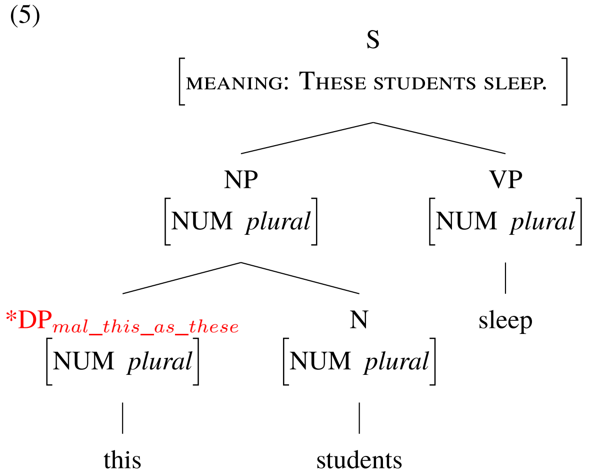
There are, potentially, two sources of ungrammaticality in (1): the first is concerned with the problem of agreement between the determiner *this* and the noun *students*; and the second is concerned with the problem of subject-verb agreement, but is dependent on how the first is corrected.

With three simple mal-rules, a grammar for English would be able to correctly predict the two corrections for (1) presented above. Their parses would be analogous to the trees shown in (4) and (5).

(4)



(5)



Each of the trees above uses a different set of *mal-rules* to produce a correction with a different assumed intended meaning. (4) uses a *lexical mal-rule* (*$N_{mal\_pl\_noun\_as\_sg}$) to change the grammatical number of the form *students* to singular, and a *syntactical mal-rule* (*$S_{head\_subj\_mal\_rule}$) to allow a third-person singular NP be the subject of the verb form *sleep*. The tree shown in (5), on the other hand, uses a single *mal lexical entry* (*$DP_{mal\_this\_as\_these}$) to resolve the ungrammaticality of sentence (1), assuming a third-person plural reading for the subject — and hence requiring only a single *mal-rule*.

For systems where the goal is simply grammatical error detection (i.e. without correction), using a grammar extended with *mal-rules* and traversing the parse tree looking for nodes where *mal-rules* were used is enough to diagnose if and in what way a sentence is ungrammatical (providing the right *mal-rules* are available in the grammar). But when dealing with grammars with generation capabilities, using *mal-rules* to reconstruct different intended meanings, as shown in (4) and (5), would also allow the generation of the corrected counterparts. As such, HPSG grammars can also be used to produce error correction systems.

In the full paper we will expand on the process of deriving (4) and (5) – showing how different levels of specificity and types of *mal-rule* can achieve different results.

3. Mal-Rules for Mandarin Chinese

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).

In this paper, the discussion of *mal-rules* for Mandarin Chinese will be done in the context of a new learner corpus collected from first year learners of Mandarin Chinese 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 in the corpus had at least one error tag assigned to them). A summary of results is shown in Table 1.

The full paper will briefly expand on the contents of this learner corpus by focusing on key grammatical errors, explaining their nature and providing minimal pairs to better understand the design of *mal-rules*.

3.2. Diagnosing Common Grammatical Errors

After discussing key common errors empirically determined by our new Mandarin Chinese learner corpus, we now provide in-depth examples of how to design *mal-rules* to detect these errors with high precision.

For this abstract, we only exemplify this discussion for the most frequent error (ID1) – 吗 *ma* "question particle" redundancy – see below. The full paper will also include a full discussion for a variety of other errors, including: common errors concerning the usage 也 *yě* "also" (ID2 and ID6); the use of copula with adjectival predicates (ID4); and the use of bare nominal predicates (ID 18).

吗 (*ma*, question particle) redundancy (ID1)

The most frequent grammatical error our learners make is the misuse of the question particle 吗 *ma*. The proper use of *ma* transforms propositions into polar (i.e. yes-no) questions. This particle, usually appearing without any other sort of syntactic evidence (e.g., the use of an auxiliary verb), often confuses learners into assuming that it is similar to a question mark (i.e. simply marking the existence of a question: this is how, e.g., the Japanese question marker か *ka* behaves). However, as can be seen in (7) and (9), this is not the case in Mandarin Chinese. In sentences where other interrogatives are used, such as (6), *ma* should not be present. A similar situation happens in (8), where the usage of a special syntactic construction (verb-NOT-verb) already marks the existence of a (polar) question. In this case, it is redundant and ungrammatical to add *ma*, as shown in (9). More

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

generally, *ma* should never be used in sentences that are, by themselves, already questions.

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

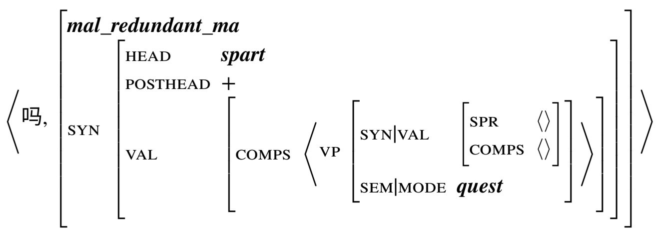
(7) *你 要 什么 吗 ?
nǐ yào shénme ma ?
2SG want Q.what Q.polar ?
(intended) 'What do you want?'

(8) 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?'

(9) *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?'

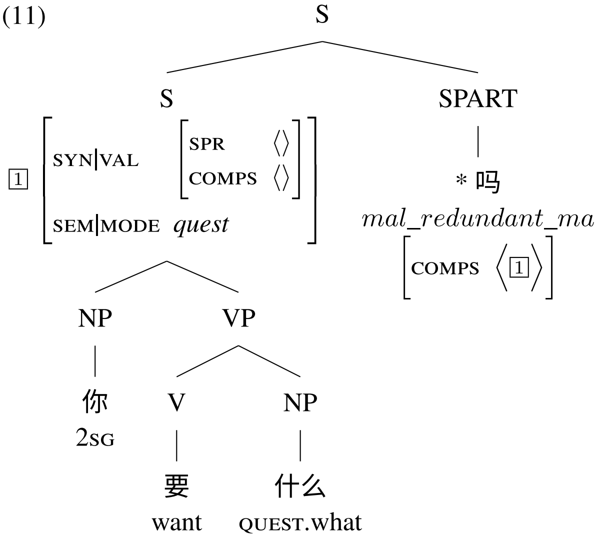
We deal with this error by adding to the grammar an extra *mal lexical entry* for 吗 *ma*, shown in (10). This *mal* lexical entry – which is identified as a mal-rule by the type's name – provides a second entry for *ma* as a sentence final particle (i.e. *spart*). This sentence particle expects a single 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 [POSTHEAD +], 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.

(10)



Using the *mal lexical entry* (10) in an existing grammar of Mandarin Chinese would allow it to parse ungrammatical sentences like the one shown in (11). All similar ungrammatical sentences, where a well formed question (i.e., explicitly marked as [SEM|MODE ***quest***]) is followed by a redundant *ma*, can be detected with this *mal lexical entry*.

(11)



4. Notes on Implementation of *Mal-rules*

In this section we will present some notes on the implementation of these *mal-rules* in Zhong – an open source Mandarin Chinese HPSG grammar (Fan et al., 2015). Zhong is a medium-sized HPSG grammar able to produce Minimal Recursion Semantics representations (Copestake et al., 2005, MRS).

Zhong currently contains more than 60 *mal rules* (including lexical entries). As such, describing each individual rule would not be possible nor desirable, as many *mal rules* share many design principles. In the full paper we will provide an overview of the kind of *mal rules* that have been implemented, clustering them by types, while also discussing why some of the common errors presented in Table 1 could not easily be detected (i.e., imposed by either theoretical limitations or by Zhong's state of development).

We will also discuss two other important ideas: i) the need to develop learner treebanks alongside *mal rules*; and ii) the importance of exploring semantics in error detection in parallel with *mal rules*. Both of these topics will be briefly introduced below.

Learner Treebanks

With the addition of *mal rules*, grammars become increasingly more ambiguous. This is not necessarily a problem in the sense that this ambiguity is reflected on the ability to predict multiple different corrections for the same ungrammatical input, as was discussed for (1) through (3), but it becomes a problem when parsing grammatical sentences – because *mal rules* will be competing with descriptive rules. Investing early in treebanks that contain learner data – which we have named Learner Treebanks (Morgado da Costa et al., 2022) – will enable the creation of *mal rule* enhanced parse ranking models (Toutanova et al., 2005). These models help ranking 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, or the use of heuristics to select the best parse such as fewest number of mal-rules needed to provide a parse).

A new Learner Treebank for Mandarin Chinese is being created in the context of work presented here. However, a full discussion of this resource would fall outside the scope of this paper. Still, this paper will expand on this topic and highlight why Learner Treebanks are an essential resource to make computational grammars ready to be used as pedagogical systems.

Error Diagnosis through Semantics

While this paper focuses mainly on *mal rules*, it is important to note that *mal rules* are not always the best solution to detect certain kinds of common errors. In particular, a large class of errors made by low proficiency language learners are not syntactic in nature but rather semantic.

One important example of this kind of error is the lexical conflation between the words *China* (中国, *zhōngguó*) and *Chinese Language* (中文, *zhōngwén*), error ID 5 in Table 1 – exemplified by sentences (12) through (14).

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

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

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

Although the sentence shown in (13)/(14) is not strictly ungrammatical, learners often use this sentence with the intended meaning shown in (14) – *I speak Chinese*. Other similar, although less frequent, errors based on awkward semantics included the the Mandarin Chinese equivalent of "I am France." or "This is the office's professor.".

While these sentences reveal a problem in the student's knowledge of Mandarin Chinese, they are not technically ungrammatical, and we expect a grammar to produce a parse for them without resorting to *mal rules*. Fortunately, one advantage of working with a grammar capable of producing computationally tractable semantics is the fact that this semantic output can also be used to identify certain kinds of problems in language usage.

Considering the problem exemplified in sentences (12) through (14), it would be possible to identify the use of non-prototypical complements using *mal lexical entries*. One could, for example, create a special *mal lexical entry* for the verb 说 (shuō, *to talk*), as shown in (15).

(15)

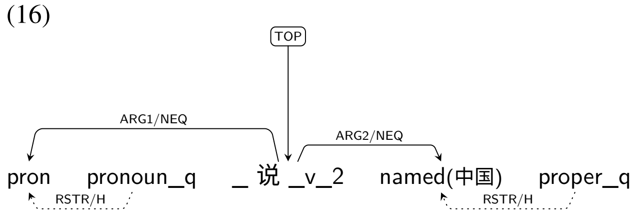
$$\left\langle \text{说}, \begin{bmatrix} \textbf{\textit{mal\_talk\_china}} \\ \text{SYN} \begin{bmatrix} \text{HEAD} & \textit{verb} \\ \text{VAL} & \begin{bmatrix} \text{SPR} & \langle \text{NP} \rangle \\ \text{COMP} & \left\langle \text{NP} \begin{bmatrix} \text{FORM} & \text{中国} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix} \\ \text{SEM} \begin{bmatrix} \text{INDEX} & s \\ \text{MODE} & \textit{prop} \end{bmatrix} \end{bmatrix} \right\rangle$$

This entry selectively chooses its complement to be 中国 (*zhōng guó*, *China*). This means that the sentence (14) would have two available parses – one using the normal entry the verb 说 (shuō, *to talk*) and another using the *mal lexical entry* described above. This ambiguity reflects the duality of the judgments seen in (13) and (14) – one of the parses should be deemed as completely grammatical, while the other can be used to raise a potential problem concerning the non-prototypical complement used with the verb 说 (shuō, *to talk*).

However, and even though the creation of an extremely restrictive *mal lexical entry* for the verb 说 (shuō, *to talk*), as the shown in (15) would not necessarily have a big impact on ambiguity (due to its restrictive complement), it would most likely have some impact on parsing time and memory.

Fortunately, these issues can be completely avoided by using the semantic output produced by the grammar instead of *mal rules*. (16) shows Zhong's simplified semantic representation for (13), as a Dependency MRS (Copestake, 2009).

(16)



Through this representation we can easily confirm 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 lexical entry for 说 (shuō), 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 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 shallower text based methods. This method is better aligned with the nature of the error (i.e. semantic and not syntactic), and thus it will be our suggested method to deal with this and similar errors.

5. Conclusion

At this particular time juncture, when scholars are trying to rediscover the role of formal linguistics in the wider field of Computational Linguistics[1] (currently dominated by statistical and, in particular, 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 Grammatical Error Detection and Correction 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, each *mal-rule* (and semantic check) will be linked to corrective feedback messages describing errors and how to correct them.

6. Acknowledgements

This research project received support from Nanyang Technological University through a Research Scholarship, and 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.

7. Bibliographical References

Bender, E. M., Flickinger, D., Oepen, S., Walsh, A., and Baldwin, T. (2004). Arboretum: Using a precision grammar for grammar checking in CALL. In *InSTIL/ICALL Symposium 2004*.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (2005). Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2-3):281–332.

Copestake, A. (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)*, pages 1–9, Athens.

Costa, F. M., Ralha, J. C. L., and Ralha, C. G. (2006). Aprendizagem de língua assistida por computador: Uma abordagem baseada em HPSG. *Revista Brasileira de Informática na Educação*, 14(1).

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

Flickinger, D. and Yu, J. (2013). Toward more precision in correction of grammatical errors. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 68–73.

Flickinger, D. (2010). Prescription and explanation–using an HPSG implementation to teach writing skills. In *Invited talk, HPSG Conference*.

Gaoqi, R., Zhang, B., Endong, X., and Lee, L.-H. (2017). IJCNLP-2017 task 1: Chinese grammatical error diagnosis. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 1–8.

Granger, S. (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. K. (1994). Unification-based parsing applications for intelligent foreign language tutoring systems. *Calico Journal*, 12(2&3):5–31.

Heift, T. (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*, volume 2, pages 508–512.

Hellan, L., Bruland, T., Aamot, E., and Sandoy, M. H. (2013). A Grammar Sparrer for Norwegian. In *Proceed-*

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

ings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), Oslo, Norway. NEALT Proceedings Series, volume 16.

Lee, L.-H., RAO, G., Yu, L.-C., XUN, E., Zhang, B., and Chang, L.-P. (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 (NLPTEA2016)*, pages 40–48, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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

Rao, G., Gong, Q., Zhang, B., and Xun, E. (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*, pages 42–51, Melbourne, Australia, July. Association for Computational Linguistics.

Schneider, D. and McCoy, K. F. (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, pages 1198–1204, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Toutanova, K., Manning, C. D., Flickinger, D., and Oepen, S. (2005). Stochastic HPSG parse disambiguation using the Redwoods corpus. *Research on Language & Computation*, 3(1):83–105.