Challenges in building a benchmark of linguistic minimal pairs for low resource languages: The case of Malay and Indonesian

Hiroki Nomoto[1] Sri Budi Lestari[2] David Moeljadi[3] Farhan Athirah binti Abdul Razak[1]
Kazuya Inagaki[4] Masashi Furihata[1]
[1]Tokyo University of Foreign Studies [2]Ritsumeikan Asia Pacific University
[3]Kanda University of International Studies [4]Nanzan University
{nomoto, farhan, furihata}@tufs.ac.jp tari0828@apu.ac.jp
moeljadi-d@kanda.kuis.ac.jp inagakik@nanzan-u.ac.jp

Abstract

This paper discusses challenges we faced in developing MALINDO BLiMP (Malay/Indonesian Benchmark of Linguistic Minimal Pairs), making a comparison with the situations faced by the developers of JBLiMP (Japanese Benchmark of Linguistic Minimal Pairs) [5]. The smaller community size and the absence of a suitable crowdsourcing platform made data collection and validation more costly. More than half of our data sources were authored by foreigners alone, which turned out to impair data reliability. At the same time, the results of the acceptability judgement experiment conducted for data validation show that data augmentation by translation is a legitimate method to alleviate the quantitative challenge at least for Malay/Indonesian.

1 Introduction

When language models have achieved certain levels of accuracy in practical downstream tasks, a scientific question arises as to whether and to what extent they possess the kind of linguistic knowledge humans have. Syntactic knowledge is one aspect of such knowledge, and various benchmarks have been developed to investigate it [1, 2, 3, 4, 5, 6, 7]. Since current large language models are capable of handling many low resource languages quite successfully, the question is now relevant not just for high and medium resource languages such as English and Japanese, but also for low resource languages. Consequently, syntactic evaluation benchmarks are needed in those languages too.

We have started building a dataset for targeted syntactic evaluation for two related low resource languages, i.e. Malay and Indonesian:[1)] MALINDO BLiMP.[2)] We closely followed the procedure adopted by the developers of JBLiMP (Japanese Benchmark of Linguistic Minimal Pairs) [5], who collected their data from a linguistics journal and created minimal pairs. This paper discusses challenges we faced in developing MALINDO BLiMP during data collection (§2) and validation (§3). Throughout the paper, we compare our experiences with those of the JBLiMP developers.

2 Procedure

2.1 Unacceptable sentence collection

First, we collected unacceptable sentences from linguistics articles, books and dissertations. Acceptable counterparts were also collected when available. Non-sentential examples were converted into sentences by applying one of the following templates: NP *ada di sini* ‘NP is here’, *Ada* NP ‘There is/are NP’, *Saya tidak tahu tentang* NP ‘I don’t know about NP’, *Saya ada di sini* PP ‘I was here PP’.

This task is more challenging in Malay/Indonesian than in Japanese. While the data for JBLiMP were collected from 28 syntax articles published in a single journal, i.e. *Journal of East Asian Linguistics* (JEAL), between 2006 and 2015, we needed to rely on as many as 70 sources written in four languages (English, Malay, Indonesian, Japanese) for a much longer period (1976–2023). This dif-

[1)] Malay (ISO 639-3: zsm) and Indonesian (ISO 639-3: ind) are two standard varieties of the macrolanguage Malay (ISO 639-3: msa). The former is the national language of Malaysia, Singapore and Brunei whilst the latter is that of Indonesia.

[2)] MALINDO BLiMP-related materials, including materials used in the acceptability judgement experiment discussed in §3, will be made available at https://github.com/matbahasa/MALINDO_BLiMP.

ference reflects the **community size difference**. Although Malay/Indonesian is one of the world's major languages with more than 300 million speakers [8], the number of linguists working on its syntax is quite small, unlike Japanese, which has a large syntactician community. JEAL during the aforesaid period published 133 articles on Japanese [5] but only two on Malay/Indonesian [9, 10]. Similarly, another area-specific journal, *Oceanic Linguistics* published only four Malay/Indonesian syntax articles during the same period [11, 12, 13, 14]. We thus collected data from books and dissertations in addition to journal articles. Moreover, some of our data sources were not available in digital format and we had to manually type the sentences.

Malay/Indonesian and Japanese syntax communities also differ qualitatively, particularly with regard to **the authors' backgrounds**. While the majority (23/28) of JBLiMP's data sources were authored by local linguists (with their foreign colleagues), more than half (39/70) of our data sources were authored by foreigners alone. Malay/Indonesian proficiency varies from author to author, which may affect the data reliability. As we will show in §3.3, this is indeed the case.

2.2 Minimal pair creation

Next, we created minimal pairs by pairing each unacceptable sentence with its acceptable counterpart, which is normally provided in the original data source. In rare cases where an acceptable counterpart was not found in the data source, we constructed one. Moreover, we modified some acceptable sentences that were too different from their unacceptable counterparts to form minimal, or at least near minimal, pairs.

2.3 Data augmentation by translation

In order to accelerate the data collection process, we augmented the data in each language by adding translations of the sentences from the other language. This sort of data augmentation is possible because Malay and Indonesian stand in a dialectal relation and hence share a large portion of their grammars. Yet, their grammars are not completely identical. Hence, this method cannot be applied when the relevant linguistic phenomenon only exists in one language, but not in the other. Moreover, acceptability judgements are not always the same in the two languages. The acceptable sentence in one language may be unacceptable in the other, and vice versa. In such cases, we reversed the acceptability judgements of the translated sentences. In other cases, the translated sentences exhibit no acceptability contrast. Such translated sentences were excluded from the dataset. Other subtle and often unnoticed syntactic differences may affect the final data reliability. We will show in §3.4 that translation is nevertheless a legitimate method.

2.4 Categorization by phenomenon

Finally, we classified the minimal pairs into 12 phenomena consisting of 45 paradigms (cf. Tables 1 and Appendix). Our categorization is modelled on that of JBLiMP, which in turn largely follows BLiMP [2]. Although the categories originally designed for English do not perfectly fit Malay/Indonesian, we prioritized ease of cross-linguistic comparison.[3)]

3 Data validation

We conducted an acceptability judgement experiment to validate the data quality. Here too, we attempted to follow JBLiMP's methodology as much as possible.

3.1 Methodology

JBLiMP developers used a crowdsourcing platform to recruit 240 native speaker participants and administer their acceptability judgement experiment. For Malay/Indonesian, however, **no suitable crowdsourcing platform** was available for recruiting a large pool of speakers for small tasks. We therefore conducted the experiment using a more traditional approach. We administered the experiment at Universiti Kebangsaan Malaysia (Malay) and Universitas Indonesia (Indonesian) with undergraduate students there as our primary participants. In addition, some participants were recruited individually to ensure that all minimal pairs receive judgements from at least 13 native (L1) speakers. The numbers of participants were 152 (L1: 135) for Malay and 156 (L1: 145) for Indonesian.[4)]

A total of 300 minimal pairs were chosen from the 1,195 pairs that had already been annotated with linguistic phenomenon categories. The breakdown of the 300 pairs is shown in Table 1 (see Appendix for a more detailed breakdown). These pairs were divided into ten questionnaires.

[3)] The existing benchmark for Indonesian, LINDSEA, uses language-specific categories such as '*-i/-kan* suffix' and '*Ada*' [7].

[4)] LINDSEA's data validation involved three speakers and 82 minimal pairs [7].

Table 1 Average item-level percentages of native speakers' choices

Phenomenon	# minimal	Malay				Indonesian			
	pairs	same	reverse	both	neither	same	reverse	both	neither
Argument Structure	56	**54.5**	11.8	21.4	12.4	**59.9**	5.9	28.3	5.9
Verbal Agreement	9	**43.5**	12.5	23.2	20.8	**43.0**	9.9	42.4	4.7
Binding	11	**72.5**	5.9	12.9	8.7	**67.9**	7.9	17.5	6.7
Ellipsis	12	**55.8**	5.0	11.2	28.0	**57.2**	9.5	18.0	15.3
Morphology	50	**63.3**	8.6	18.2	9.9	**66.1**	3.5	20.9	9.5
Quantifiers	10	**94.0**	0.7	5.3	0.0	**73.4**	4.0	17.6	5.0
Island Effects	27	27.3	12.8	2.5	**57.4**	30.2	10.7	3.4	**55.7**
Filler-Gap	41	**52.0**	9.3	16.9	21.5	**61.8**	3.6	10.0	24.6
NPI Licensing	9	**82.6**	0.0	16.5	0.9	**80.4**	0.0	12.4	7.2
Nominal Structure	11	**63.4**	4.7	30.6	1.3	**75.7**	1.2	22.4	0.6
Control/Raising	24	**55.1**	7.3	15.3	22.3	**60.9**	5.7	18.7	14.7
Misc./Unknown	40	**66.5**	5.9	13.9	13.7	**66.6**	3.3	12.2	17.8
Total	300	**57.7**	8.5	16.1	17.8	**60.7**	5.2	18.2	15.9

The participants were each asked to judge 30 pairs and to answer several questions about their demographic and linguistic backgrounds using Google Forms. They were explicitly cautioned to base their judgements on their own intuitions, which do not necessarily conform to the standardized normative language taught at school (*bahasa baku*). Every pair, whose sentences were labelled as A and B, was presented with the instruction 'Which option is acceptable?' and the following four response options (both in Malay/Indonesian): A only, B only, A and B, Neither. In this respect, our methodology differs from JBLiMP's, which only provides the first two options. We included the other two options because we as linguists wanted to know more fine-grained judgements. Further, a binary-choice task does not guarantee that one sentence is acceptable and the other unacceptable; it could also be the case that both/neither are acceptable but one sounds better than the other. The order of the minimal pairs and the vertical placement of acceptable and unacceptable sentences within each pair were randomized.

3.2 Results

Table 1 shows the average item-level percentages of the native speakers' choices. Same indicates that the participants selected only the sentence claimed to be acceptable in the original data source whereas reverse indicates that they chose only the sentence claimed to be unacceptable. Both and neither indicate that they selected 'A and B' and 'Neither', respectively. As can be seen in the table, most phenomena received the same judgements as originally reported from more than half speakers. Island Effects exhibits a distinct pattern in that more than half speakers judged both sentences as unacceptable. It is possible that the relevant sentences have empirical issues, though the cause may lie with the speakers, who are not trained linguists, as sentences showing island effects tend to be long and complex.

The following values can be used to evaluate the validity of a minimal pair:

A. same + both (= percentage of speakers who regarded the acceptable sentence(s) as acceptable)

B. same + neither (= percentage of speakers who regarded the unacceptable sentence(s) as unacceptable)

C. both + neither (= percentage of speakers who found no contrast)

We consider a minimal pair to be valid if $A, B \geq 55\%$ and $C < 50\%$. Those pairs which do not satisfy these conditions will be excluded from MALINDO BLiMP. For pairs for which $\mathsf{same} > 0.3(\mathsf{same} + \mathsf{reverse})$, we calculated these values by swapping same and reverse. If they satisfy the conditions above, they will be included in MALINDO BLiMP after reversing the acceptability judgements of the original data sources. Table 2 shows the numbers of minimal pairs that will be included in MALINDO BLiMP. The percentages can be used as human baseline accuracies when evaluating various language models in the future. Note that they are lower than the corresponding figures for JBLiMP, which range from 70.00 to 97.68, except for Quantifiers and Binding in Malay. Multiple factors likely contributed to lower acceptance rates: a method-

Table 2 Number of accepted minimal pairs by phenomenon

Phenomenon	Malay		Indonesian	
Argument Structure	33	(58.9%)	35	(62.5%)
Verbal Agreement	3	(33.3%)	3	(33.3%)
Binding	11	(100.0%)	9	(81.8%)
Ellipsis	6	(50.0%)	7	(58.3%)
Morphology	31	(62.0%)	36	(72.0%)
Quantifiers	10	(100.0%)	8	(80.0%)
Island Effects	5	(18.5%)	10	(37.0%)
Filler-Gap	20	(48.8%)	24	(58.5%)
NPI Licensing	8	(88.9%)	8	(88.9%)
Nominal Structure	7	(63.6%)	9	(81.8%)
Control/Raising	13	(54.2%)	14	(58.3%)
Misc./Unknown	27	(67.5%)	26	(65.0%)
Total	174	(58.0%)	189	(63.0%)

ological difference (four vs. two options), a stricter condition (with vs. without the contrast precondition $C < 50\%$), quality differences of the original data sources, etc.

3.3 Influence of the authors' backgrounds

Figure 1 shows the average item-level percentages of speakers' choices for three types of author backgrounds: local, foreign and mixed. The numbers of data sources included in the three groups are 21, 39 and 10, respectively. The three groups exhibit distinct patterns. The local group attempts to discuss subtle differences between two acceptable sentences (both: 20.7%) more frequently than the other two groups. By contrast, the foreign group tends to compare two unacceptable sentences (neither: 23.5%). In terms of strict replicability (same), the mixed group performs best, followed by the local and foreign groups. The significant difference between the foreign group and the local/mixed groups (Welch's t-test, $t = -2.29/-2.25, p = 0.023/0.026$) suggests the important role played by local linguists in producing reliable language data.[5)]

3.4 Influence of translation

Figure 2 shows the average item-level percentages of speakers' choices for minimal pairs taken directly from the original data source or derived via translation. Translation has no effect on the strict replicability (same) in Indonesian. In Malay, although a slight difference is observed, it does not reach statistical significance (Welch's t-test, $t = 1.24, p = 0.216$). We therefore conclude that translation is a legitimate data augmentation method for Malay/Indonesian.

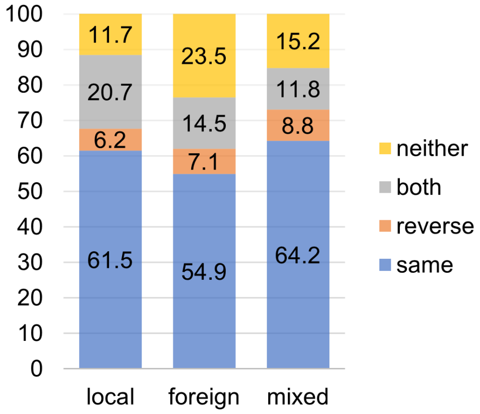

Figure 1 Choice distributions by authors' background

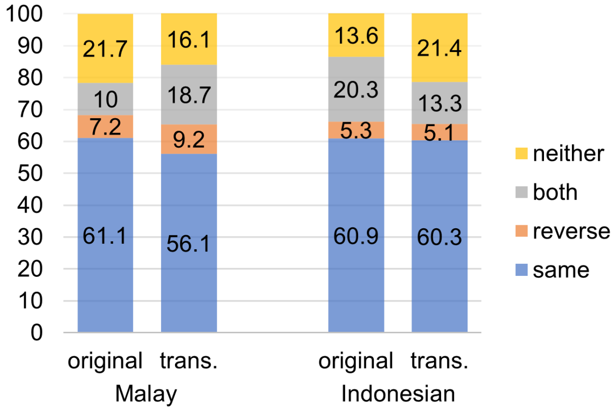

Figure 2 Choice distributions for original vs. translated pairs

4 Conclusion

Working on low resource languages such as Malay/Indonesian presents quantitative and qualitative challenges that researchers working on high or medium resource languages may not face, even when the final product (e.g. benchmark of linguistic minimal pairs) is nominally the same and may be of lower overall quality. The smaller community size makes data collection and validation more costly in terms of time, effort and money. Syntactic studies on low resource languages are often conducted more—sometimes only—by foreign researchers. In our study, the data presented by foreign researchers alone turned out to be less reliable compared to those presented by local researchers (with their foreign colleagues). Data augmentation by translation was shown to be a legitimate way to alleviate the quantitative challenge at least for related languages that share a large portion of their grammars such as Malay and Indonesian.

[5)] No significant difference exists between the local and mixed groups.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23K25336. We thank Normalis binti Amzah, Niken Pramanik and Himawan Pratama for their assistance in recruiting the participants and securing the rooms for the acceptability judgement experiment in Malaysia and Indonesia. We also thank all the speakers who participated in the experiment. Any remaining errors are our own.

References

[1] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 625–641, 2019.

[2] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, 2020.

[3] Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. CLiMP: A benchmark for Chinese language model evaluation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 2784–2790. Association for Computational Linguistics, 2021.

[4] Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. A systematic assessment of language models with linguistic minimal pairs in Chinese. arXiv:2411.06096, 2025.

[5] Taiga Someya and Yohei Oseki. JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs. In **Findings of the Association for Computational Linguistics: EACL 2023**, pp. 1581–1594, Dubrovnik, Croatia, 2023. Association for Computational Linguistics.

[6] Taiga Someya, Yushi Sugimoto, and Yohei Oseki. JCoLA: Japanese Corpus of Linguistic Acceptability. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9477–9488, Torino, Italia, 2024. ELRA and ICCL.

[7] Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. BHASA: A holistic Southeast Asian linguistic and cultural evaluation suite for large language models. arXiv:2309.06085, 2023.

[8] Hiroki Nomoto. Issues surrounding the use of ChatGPT in similar languages: The case of Malay and Indonesian. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, **Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 76–82. Association for Computational Linguistics, 2023.

[9] Hooi Ling Soh and Hiroki Nomoto. The Malay verbal prefix *meN-* and the unergative/unaccusative distinction. **Journal of East Asian Linguistics**, Vol. 20, pp. 77–106, 2011.

[10] Yosuke Sato. P-stranding under sluicing and repair by ellipsis: Why is Indonesian (not) special? **Journal of East Asian Linguistics**, Vol. 20, No. 4, pp. 339–382, 2011.

[11] Dwi Noverini Djenar. On the multifunctionality of compound prepositions in Indonesian. **Oceanic Linguistics**, Vol. 45, No. 2, pp. 404–428, 2006.

[12] Peter Cole, Gabriella Hermon, and Yassir Tjung. Is there *pasif semu* in Indonesian? **Oceanic Linguistics**, Vol. 45, No. 1, pp. 64–90, 2006.

[13] Hooi Ling Soh and Hiroki Nomoto. Progressive aspect, the verbal prefix *meN-*, and the stative sentences in Malay. **Oceanic Linguistics**, Vol. 48, No. 1, pp. 148–171, 2009.

[14] Hiroki Nomoto and Kartini Abd. Wahab. *Kena* adversative passives in Malay, funny control, and covert voice alternation. **Oceanic Linguistics**, Vol. 51, No. 2, pp. 360–386, 2012.

[15] Peter Cole and Gabriella Hermon. The typology of wh-movement: Wh-questions in Malay. **Syntax**, Vol. 1, No. 3, pp. 221–258, 1998.

[16] Alec Marantz. Generative linguistics within the cognitive neuroscience of language. **The Linguistic Review**, Vol. 22, No. 2-4, pp. 429–445, 2005.

Appendix. Detailed version of Table 1

Phenomenon/	# minimal	Malay				Indonesian			
Paradigm	pairs	same	reverse	both	neither	same	reverse	both	neither
Argument Structure	56	**54.5**	11.8	21.4	12.4	**59.9**	5.9	28.3	5.9
passive	8	**48.4**	14.1	10.4	27.1	**77.6**	1.5	9.9	11.0
scrambling	8	**79.7**	1.9	13.0	5.4	**76.6**	0.0	19.1	4.4
animacy	8	**36.7**	22.2	34.8	6.3	41.8	8.5	**46.3**	3.4
aspect	8	36.5	15.1	**42.1**	6.3	38.4	8.8	**47.1**	5.8
internal argument	8	**46.8**	16.8	25.8	10.6	**52.6**	6.8	31.9	8.7
applicative	8	**60.9**	4.7	11.3	23.1	**61.4**	8.3	27.7	2.6
transitivity	8	**72.4**	7.7	12.2	7.7	**70.9**	7.4	16.3	5.4
Verbal Agreement	9	**43.5**	12.5	23.2	20.8	**43.0**	9.9	42.4	4.7
person constraint	9	**43.5**	12.5	23.2	20.8	**43.0**	9.9	42.4	4.7
Binding	11	**72.5**	5.9	12.9	8.7	**67.9**	7.9	17.5	6.7
anaphor	9	**70.6**	6.4	13.2	9.8	**62.5**	8.7	20.6	8.2
reciprocal	2	**81.0**	3.8	11.5	3.6	**92.3**	3.8	3.8	0.0
Ellipsis	12	**55.8**	5.0	11.2	28.0	**57.2**	9.5	18.0	15.3
nominal ellipsis	3	**85.0**	2.6	5.1	7.3	**90.6**	0.0	9.4	0.0
sluicing	9	**46.0**	5.9	13.2	34.9	**46.0**	12.7	20.9	20.4
Morphology	50	**63.3**	8.6	18.2	9.9	**66.1**	3.5	20.9	9.5
part of speech	8	**63.0**	1.9	28.6	6.5	**59.4**	1.0	28.7	11.0
idiom	2	**100.0**	0.0	0.0	0.0	**52.1**	3.8	25.4	18.7
reflexive	8	**69.9**	2.9	15.4	11.8	**59.5**	4.5	35.9	0.0
inflection	8	**75.4**	6.7	17.0	0.9	**71.6**	8.0	16.9	3.5
nominalization	8	**47.6**	15.9	20.8	15.7	**75.1**	1.9	18.2	4.8
voice	8	**56.8**	11.3	10.2	21.7	**76.2**	0.0	7.7	16.2
causativization	8	**58.1**	14.8	21.6	5.5	**58.5**	5.7	16.8	19.1
Quantifiers	10	**94.0**	0.7	5.3	0.0	**73.4**	4.0	17.6	5.0
floating quantifiers	1	**100.0**	0.0	0.0	0.0	**69.2**	0.0	30.8	0.0
classifier	9	**93.3**	0.8	5.9	0.0	**73.9**	4.4	16.2	5.5
Island Effects	27	27.3	12.8	2.5	**57.4**	30.2	10.7	3.4	**55.7**
complex-NP island	8	35.0	1.9	1.9	**61.2**	25.6	5.1	0.0	**69.3**
adjunct island	4	**48.2**	14.7	0.0	37.1	**61.8**	1.9	7.4	28.8
negative island	2	11.5	36.5	3.8	**48.1**	2.5	39.0	12.7	**45.8**
factive island	3	14.7	12.5	4.9	**67.9**	**55.2**	1.7	4.2	38.9
subject island	2	11.0	14.3	0.0	**74.7**	0.0	26.5	0.0	**73.5**
wh- island	8	21.8	16.6	3.6	**58.0**	24.1	13.1	2.9	**59.8**
Filler-Gap	41	**52.0**	9.3	16.9	21.5	**61.8**	3.6	10.0	24.6
intervention effects	1	38.5	0.0	7.7	**53.8**	23.1	0.0	15.4	**61.5**
relative clause	8	**55.8**	10.0	23.0	11.1	**58.3**	9.2	15.5	16.9
cleft	8	**84.1**	2.7	3.7	9.5	**67.5**	2.6	2.6	27.3
resumptive pronoun	8	36.5	8.2	16.8	**38.5**	**55.7**	3.9	11.5	29.0
wh- question	8	**50.9**	14.1	8.4	26.6	**62.1**	0.0	8.4	29.5
topicalization	8	34.3	16.1	**34.4**	16.1	**65.3**	3.3	17.9	13.5
NPI Licensing	9	**82.6**	0.0	16.5	0.9	**80.4**	0.0	12.4	7.2
NPI	9	**82.6**	0.0	16.5	0.9	**80.4**	0.0	12.4	7.2
Nominal Structure	11	**63.4**	4.7	30.6	1.3	**75.7**	1.2	22.4	0.6
modifier	9	**56.2**	5.7	36.5	1.6	**71.2**	1.5	26.6	0.7
measure phrase	2	**96.2**	0.0	3.8	0.0	**96.2**	0.0	3.8	0.0
Control/Raising	24	**55.1**	7.3	15.3	22.3	**60.9**	5.7	18.7	14.7
subject control	8	**40.2**	4.7	12.0	43.1	**45.5**	6.9	18.6	29.0
object control	8	**87.6**	1.1	4.2	7.1	**95.5**	1.0	0.7	2.9
complementation	8	**47.5**	12.6	30.4	9.5	**49.4**	8.6	33.7	8.3
Misc./Unknown	40	**66.5**	5.9	13.9	13.7	**66.6**	3.3	12.2	17.8
pronoun	8	**51.6**	2.9	29.5	16.0	**45.8**	3.5	22.7	28.0
verbal structure	8	**78.5**	3.8	11.5	6.3	**82.1**	0.8	10.8	6.2
focus	8	**56.5**	14.4	7.3	21.8	**53.8**	5.6	11.9	28.7
negation	8	**74.2**	2.7	13.8	9.3	**78.5**	1.7	6.0	13.9
copula	8	**71.8**	5.6	7.6	15.0	**72.7**	5.0	9.8	12.4
Total	300	**57.7**	8.5	16.1	17.8	**60.7**	5.2	18.2	15.9