Annotation of modal expressions in Indonesian

Hiroki Nomoto[1] Jozina Vander Klok[2] David Moeljadi[3]
[1]Tokyo University of Foreign Studies [2]Humboldt-Universität zu Berlin
[3]Kanda University of International Studies
nomoto@tufs.ac.jp jozina.vander.klok@hu-berlin.de moeljadi-d@kanda.kuis.ac.jp

Abstract

We developed an annotated dataset to investigate six modal expressions in Indonesian, i.e. *harus* 'must', *harusnya*, *seharusnya* 'should', *mesti* 'must', *mestinya*, *semestinya* 'should'. The data consists of 600 sentences. Three native speaker annotators annotated them with tags concerning modal force, modal flavour, mood, etc. The annotation results provide quantitative information about the relevant modal forms, which have been missing in the literature. Moreover, they validate some previous qualitative descriptions, but invalidate others.

1 Introduction

Corpus-based investigations into modality (= the semantic category pertaining to necessity and possibility expressed by words such as *must* and *can* in English) have been increasingly popular in recent years to better understand the range of uses of modality in natural language (see [1], [2] and the references therein). We thus created a dataset to investigate modality in Indonesian (ISO 639-3: ind), in particular the syntactic and semantic contribution of the derivational affixes *-nya* and *se-. . . -nya*. This paper reports how we created the dataset and some initial findings based on it. The resources we developed are available on our project github page.[1)]

2 Background

2.1 Modality

Modality is a semantic category concerning necessity and possibility. Expressions conveying modality are called 'modal expressions'. The word *modal* is also used as a noun referring to a class of modal expressions occurring in a particular position in a sentence. English modals include words such as *must*, *should* and *can*.

Modal expressions have two main meaning components, modal force and modal flavour [3, 4, 5, 6, 7]. Modal force concerns possibility and necessity, and an additional meaning component, modal strength can also distinguish beyond these, such as weak necessity:

(1) a. **Possibility:** Ann *may* buy a lottery ticket because she turned 18 years old.
b. **Necessity:** Ann *must* buy a lottery ticket because her boss ordered her to.
c. **Weak necessity:** Ann *ought to* buy a lottery ticket today if the odds are good.

Modal flavour concerns the type of modality. Although various flavours have been identified in the literature, this study distinguishes just two, i.e. epistemic and root modality. Epistemic modality expresses necessity/possibility relative to the speaker's knowledge whereas root modality expresses necessity/possibility relative to other aspects, which include rules, regulations and facts of the actual world. In English, a modal form can express either epistemic or root modality, as exemplified by (2).

(2) a. **Epistemic:** Ann *may* buy a lottery ticket today because she's feeling lucky.
b. **Root:** Ann *may* buy a lottery ticket today because she just turned 18 years old!

2.2 Indonesian modals

The basic word order of an Indonesian clause is 'subject-predicate' and modals canonically occur at the front of predicate, as in (3a). Modal adverbs can also occur in other positions, as in (3b).

(3) a. Besok saya *harus/seharusnya* ke kedutaan.
tomorrow I must/should to embassy
'I must/should go to the embassy tomorrow.'

1) https://github.com/matbahasa/IndoModal

b. *Seharusnya* besok saya ke kedutaan.
should tomorrow I to embassy
'I should go to the embassy tomorrow.'

Some modals can derive into modal adverbs by means of the circumfix *se-. . . -nya*, as is the case with *seharusnya* above. In addition, they can be suffixed by *-nya* to form another type of modal expression.[2)] The forms with *-nya* are considered colloquial [9]. This study focuses on *harus* and *mesti* as well as their derivatives: *harus-harusnya-seharusnya* and *mesti-mestinya-semestinya*.

The semantics of these forms is not fully understood. Regarding modal force, researchers agree that the derived forms express weak necessity. However, consensus does not seem to exist about the stem form. Some regard it as expressing just necessity. For instance, [10] translate *harus* just as *must*. Others regard it as covering both necessity and weak necessity. For instance, the English equivalents of *harus* provided by [11] and [12] include not only *must* and *have to* but also *should* and *ought to*. Similarly, the Indonesian-Japanese/Japanese-Indonesian dictionary by [13] treat *harus* as ambiguous between two senses, i.e. *shinakereba naranai* (necessity) and *subekidearu* (weak necessity). Regarding modal flavour, Indonesian modals can be either epistemic or root. However, it is not as obvious as with English modals. While recognizing both meanings for *mesti* and *semestinya*, most dictionaries only show the root meaning for *harus* and *seharusnya*. Furthermore, the *(se-. . .)-nya* forms are sometimes associated with counterfactual mood [12, 8]. Thus, [8] presents the contrast in (4).

(4) a. Kamu *harus* datang.
you must come
'You should come.'
b. *Harus-nya* kamu datang.
must-NYA you come
'You should have come.' (Arka 2013: (37))

Whether the *-nya* forms are included and, if they are, what they mean vary from dictionary to dictionary. Corpus data should shed some light to these unclear areas.

2) The syntactic category of the derived form is unclear. It is a noun if *-nya* is a nominalizer (cf. *kapan beli**nya**?* 'When did you buy it? [lit. When was the buy**ing** (event)?]') [8], but it is an adverb if *-nya* is an adverbializer (cf. *biasa**nya*** 'normal**ly**') or if *-nya* is a contracted form of *se-. . . -nya* [9].

Table 1 Target modal forms and their frequencies

Stem form		*-Nya* form		*Se-. . . -nya* form	
harus	27,245	*harusnya*	224	*seharusnya*	2,057
mesti	524	*mestinya*	314	*semestinya*	247

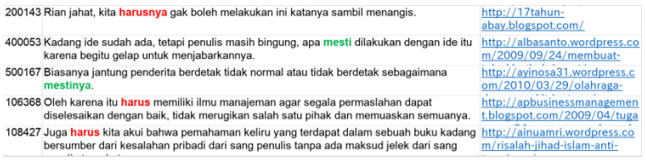

Figure 1 Snapshot of the annotation file

3 Methodology

3.1 Data

We employed three Indonesian subcorpora of the Leipzig Corpora Collection (LCC) [14], i.e. mixed-tufs4, web-tufs13 and wikipedia-tufs16. Each subcorpus contains 300K sentences (4,823,624 tokens on average) automatically collected from the Internet by web crawling. The subcorpora we employed have underwent an additional language identification process to minimize Malay sentences [15].

We used the MALINDO Conc concordancer[3)] [16] to search the three subcorpora for sentences containing the six target forms. Table 1 summarizes the frequency of each form in the entire subcorpora.

Next, we randomly selected 100 sentences for each form, thereby building a dataset comprising 600 sentences in total. These sentences were randomly presented with their source URLs in an MS Excel sheet (Figure 1). The URLs allow the annotators to refer to the discourse context in which the sentence is used. Note that some URLs had become obsolete when the annotation was conducted.

3.2 Annotators and training session

This dataset was annotated by the following three native speakers of Indonesian:

- Annotator 1: university lecturer, Ph.D. degree in linguistics, from Java
- Annotator 2: university lecturer, Master's degree in linguistics, from Bali
- Annotator 3: undergraduate student of Japan studies, from Riau

3) https://malindo.aa-ken.jp/conc/

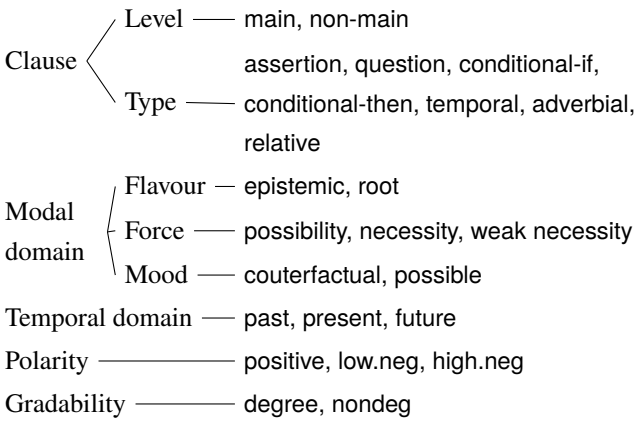

Figure 2 Tagset

A training session was conducted in Indonesian before the annotation task, during which the annotation guidelines were explained and 24 sentences (4 sentences × 6 forms) were annotated together. Follow-up discussions and clarifications were conducted afterwards.

3.3 Tagset

We used 24 annotation tags, grouped into five categories, as outlined in Figure 2. Our tagset is mostly based on those of [1] and [17]. The Clause category pertains to the syntactic context in which the modal form in question occurs. The tags high.neg and low.neg in the Polarity category capture the relative position between the modal and negator, i.e. whether the negator occurs before (high.neg) or after (low.neg) the modal. They are not about semantic scope. The tag degree in Gradability is assigned when the degrees of necessity/possibility are compared, as in *You should call Barbara more than (you should call) Alice* [18].

4 Results and discussion

4.1 Modal force

The annotation results showed a clear flip between *harus/mesti* and their affixed derivatives with regard to force (Figure 3). The sentences with the former were mostly annotated with necessity and those with the latter with weak necessity (none was annotated with possibility).

The results endorses the previous qualitative descriptions reviewed in section 2.2. The affixes *-nya* and *se-. . . -nya* indeed indicate weak force. It is notable, however, that moderate inter-annotator variability was found with *mesti* and *semestinya*, as shown in Table 2. Note that standard

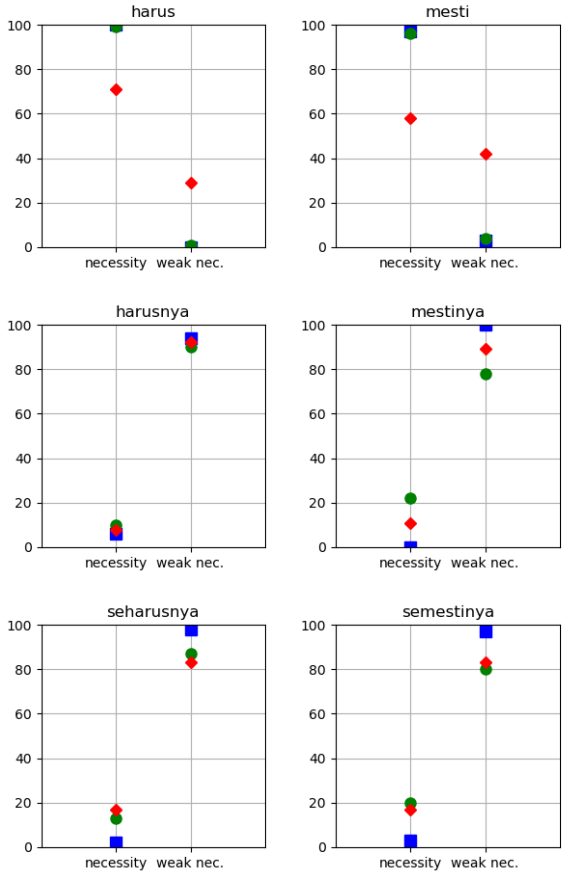

Figure 3 Annotation results: Modal force
(■ Annotator 1, ● Annotator 2, ♦ Annotator 3)

Table 2 Inter-annotator agreement: Modal force

Metric	*Harus*	*H-nya*	*se-H-nya*	*Mesti*	*M-nya*	*se-M-nya*	Total
%	80.0	86.7	90.0	71.3	91.7	76.7	80.3
κ	−0.01	0.09	0.21	0.06	0.15	0.01	0.58

metrics such as Fleiss' κ and Krippendorf's α are known to be unreliable "when the units to be rated are not well-distributed across the rating categories" [19]. The latter situation applies to our data when individual forms are considered. We thus present simple agreement percentage too, although it has the problem of ignoring the possibility of chance agreement. The pattern shown by Annotators 1 and 2 matches the description whereby the stem form expresses necessity, but not weak necessity. By contrast, the pattern shown by Annotator 3 matches the description whereby they express both necessity and weak necessity.

4.2 Modal flavour

The results for modal flavour are shown in Figure 4, along with the inter-annotator agreement scores in Table 3. Unlike modal force, modal flavour is not affected by affixation.

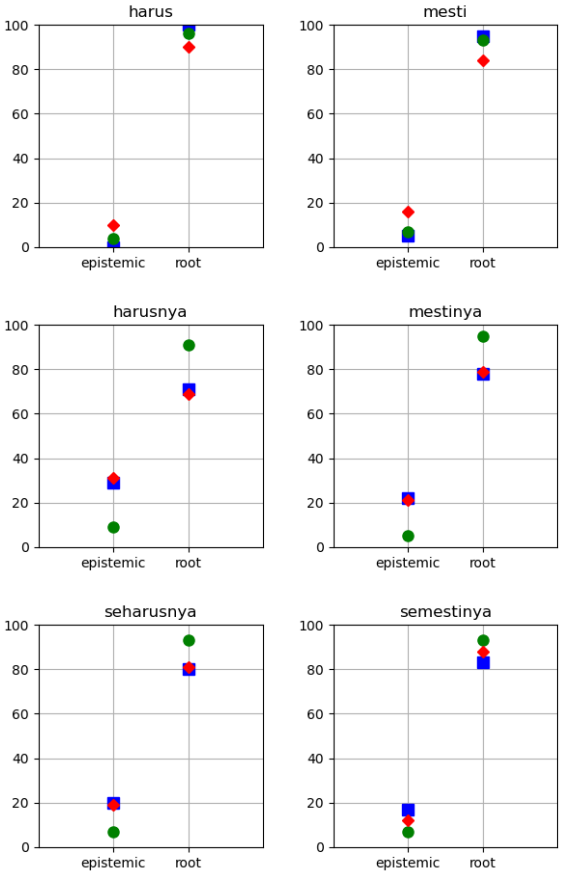

Figure 4 Annotation results: Modal flavour (■ Annotator 1, ● Annotator 2, ♦ Annotator 3)

Table 3 Inter-annotator agreement: Modal flavour

Metric	*Harus*	*H-nya*	*se-H-nya*	*Mesti*	*M-nya*	*se-M-nya*	Total
%	91.3	65.3	86.7	82.7	91.3	81.3	79.3
κ	0.04	0.05	0.11	−0.01	0.21	0.12	0.11

For all forms and all annotators, the numbers of sentences annotated with root are far larger than those of sentences annotated with epistemic. Again, this makes standard inter-annotator agreement metrics unreliable.[4] Moreover, such a skewed distribution may be the reason why the existence of the epistemic use is not asserted explicitly in the literature. Nevertheless, there are sentences that were annotated with epistemic by all annotators (*harus* 1, *harusnya* 6, *seharusnya* 2; *mesti* 2, *mestinya* 2, *semestinya* 1).[5]

4) Indonesian differs from English in this respect. As epistemic and root readings are more well-distributed in English, similar modal annotation studies on English were able to report high inter-annotator agreement scores for modal flavour: $\kappa = 0.84$ for epistemic vs. root [20] and $\alpha = 0.89$ for priority vs. non-priority [1].

5) The exact sentences are given in Appendix A.

4.3 Mood

Although the *(se-. . .)-nya* forms are sometimes associated with counterfactual mood, the results contain numerous counterexamples. The numbers of sentences annotated with possible (i.e. non-counterfactual) by all annotators are as follows: *harusnya* 61, *seharusnya* 67, *mestinya* 53, *semestinya* 59.[6] This fact suggests that the relevant forms can express counterfactuality, but they do not necessarily do so, contra the proposal by [8].

5 Conclusion

To the best of our knowledge, this study is the first attempt at annotating Indonesian texts with respect to modality. We developed a set of annotation tags, adapting for Indonesian those employed by previous modal annotation studies on English and Oceanic languages of Melanesia. We believe our tagset and its associated guidelines are useful for others interested in modality in Indonesian and related languages in the region.

The results of our modal annotation provide support for some previous qualitative descriptions, but counterexamples for others. Moreover, they present quantitative information about the six target modal forms (cf. Table 1) and various modal meanings they convey (cf. Figures 3–4). As can be seen from the figures, the distributions of the meaning categories are considerably skewed in Indonesian. In terms of modal force, the stem forms and the affixed forms are biased towards necessity and weak necessity, respectively. In terms of modal flavour, the root interpretation is much more frequent for all forms. Therefore, extra care is required in analysing the infrequent meanings.

Lastly, our study has at least two limitations. First, it only deals with *harus*, *mesti* and their derivatives. Indonesian has more modal expressions such as *bisa* ‘can’, *pasti* ‘certain’ and *mau* ‘want, will’. Sentences with these other modal expressions can be annotated using our tagset. Second, our dataset consists only of 600 sentences. Although it may be considered sufficiently large for the purpose of certain types of linguistic studies, it is small for NLP tasks. Moreover, the sentences appear without their surrounding discourse contexts. A large-scale annotation project working on a collection of documents, ideally an existing corpus or corpora, will certainly overcome these shortcomings.

6) See Appendix B for example sentences.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23H00639.

References

[1] Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz, and Paul Portner. Toward fine-grained annotation of modality in text. In Paul Portner, Aynat Rubinstein, and Graham Katz, editors, *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pp. 38–46, Potsdam, 2013. Association for Computational Linguistics.

[2] Helena Bermúdez-Sabel, Francesca Dell'Oro, and Paola Marongiu. Multi-layered semantic annotation and the formalisation of annotation schemas for the investigation of modality in a Latin corpus. *Language Resources and Evaluation*, Vol. 58, 2024.

[3] Angelika Kratzer. What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, Vol. 1, pp. 337–355, 1977.

[4] Angelika Kratzer. The notional category of modality. In H.-J. Eikmeyer and H. Rieser, editors, *Words, Worlds, and Contexts: New Approaches in Word Semantics*, pp. 38–74. Mouton de Gruyter, Berlin, 1981.

[5] Angelika Kratzer. Modality. In Arnim von Stechow and Dieter Wunderlich, editors, *Semantics: An International Handbook of Contemporary Research*, pp. 639–650. Mouton de Gruyter, Berlin, 1991.

[6] Paul Portner. *Modality*. Oxford University Press, Oxford, 2009.

[7] Valentine Hacquard. Modality. In Klaus von Heusinger, Claudia Maienborn, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, Vol. 2, pp. 1484–1515. de Gruyter, Berlin, 2011.

[8] I Wayan Arka. On the typology and syntax of TAM in Indonesian. *NUSA*, Vol. 55, pp. 23–40, 2013.

[9] Hasan Alwi. *Modalitas dalam Bahasa Indonesia*. Penerbit Kanisius, Yogyakarta, 1992.

[10] James Neil Sneddon, Alexander K. Adelaar, Dwi N. Djenar, and Michael Ewing. *Indonesian: A Comprehensive Grammar*. Routledge, London, 2nd edition, 2010.

[11] George Quinn. *The Learner's Dictionary of Today's Indonesian*. Allen & Unwin, Sydney, 2001.

[12] Alan M. Stevens and A. Ed. Schmidgall-Telling. *A Comprehensive Indonesian-English Dictionary*. Ohio University Press, Athens, OH, 2004.

[13] Kyoko Funada, Yoshihiro Takadono, and Masanori Sato, editors. *Puroguresshibu Indoneshiago Jiten [Progressive Indonesian Dictionary]*. Shogakukan, Tokyo, 2018.

[14] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 759–765, Istanbul, 2012. European Language Resources Association.

[15] Hiroki Nomoto, Shiro Akasegawa, and Asako Shiohara. Reclassification of the Leipzig Corpora Collection for Malay and Indonesian. *NUSA*, Vol. 65, pp. 47–66, 2018.

[16] Hiroki Nomoto, Shiro Akasegawa, and Asako Shiohara. Building an open online concordancer for Malay/Indonesian. Paper presented at the 22nd International Symposium on Malay/Indonesian Linguistics (ISMIL), 2018.

[17] Annika Tjuka, Lena Weißmann, and Kilu von Prince. Tagging modality in Oceanic languages of Melanesia. In Annemarie Friedrich, Deniz Zeyrek, and Jet Hoek, editors, *Proceedings of the 13th Linguistic Annotation Workshop*, pp. 65–70, Florence, 2019. Association for Computational Linguistics.

[18] Paul Portner and Aynat Rubinstein. Extreme and non-extreme deontic modals. In Nate Charlow and Matthew Chrisman, editors, *Deontic Modality*. Oxford University Press, Oxford, 2016.

[19] David Quarfoot and Richard A. Levine. How robust are multirater interrater reliability indices to changes in frequency distribution? *The American Statistician*, Vol. 70, No. 4, pp. 373–384, 2016.

[20] Valentine Hacquard and Alexis Wellwood. Embedding epistemic modals in English: A corpus-based study. *Semantics and Pragmatics*, Vol. 5, No. 4, pp. 1–29, 2012.

A Sentences annotated with 'epistemic' by all annotators

A.1 *Harus*

1. Justru mereka yang tidak sungguh-sungguh memuji Tuhan dan mengolok-olok teman lain dengan kata fanatik, dialah yang *harus* bertobat. [ID: 101357]
   'In fact, those who do not truly praise God and ridicule other friends with the word fanatic are the ones who *must* repent.'

A.2 *Harusnya*

1. Kemiskinan, kemelaratan, minimnya akses informasi, *harusnya* itu semua justru bisa kita atasi agar umat ini bisa maju BERSAMA. [ID: 200024]
   'Poverty, destitution, lack of access to information, we *should* be able to overcome all of them so that these people can move forward TOGETHER.'
2. Album ini memuat 13 lagu dengan hits Pencuri Hati, *Harusnya* Kau Sadari, dan Senja di Jakarta. [ID: 200139]
   'This album contains 13 songs with the hits Thief of Hearts, You *Should* Be Aware, and Dusk in Jakarta.'
3. Ya terakhir saya melihat dia sedang menghina kangen band khususnya vocalistnya yang bertampang jelak padahal sesama muka pas-pasan *harusnya* saling mendukung! [ID: 200142]
   'Yes, the last time I saw, he was insulting the band, especially the vocalist, who had an ugly face, even though people with mediocre faces *should* be supporting each other!'
4. Perlajaran SD yang *harusnya* di ajarkan sekolah SD dulu tidak secara kaffah. [ID: 200166]
   'The elementary school lessons that *should* have been taught in elementary schools were not taught in a comprehensive manner.'
5. Lupa, *harusnya* tadi bilang kalau mau pesan bagian dada. [ID: 200195]
   'I forgot, I *should* have said earlier that I wanted to order the chest part.'
6. Dan tidak lupa, mereka juga merengek minta jajan (di kamusku yang *harusnya* traktir adalah mereka). [ID: 200203]
   'And don't forget, they also whine for snacks (in my dictionary the one who *should* treat is them).'

A.3 *Seharusnya*

1. *Seharusnya* kita seperti muslim Timur Tengah yang militan? [ID: 300155]
   '*Should* we be like militant Middle Eastern Muslims?'
2. Harry menolak untuk membicarakan tentang kematian Cedric dan menawarkan uang hadiah Turnamen Triwizard kepada orangtua Cedric, dan mengatakan bahwa uang itu *seharusnya* milik Cedric. [ID: 301301]
   'Harry refused to talk about Triwizard and offered the Tournament prize money to Cedric's parents, and said that the money *should* belong to Cedric.'

A.4 *Mesti*

1. Mereka para wakil rakyat belum banyak tahu apa yang *mesti* dilakukan dan dikerjakan. [ID: 400008]
   'They, the people's representatives, don't really know what *must* be done and carried out.'
2. Akan tetapi, tanpa ada dorongan ini pun *mesti* diakui bahwa manusia membutuhkan orang lain dan secara alami membangun kontak sosial. [ID: 400500]
   'However, without any encouragement *must* be acknowledged that humans need other people and naturally build social contacts.'

A.5 *Mestinya*

1. Bahkan tahun wafatnya pun yang *mestinya* diketahui dengan jelas oleh para pengikutnya, juga belum bisa dipastikan hingga hari ini. [ID: 500159]
   'Even the year of his death, which *should* be clearly known to his followers cannot be confirmed to this day.'
2. Menurutnya, warga yang sudah sadar pajak *mestinya* bisa dilayani lebih baik dan mudah. [ID: 500229]
   'According to him, citizens who are aware of taxes *should* be able to be served better and easier.'

A.6 *Semestinya*

1. *Semestinya* begitu, tapi entahlah ayah tidak mau. [ID: 600173]
   'That's how it *should* be, but somehow father doesn't want to.'

B Examples of non-counterfactual *harusnya* and *mestinya*

1. Rumus yang dikembangkan oleh peneliti dan pakar statistik dari University of New South Wales diklaim bisa memprediksi usia ideal kapan seseorang akan atau *harusnya* menikah. [ID: 200179]
   'The formula developed by researchers and statisticians from the University of New South Wales is claimed to be able to predict the ideal age when someone will or *should* get married.'
2. [. . .] pada saat kita hendak marah kepada orang lain *mestinya* kita segera mengingat Allah sehingga tidak melampiaskan kemarahan dengan hal-hal yang tidak benar. [ID: 500243]
   '[. . .] when we want to be angry with others we *should* immediately remember Allah so as not to vent our anger with things that are not true.'