[Last modified: 30 July 2021]

Modal tagging – Annotation guidelines for annotators

Jozina Vander Klok, Hiroki Nomoto & David Moeljadi

1. Extraction of modal examples from MALINDO Conc.

This study investigates the following target words, which are **modals**:

- *mesti, mestinya, semestinya, (?)se-mesti*
- *harus, harusnya, seharusnya, (?)se-harus*

2. Annotation guidelines

Guidelines are explained for each set of tags, with examples in English and also some in Indonesian. Please keep in mind we are only focusing on tagging *mesti, harus* (and their variants), but not (yet) other modals in Indonesian which include *pasti, perlu, bisa, boleh,* etc..

2.1 Guidelines for CLAUSE category

- This category has two sections: CLAUSE LEVEL and CLAUSE TYPE.
- **Clause level** is about where the modal (*mesti, harus…*) is found: either in the main clause or not.
    - If the modal is in the main clause, choose the tag **main**.
    - If the modal is in the embedded clause, choose the tag **non-main**.
- **Clause type** concerns the *type* of main or non-main clause.
    - The tag **assertion** applies to either main or non-main clauses. Another way to describe this type can be a 'declarative'. With non-main clauses, an overt complementizer could include *bahwa, kalau, supaya, agar*.
    - The tag **question** applies to either main or non-main clauses. With **non-main clauses**, an overt complementizer that introduces the embedded question could include *apakah, jika, kalau*

    The following tags only are for sub-types of NON-MAIN clause levels.
    - **conditional-if** is chosen when the modal is found within the 'if-clause', which we take to be a non-main clause level. In Indonesian, the 'if-clause' may be headed by *jika, kalau, sekiranya, seandainya,* etc.
    - **conditional-then** is chosen when the modal is found within the 'then-clause', which we also take to be a non-main clause level.
    - **temporal** is chosen when the modal is found within a **non-main clause** headed by some temporal marker such as 'when, before, after' (Indonesian: *apabila, saat, ketika, sebelum, setelah, sampai, (se)hingga,* etc.).
    - **adverbial** (non-temporal) is chosen when the modal is found within a non-main clause headed by some adverbial that is non-temporal, such as 'because, since, for' (Indonesian: *karena, untuk, meskipun, walaupun).* This is for e.g. purposive, concessive, etc. embedded clause types.

- **relative** is chosen when the modal is found within a relative clause, which modifies a noun phrase and can be overtly headed by *yang* in Indonesian.

Category	Name	Tags	Examples (not exact translations)
Clause	Clause level	**main**	John must be home at 6pm. *Tono harus pergi sekarang.*
		non-main	Amina knows that John must/should be home at 6pm. *Amina tahu bahwa Tono harus pergi sekarang.*
Clause	Clause type	**assertion** (applies to main and non-main)	John must be home at 6pm. [MAIN] *Tono harus pergi sekarang.* You must/should be home at 6pm! [MAIN] *Anda seharusnya pergi sekarang.* Amina knows that John must/should be home at 6pm. [NON-MAIN] *Amina tahu bahwa Tono harus/seharusnya pergi sekarang.*
		question (applies to main and non-main)	Must John be home at 6pm? [MAIN] *Apakah Tono harus pergi sekarang?* *Haruskah Tono pergi sekarang?* Amina doesn't know whether John must be home. [NON-MAIN] Amina wondered whether/if John must be home. *Amina tidak tahu apakah Tono harus pergi sekarang.* *Amina ragu-ragu kalau Tono bisa pergi sekarang.*
		conditional-if (non-main only)	**If** John **must** be home at 6pm, then I will be home later. ***Kalau*** *Tono* ***harus*** *pergi sekarang, saya akan pergi nanti.*
		conditional-then (main only)	If it is 6pm, (**then**) John **must** be home. *Kalau sudah jam enam,* (***maka***) *Tono* ***harus*** *pergi.*
		temporal (non-main only)	{**When/before/after**} John **must** be home, we will play football. We will play football {**when/before/after**} John **must** be home ***Sebelum*** *Tono* ***bisa*** *pergi ke Jepang, kami sudah ke sana dua kali.* *Kami sudah ke Jepang dua kali* ***sebelum*** *Tono* ***bisa*** *pergi ke sana.*
		adverbial (non-temporal)	I am happy {**because/since/for**} John **must** be home now.

		(non-main only)	*Saya senang **karena** Tono **bisa** pergi ke Jepang.*
		relative (non-main only)	The student **that must** finish their homework today is sick. *Mahasiswa **yang harus** menyelesaikan tugasnya hari ini sakit.*

2.2 Guidelines for MODAL and TEMPORAL DOMAIN categories

These two categories are explained together because modal-temporal relations are closely connected in natural language.

- Modal markers or constructions across all of the world's languages express at least two dimensions of meaning: (i) whether the modal expresses a possibility or necessity claim (called **modal force**), and (ii) what the possibility or necessity claim is based on (e.g. based on inference, or based on rules), called **modal flavour** or type of modality. A third dimension of meaning is about whether or not the possibility or necessity claim is 'strengthened' or 'weakened', called **modal strength**
- **The target words are *harus, mesti* and their variants. We will use other modal words in Indonesian (e.g., *bisa, boleh, perlu*) for some example sentences to illustrate for these tags. Other example sentences are only illustrated in English.**

Category	Name	Tags	Examples
Modal domain	Modal flavour	**epistemic**	John **must** be home (since I know he always comes home at this time.) [Based on available knowledge; inference] *Tono pasti di rumah.*
		root (e.g., non-epistemic)	John **must** be home (because the judge ordered that he is under house arrest) [Root → e.g., based on rules/regulations] John has not been able to go to the washroom for 5 hours! He **has to** go pee. [Root → e.g., participant-internal necessity] John **must** take Highway 22 to reach Toronto (because there is only one highway) [Root → e.g., goal-oriented, in order to reach Toronto by car]
	Modal force /strength	**possibility**	John **may/might** be home (since I know he sometimes comes home at this time, but its not always the case.) [EPISTEMIC] *Tono **mungkin** di rumah sekarang*.

			Since John is now 18 years old, he **may/can/is allowed to** buy a lottery ticket (as based on the laws of Canada) [ROOT]
		weak_necessity	John **should** be home (since I know he usually comes home at this time, but lately he has been working late) [EPISTEMIC; based on speaker's knowledge] John **should/ought to** clean his room (since his mother asked him to) [ROOT; based on expectations of his mother] John **should/ought to** wash his hands before eating [ROOT; based on cultural norms]
		necessity	John **must** be home (since I know he always comes home at this time.) [EPISTEMIC] John **must** be home (because he is under house arrest as ordered by the court of law) [ROOT; based on the law]
	Mood	**counterfactual** (CF)	If John had played yesterday, he **should** have passed the ball to Amina (but he didn't play). / If John hadn't been sick, he **should** have played yesterday. [PAST CF] if John were playing right now, he **should/must** pass the ball to Amina (but it is known that he is not actually playing now) [PRESENT CF] If John played tomorrow, they **should** win. (but he will not play because he is out of town) [FUTURE CF]
		possible	*all non-counterfactual ones, see examples below across epistemic and root modal flavours*
Temporal domain	Temporal domain	**past**	Uttered at 11pm at night: John **must have** been home all day (the lights were always on). [PAST, EPISTEMIC, POSSIBLE] John **must have** been home all day (because he is under house arrest as

			ordered by the court of law) [PAST, ROOT, POSSIBLE]
		present	John **must** be home now (I see the lights on) [PRESENT, EPISTEMIC, POSSIBLE] John **must** take his medicine now (e.g. because the doctor prescribed it to be taken at the same time every day, and it is that time now → due to doctor's orders) [PRESENT, ROOT, POSSIBLE] This tag also includes generic state-of-affairs which are hold regardless of time, e.g. John **must** wear a helmet (when riding a motorbike, as the law states) [PRESENT, ROOT, POSSIBLE]
		future	John **must** be home at 8pm (now its 1pm) (because he always comes home at 7pm; based on my inference of John's habits) [FUTURE, EPISTEMIC, POSSIBLE, ASSERTION] John **must** be home at 8pm (now it's 1pm) (because that is his curfew as based on his parents' orders) [FUTURE, ROOT, POSSIBLE] If John marries the tall woman, they **will** have lots of children. [FUTURE, EPISTEMIC, POSSIBLE, CONDITIONAL-THEN]

The tags for **modal flavour** for this study will be broadly construed, with EPISTEMIC and ROOT (or non-epistemic).

- The tag **epistemic** is chosen when the context makes it clear that the modal claim is based on inference or based on the available knowledge. It can be due to the speaker of the sentence (e.g. 'John must be home now' is based on the speaker's knowledge), or someone else (e.g. 'Ben thinks that John must be home now.', is based on Ben's knowledge of the state of affairs, not the speaker).
    - In Indonesian, substitution of the target modal with "pasti" can be used as a strategy to identify **epistemic** modal flavour.
- The tag **root** is chosen when the context makes it clear that the modal claim is NOT epistemic, e.g., not inference/knowledge-based.
    - This can be based on a variety of different types of modality/modal flavours: e.g. based on official rules, regulations or also cultural/societal traditions/rules (e.g., 'obligation'), or based on participant-internal necessity (e.g., *have to sneeze*), or based on goal-oriented necessity (e.g., *there is only*

one highway to Toronto, so John must take that one in order to reach the goal of getting to Toronto).

The tags for **modal force/strength** for this study will have 3 choices: POSSIBILITY, WEAK NECESSITY, NECESSITY

- Modal markers that have a **possibility** claim can conjoin 2 mutually exclusive propositions without it being a contradiction: "*You can stay and you can go*." (Indonesian: *Kamu bisa tinggal dan kamu bisa pergi*) → it is not a contradiction.
- However, modal markers that have a **necessity** or **weak_necessity** claim cannot conjoin 2 mutually exclusive propositions. This will result in a contradiction (indicated as not semantically appropriate/infelicitous with #):
  #*You must stay and you must go*. / # *Kamu perlu tinggal dan kamu perlu pergi.*
  #*You should stay and you should go.*
- Another example is given here to illustrate possibility vs. necessity/weak necessity:
  - Context: *It is raining. I know that Anne doesn't care about the weather; she will go about her business as usual*.
    a. *Anne may be inside and she may be outside*.
    b. #*Anne must be inside and she must be outside*.
    c. #*Anne should be inside and she should be outside* (Vander Klok & Hohaus 2020)
- We can understand that WEAK NECESSITY is different from both NECESSITY and POSSIBILITY modals by using entailment. Thus, in the example below, if it is true that *Jordan must stay inside*, then is it also true that *Jordan should stay inside*, but not vice versa. And similarly, if it is true *that Jordan should stay inside*, then is it also true that *Jordan may stay inside*, but not vice versa.
  - ROOT context: *It is raining cats and dogs. Jordan's mom is worried about him getting sick. She tells her partner*:
    a. *Jordan must stay inside*.
    b. → *Jordan should stay inside*.
    c. → *Jordan may stay inside*.
- The above example is with ROOT modal flavour (based on Jordan's mother's orders/rules/expectations for her child).
- With EPISTEMIC modal flavour, **necessity** can be a claim that one is very sure about, while **weak_necessity** is a claim that one is quite sure, but a bit less sure than 'very certain', and **possibility** is a claim that one has even less knowledge/certainty about.
- English or Japanese translations, which use different lexical words for the two kinds of necessity, are helpful in distinguishing between these:
  - English: *must, have to* (necessity) vs. *should, ought to* (weak necessity) (used across epistemic/root modal flavours)
  - Japanese: ROOT modals: *nakereba naranai* (necessity) vs. *bekida* (weak necessity); EPISTEMIC modals: *nichigainai* (necessity) vs. *nohazuda* (weak necessity).

The tags for **Mood** includes two choices: **counterfactual** or **possible**.

- The tag **counterfactual** is about the event/state-of-affairs NOT happening. In other words, the event/state-of-affairs is known that it will not actually be realized at the time of utterance (in the past, present, or future).
    - This last part concerns the temporal domain:
        - The counterfactuality can be located in the **past**: e.g. *If John had played yesterday, he **should** have passed the ball to Amina (but it is known that he didn't play)*) [PAST COUNTERFACTUAL]
        - in the **present**: e.g., *if John were playing right now, he **should/must** pass the ball to Amina* (but it is known that he is not actually playing now) [PRESENT COUNTERFACTUAL]
        - or in the **future**: e.g. *If John played tomorrow, they **should** win. (but it is known that he will not play because he is out of town*) [FUTURE COUNTERFACTUAL]
- The tag **possible** is about the event/state-of-affairs being possible (opposite of counterfactual, in which it is known that the event/state-of-affairs is false at the time of utterance).
    - In conjunction with **the temporal domain**, and in other words,
        - it is possible that the event/ state-of-affairs has been realized or that there is a past obligation, in this case you use the tag **past** (PAST POSSIBLE),
        - it is possible that the event/ state-of-affairs is being realized (PRESENT POSSIBLE); in this case you use the tag **present**
            - Also use the tag **present** for generic statements, which hold regardless of the time: e.g. John must wear a helmet when riding a motorbike (as the law states).
        - it is possible that the event/state-of-affairs will be realized, in this case you use the tag **future** (FUTURE POSSIBLE)

The tags for **Temporal domain** has three choices: **past**, **present**, or **future**. See the above descriptions for how they are used in conjunction with **mood**.

2.3 Guidelines for POLARITY category

This category has 3 tags: **positive**, **high.neg**, and **low.neg**

- This category is concerned with whether or not the target word appears in a positive context (with no negators → **positive**) or a negative context (→ **high.neg**), or takes scope above a negative proposition (→ **low.neg**). In Indonesian, we are assuming the word order entails this distinction (unlike English); see the examples in the table.
- Negative contexts usually involve a negator such as *tidak/nggak* or *bukan*. Other negators and the like: *tidak*/*tak*, *gak*/*ga*, *kurang*, *bukan*, *jangan*, *belum*, *tanpa*, *jarang(-jarang)*.
- Note also that this category is not about whether the speaker/writer has a negative attitude towards the content described.

Some additional examples:

a. *Tono **harus** makan kue ini.* **positive**
b. *Tono **tidak harus** makan kue ini.* **high.neg** (talking about what's not *harus*)
c. *Semua **harus** mengetahui bahwa Tono kurang sehat hari ini.* **positive** (talking about what's *harus*)
d. *Tono **mesti tidak** suka kue ini.* **low.neg** (talking about what's *mesti*, which is a negative assertion)

Category	Name	Tags	Examples
Polarity	Polarity	**positive**	John **must/should** drink lots of water.
		low.neg (MODAL > NEG)	John **must not** have shellfish (because he is allergic). John **should not** have dairy (because his stomach is sensitive). *Tono **harus tidak** makan* (because he is fasting).
		high.neg (NEG > MODAL)	John **does not have to** remember his score. *Tono **tidak harus** ingat nomer hp* (because it is saved in the mobile phone.)

2.4 Guidelines for GRADABILITY category

The **gradability** category has two choices: **degree** and **nondeg**

- The tag **degree** is about whether the target modal word is in a degree construction, which can include e.g. comparatives or equatives
  - Comparatives occur with *daripada* in Indonesian, such as *Saya lebih suka kue **daripada** (saya suka) sayuran.* If the modal *harus, mesti,* etc. occurs in this type of construction, choose the tag **degree**.
  - Degree questions are with the modal word embedded under ***(Se)bagaimana***
    - **Important! It does **not** have the manner interpretation, but it is about the degree of importance or certainty.**
  - Equatives can be formed with *sama seperti* (e.g., *Saya suka kue **sama seperti** saya suka sayuran*) or with other ways. If the modal *harus, mesti,* etc. occurs in this type of construction, choose the tag **degree**.
- If the target modal word is not in a degree construction, then it is tagged as **nondeg**.

Category	Name	Tags	Examples
Gradability	Degree	**degree** (occurs in a comparative or equative construction)	comparative: I **should** call my doctor more than I **should** call my neighbor for medical advice. (comparing a degree of necessity) degree question: How **important** is it to call my doctor? How **certain** is it that John is home now?

			equative (examples from Portner & Rubinstein 2014) (the degree of necessity is the same to call either her mother or father) a) Susan **must/should** call her mother just as much as she **must/should** call her father. b) It is as **crucial/important** for Mary to call her mother as it is for her to call her father. c) It is as **certain/likely** that Mary will call her mother as it is that she will call her father.
		nondeg	John **must** be home at 6pm. John **should** eat vegetables.