Word Graph Construction on Certain Aspects of Indonesian Language

Sri Nurdiati[1] and Cornelis Hoede[2],

[1] Department of Mathematics, Bogor Agricultural University,
Jln. Meranti, IPB Darmaga, Bogor 16680, Indonesia, nurdiati@ipb.ac.id
[2] Department of Applied Mathematics, Twente University,
PO BOX 217, Enschede, The Netherlands, hoede@math.utwente.nl

Abstract. Knowledge graph theory can be considered to be one of the methods to deal with natural language processing. The theory belongs to the theories about semantic network, but has an advantage that it uses a very restricted ontology. Knowledge graphs have been successfully applied to represent almost any characteristic feature in English and Chinese. Indonesian language, on the other hand, has a very different structure as compared with English and Chinese. In this research we investigate the application of knowledge graphs to represent some characteristic features of Indonesian language. The characteristic features to be considered are active and passive form of verbs and the derived nouns. The result shows that knowledge graph can also represent those features effectively. It can be concluded that knowledge graphs can be used to represent various languages with different characteristics effectively.

Keywords: knowledge graph, natural language processing, semantic network, conceptual structure

1 Introduction

Knowledge graph theory was developed in 1982 at the University of Twente and Groningen. The theory can be considered to belong to the theories about semantic networks. However, the theory is essentially different from other theories, in particular in the fact that a very restricted ontology is used. For graph theoretical terminology we refer to the book of Bondy and Murty [1] or any other of the many books on graph theory. Knowledge graphs consist of a set V of unlabeled vertices, called *tokens* and represented by squares. The knowledge graph is usually a mixed graph with edges and arcs that are labeled and represented by lines respectively arcs. In the theory, so far, 8 types of labels are distinguished. Next to these, 4 types of *frames* are distinguished, the contents of which are knowledge graphs.

For an introduction to the theory, we refer to the theses of Willems [2], van den Berg [3], Liu [4] and Zhang [5]. Most of the background needed can be more easily found in The Proceedings of the International Conference on Conceptual Structures (ICCS) series, see Hoede [6], Hoede and Li [7], Hoede and Liu [8], Hoede and Zhang [9] and Zhang and Hoede [10].

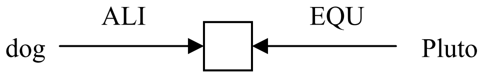
The theory of *conceptual structures* was presented by Sowa [11] in 1984. The two theories are related but essentially different. At the same series of conferences many papers can be found on the theory of *formal concept analysis*, developed by the group of Wille [12]. That theory differs essentially too, also from that of conceptual structures. We will not mention any of the many other theories of semantic networks.

The basic idea of the theory is that in the mind representation of the world is present that has a discrete mathematical nature, so can be modeled by a knowledge graph, that is called *mind graph*. The vertices of this graph correspond to somethings, the *genus* of all *concepts*. "Something" may be a perception unit, then is represented by a single token but, more generally, will be a complex structure of tokens that are linked by links of certain types. So then a subgraph of the mind graph is considered, the elements of which are "taken together" so, literally, form a concept. The first type of frame is used to indicate that the subgraph is seen as a unit. That frame may be called an AND-frame, an idea that goes back to Peirce [13] and his theory of *existential graphs*. Van den Berg [3] has shown that by introducing three other types of frames, the NEG-frame, the POS-frame, and the NEC-frame (for negation, possibility and necessity) various logical system can be represented in the formalism of knowledge graphs.

We will now focus on the 8 types of links, 3 edges and 5 arcs. For the choice of these 8 elements of the 13-elements ontology (1 token, 8 links, 4 frames) we refer to Hoede [14], where it is argued that neural networks in the brain can recognize only a restricted number of types of relationships, namely: EQU, SUB, ALI, DIS, ORD, CAU, PAR, SKO. EQU, ALI and DIS are labels of edges; SUB, ORD, CAU, PAR and SKO are labels of arcs. The relationship between an element of the AND-frame and that frame as a token is said to be of type FPAR, for F(rame) (PAR)t. In the theory there are three *merological* types of relationships: SUB : part of ; PAR : attribute of ; FPAR : property of. So far no words come into consideration. They come in as names of tokens in two ways.

One slogan of the theory is: "FRAMING AND NAMING". Concepts are seen as contents of a frame that is then named, i.e. to which is then attached a word. Note that this procedure is considered to be the same in any language. A second slogan is: "THE STRUCTURE IS THE MEANING". This is an extremely important pillar of the theory. The semantic aspect is equated to the structure of the mind graph. The meaning of a word is the associated structure in the mind of the interpreter of the word. Most linguistic theories try to keep the speaker or listener, i.e. the human, out of the theory.

Attaching a word is represented in two ways by directed links of type ALI and EQU. Note that these links do not connect two tokens but a word with a token. A simple example is :



a knowledge graph that is to be read as “something” of type dog instantiated by “Pluto”.

We have herewith shortly described some background and the formalism. The rest is playing with structures, for which the third slogan holds: "THINKING IS LINKING SOMETHINGS".

2 Word Graph

In knowledge graph theory every word has a corresponding *word graph*, expressing the meaning of the word and therefore called a semantic word graph. Next to that, a word has a certain type, like noun or verb, and in each language ways of linking to other words. In English "the cat", a determiner followed by a noun, is possible. "Cat the" is not a linguistic formation that corresponds to a grammatical rule. The rules of a generative grammar determine for a word type in which way the word can be linked to other words. The arising graph is called *syntactic word graph*; see Zhang [5] or Zhang and Hoede [10].

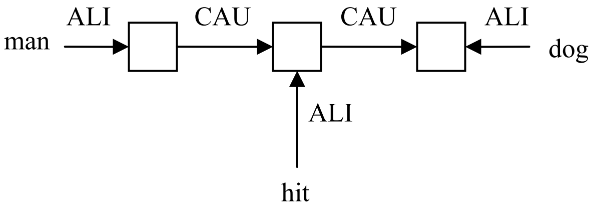
Combining semantic word graphs of words in a sentence leads to a *sentence graph*. The graph representing the combination of sentence graphs of sentences in a text is a *text graph*, expressing the knowledge described by the text.

In most linguistic theories the accent lies on the syntax, the way correct sentences are generated. Such theories are strongly influenced by the theories of *formal (computer) language*. Putting the accent on the semantics from the beginning allows partial sentences to be represented and to be *interpreted*. For interpretation, i.e. giving meaning to, grammatical rules are of minor importance as, up to the simplest single word, meaning is identical with the word graph. Languages may differ strongly in syntactical respect, but follow the same semantical procedure in knowledge graph theory.

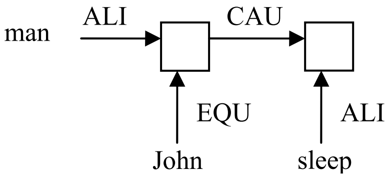
This claim of universality across languages led to the choice of Chinese as object of study, in particular the very specific features of that language, see Liu [4] and Zhang [5]. Knowledge graph theory should be able to incorporate such specific features of a language quite different from e,g. English. The study led to a somewhat different view on word classes. The three papers on word graphs dealt with the following classes. In [7] the classes nouns, verbs and prepositions were discussed. Nouns and verbs are describing somethings that are usually of considerable complexity. Describing their precise meaning leads to complex word graphs. The situation is like for dictionaries. In dictionaries definitions of words are given, but dictionaries differ considerably in:

- The extent to which an explanation is given.
- The definitions themselves.

Two lexicons of word graphs may differ in the same way, in the complexity of the word graph and in the structure of the word graph. Most dictionaries try to keep things simple, to grasp the "essential" meaning of nouns and verbs. These two classes are very closely related. Verbs differ from nouns in that a time aspect is explicitly understood to be present. This close relationship will be subject of discussion in a later section. In the representation, the difference comes forward in the CAU-relationships used to link subject and object to the verb. Consider the sentence graph:



"Hit" is a transitive verb. The sentence "John sleeps" would be represented by:



where now "sleep" is intransitive and only has an incoming CAU-arc and "John" is seen as instantiation of "man" (a dictionary might give "name of a man", when looking up "John"). The prepositions are of a quite different nature. They form the "glue" of a language. This is particularly clear in Japanese. Nouns and verbs are hardly differentiated, but *particles* play a dominant rule, see Hoede (2005). Their word graphs are very small and can be expressed by the links in knowledge graph ontology.

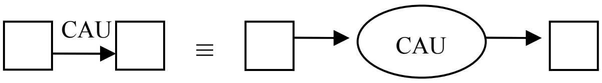
The second set of word graphs of Hoede and Liu (1998) focused on words that attach to nouns and verbs, like adjectives and adverbs, but in particular on *classifier* words, a linguistic feature of Chinese, and somewhat less often used in Indonesian Language. Classifiers have to be expressed to indicate a typical aspect of a concept, see Liu (2002) or Hoede and Liu (1998). In Indonesian Language the word "ekor" means "tail" and is used whenever an animal is mentioned. Other classifiers are "orang" for people and "buah" for things, like in "tiga buah pisang", "three banana(s)".

The third set of word graphs of Hoede and Zhang(2001), focused on those words that express logical aspects in language, see also van den Berg (1993).

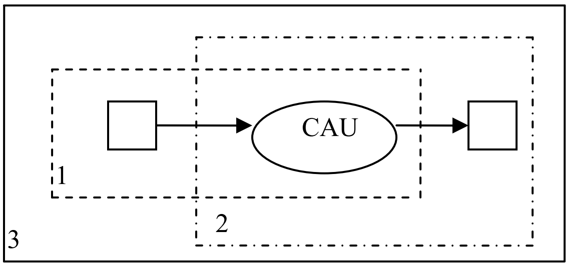
The reader should have enough information now to follow the discussion of some specific features of Indonesian Language, the use of prefixes and suffixes to express the formation of nouns, respectively active and passive forms of verbs.

3 Formation of nouns

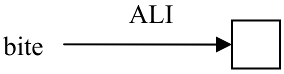
The noun is indeed a basic type of word. The token seen as AND-frame with concept C as content can be interpreted as expressing "C being", so also might be called "BE-frame". As such, framing a part of the mind graph leads to nouns. Let us make this clear by an example. We consider a CAU-arc in total graph form, which means that also the arc is represented by a vertex linked to the incident vertices with auxiliary unlabelled arcs, like in the following figure:



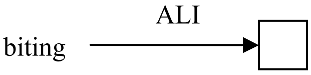
We choose this form to make our procedure clearer with respect to the CAU-arc. We consider three frames:



Frame 1 can be named "cause", frame 2 can be named "effect" and frame 3 can be named "causation". So, all three frames get names that are nouns. If a certain process is framed in English, the process is referred to by the ending -ing. So instead of



we may also describe by



where the word now is a noun. As we remarked before, noun and verb do not differ very much. This also diminishes the difference between adjectives and adverbs. Consider "nice dog", "nice skating", and "skating nicely" as an example why both types of words were collected as "adwords".

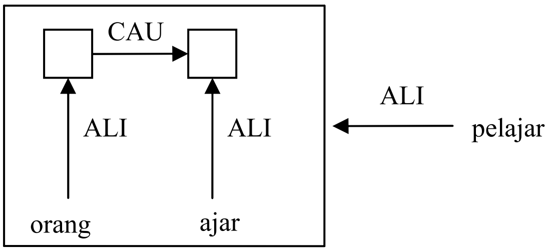
We will now go over to a systematic account of noun formation in Indonesian Language.

3.1 The prefix pe-

Kata jadian and *kata benda* stand for derived word and noun, respectively. One of the important ways to derive a noun is by the prefix pe-. The basic meaning of pe- is "he/she who ... ", where the dots are to be filled in according to the type of word of which the noun is derived.

Derivation from verbs

We consider an example: pe- *ajar*, which changes into *pe(l)ajar*. Ajar is the verb for study(ing), so pe-*ajar* is "he/she who studies/is studying", i.e. a student. The word graph is as follows:

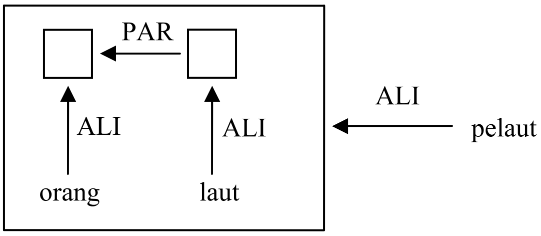


The inclusion of *orang* = man/woman is due to the fact that it is felt as definitely belonging to the concept of student.

Derivation from nouns

This at first seems a somewhat strange situation when we see "he/she who ... " as basic meaning of the prefix pe-. We consider the example pe-*layar* = sailor, where *layar* = sail, but as a noun. It is remarkable that in English sail gets a suffix -or with the same function as pe- in Indonesian Language. The explanation, of course, is that "using a sail" on a boat has been shortened to "sailing", from which a verb "to sail" has developed. The filling in of the dots therefore maybe given as "he/she who uses a sail". The "expansion" of the concept sail is not "brought under words".

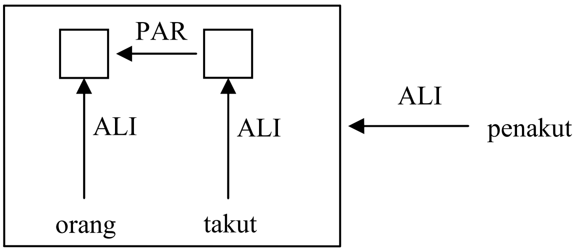
An even more striking example is pe-*laut*, *laut* = sea. The meaning is "seaman". So in English too the combination of two nouns can give a new noun. The derivation in graph theoretical sense would be more complicated than for pe-*layar* and a complicated word graph for pe-*laut* would in principle be needed. However, we can now fill in the dots as "he/she who is associated with the sea" and represent the association with a PAR-link. The graph then becomes:



and for pe-*layar* the same graph with "laut" replaced by "layar" can be given.

Derivation from adjectives

As third major use of pe- we mention the combination with an adjective. Consider "he/ she who is ..." and the possibility of deriving a noun from an adjective is clear. We give only one example. *Takut* = afraid, so pe- *takut* is somebody who is afraid, i.e. a coward. The word graph is



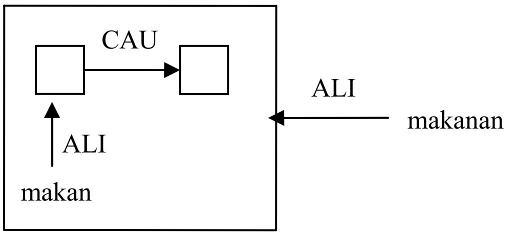
The counterpart of the prefix pe- in English is the suffix -ard. In Dutch *takut* = laf and the word for "coward" is "laf-aard".

3.2 The suffix -an

A rather universal way to derive a noun is by the suffix -an. The meaning can be described by "that what..."

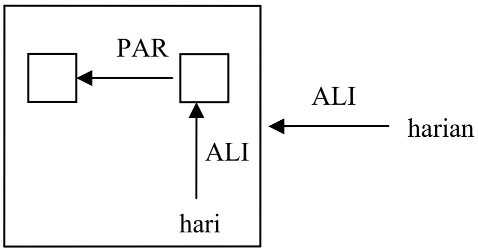
Derivation from verbs

With respect to transitive verbs the suffix-an can be said to have the same function with respect to the patient as the prefix pe- has with respect to the agent. Simple examples are the verbs *makan* = eat and *pikir* = think. *Makan-an* = food, whereas *pikir-an* = "that what is thought", i.e. "thought". The wordgraph for *makan-an* is



Derivation from nouns

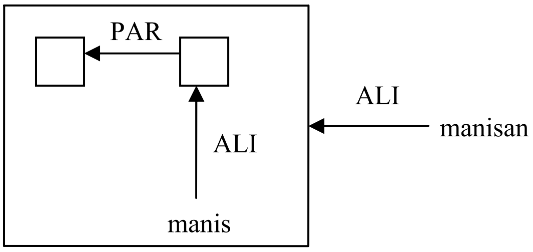
Take, for example, *hari* = day. Adding a suffix -an to this word we get *hari-an* = daily. In this case “hari” is attributed to something, so the word graph is



In Dutch dag = day and blad = journal = majalah. A daily journal is called a "dagblad", so two nouns are simply joint. In English and in Indonesian Language the nouns day respectively *hari* are modified when joint with journal respectively *majalah*; "daily journal" and "majalah hari-an". The word *hari-an* gets the function of an adjective.

Derivation from adjectives

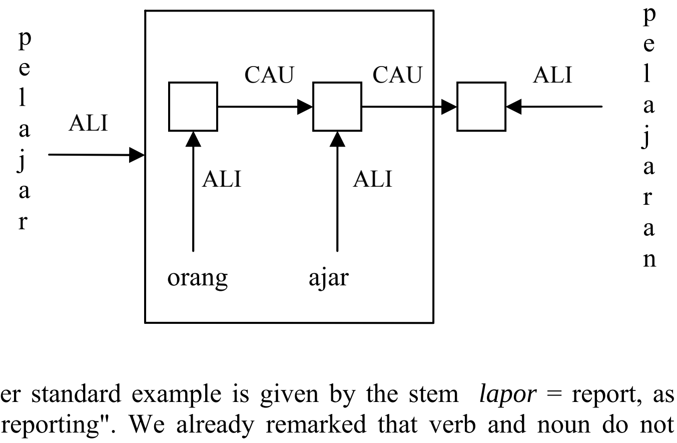
Here the suffix -an creates a noun again. *Manis* = sweet, whereas *manis-an* is "that what is sweet". The word graph is



3.3 Words with prefix pe- and suffix -an

According to the meaning given to pe- and -an as respectively "he/ she who ..." and "that what ...", the combination should express "something, -an, that is associated to somebody, pe-, in relation to a stem word". A nice example is *pe-ajar-an*. The stem is *ajar* = study, *pe-ajar* is, as we have seen before, a student. The something associated to the student, *pe-ajar-an*, is a lesson.

Combining the graph representations we give as word graph for *pe-ajar-an*



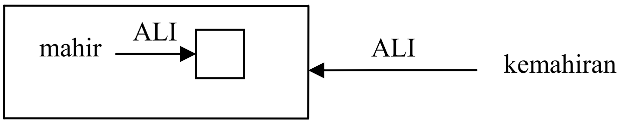
A rather standard example is given by the stem *lapor* = report, as a verb or an activity "reporting". We already remarked that verb and noun do not differ much.

Kerja might also be equated with "working". Pe- *lapor* is a reporter and *pe-lapor-an* is the report as that what is produced. Note that in English report both refers to the activity and to the result of the activity. Indonesian Language makes more distinction here.

3.4 Words with prefix ke- and suffix -an

Consider words in English ending on -ty, like e.g. ability derived from "able". In Indonesian Language, *mahir* = able and the noun *ke-mahir-an* = ability.

The meaning of the combination of prefix ke- and suffix -an can be given as "property of being ...". We therefore use a BE-frame, i.e. an AND-frame without content, and let it contain whatever is filled in for the dots. So the word graph for *ke-mahir-an* is given as



In Dutch often the suffix -heid is used. Bekwaam = able and bekwaam-heid = ability. In German the suffix -keit is often used, like in "ewig-keit" = eterni-ty = *ke-abadi-an*, where *abadi* = eternal. A funny example is given by the stem *ada* = being. *Ke-ada-an* then has the meaning "property of being being" i.e, *ke-ada-an* = situation.

We herewith conclude our account on the formation of nouns by prefixes and/or suffixes.

4 Prefixes of verbs

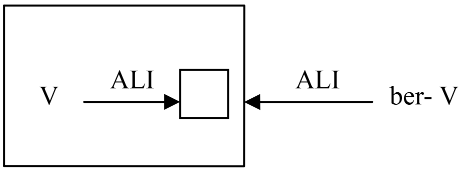
In most languages the verb is the most complicated type of word. Indonesian Language Indonesia is no exception, but is comparatively simple due to a rather systematic use of prefixes. Like for noun formation various stems can be chosen in verb formation. Next to that, prefixes are used to express active and passive form. We will discuss the prefixes ber-, me- and di-, again without discussing the morphological issues of spelling changes. These will come forward in the word graphs given.

4.1. The prefix ber-

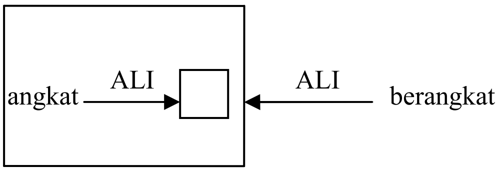
A verb formed with ber- is an active form that does not have a patient and indicates the situation in which the agent is.

Formation from a verb

The verb ber-V has the extra meaning "being in the process of V-ing". The word graph is therefore given as

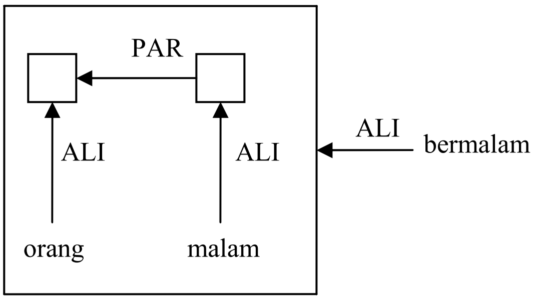


An example of the change in meaning induced by ber- is given by *angkat* = lift. *Angkat besi* = lift iron, i.e. weight lifting. *Ber-angkat* is an active form without patient and describing a process. The meaning is "leave". “Bus ber-angkat” says that "the bus is in the process of lifting", lifting itself so to say. The word graph is



Formation from a noun

Like for the formation of a noun from a noun, recall *pe-laut* = seaman, the derivation from the basic meaning can be rather complicated. A good example is given by the stem *malam* = night. The verb *ber-malam* describes an activity associated with "night". The English description is "stay overnight". As word graph we give



Other formations

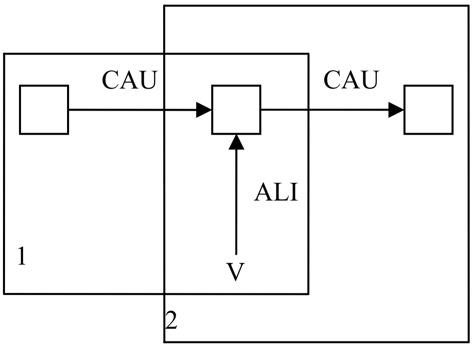
There are interesting other formations with ber-. It can be combined with an adjective like in *ber-gembira*, where *gembira* = happy, and the verb expresses "being happy". In *ber-dua* = with two, we see a combination with a number word.

Ber- with a doubling of the stem adds another piece of meaning. *Ber-dua-dua* = in groups of two, *ber-puluh-puluh* = in groups of ten. Doubling can also be used for expressing intensity. *Tahun* = year and *ber-tahun-tahun* is best translated as "year in year out". We only want to stress here that ber- clearly adds that the situation is considered.

4.2 Formation with me- or di-

The prefix me- knows many morphological changes that we will not discuss. It gives a verb in active form in which the focus is on the action. Di-, on the other hand, gives a verb used in a passive sentence where the focus is on the patient.

The main difficulty we meet is the representation of the focus. Let us consider verb V and the knowledge graph



The frames 1 and 2 focus on the agent of V respectively the patient of V. However, we already met these frames in the formation of nouns by the prefix pe- and the suffix -an.

We will now give some examples without giving word graphs as the structure of these should be clear.

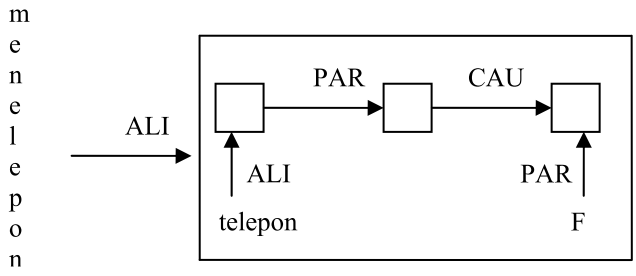
Formation with me- and verb

For this formation the verb may have a patient or not. *Beli* = buy and *me-beli* = buy + extra meaning as described before. There usually is a patient, the verb is transitive. *Desah* = sigh and *me-desah* = sigh + extra meaning as described before. There is no patient, the verb is intransitive.

There are several basic verbs where me- might not be used as a prefix. *Tidur* = sleep is one such verb, *makan* = eat, *minum* = drink or *ingin* = want, are others.

Formation with me- and other word types

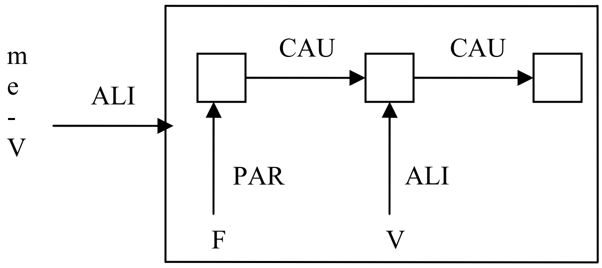
Me- plus noun is perhaps clearest in *me-telepon* = phone (use the telephone). The word graph is



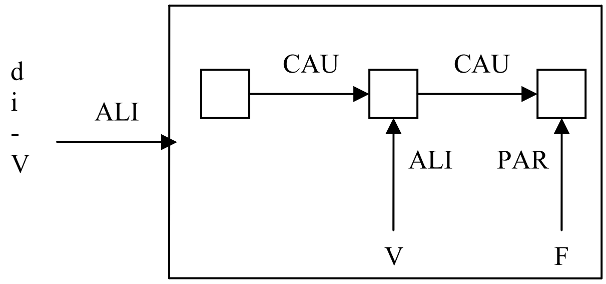
Here, as an alternative for using total graphs, we introduce an extra element in the ontology of knowledge graphs to express focus. The symbol F, like the types of the links to be seen on the meta-level and not on the word-level, might be attributed to a token. The explicit connotation of focus when using me- and di-, as well as intonation, so far not been dealt with in the theory, seems to force the introduction of a link between the symbol F and a token. As focus is typically attributed by the speaker, the presentation chosen is by a PAR-link, like in the word graph given above.

The link between symbol and token may therefore in our theory now be an ALI-arc, for typing, an EQU-arc, for instantiation or a PAR-arc, for focusing, seen as a verbal expression.

For me- plus verb and di- plus verb we would now give

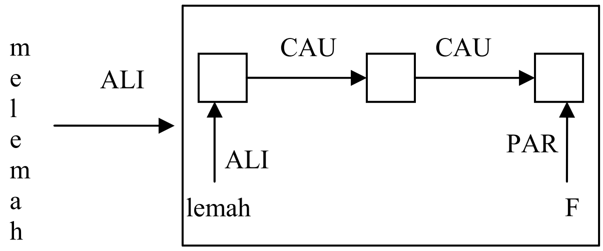


respectively



This gives a clear simplification of the formalism, at the cost of adding a fourteenth element to the ontology.

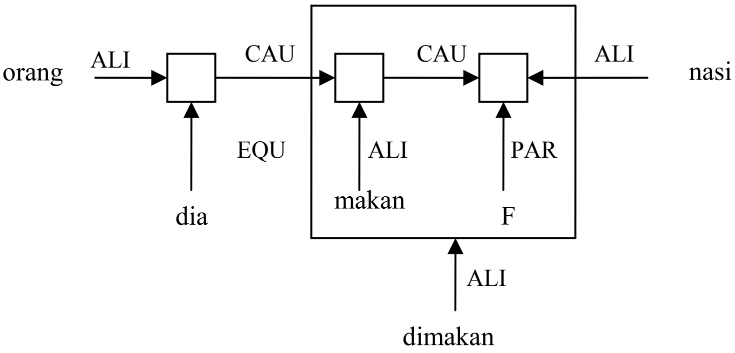
Me- plus adjective is exemplified by *lemah* = weak and *me-lemah* = weaken, an intransitive verb, with, now, word graph



4.3 The prefix di-

We can be rather short now about the prefix di- that is used to express a passive sentence about a patient on which the focus lies.

Consider the sentence: *Dia makan nasi* = he eats rice. The passive form reads *nasi dimakan dia*. The sentence graph is simply



focus on "dia" would lead to the active form and one would expect the prefix me-. However, in combination with "makan" this is not used, as we remarked before.

Note that the focus also has influence on the utterance path. In the passive form the sentence has to start with "nasi" and instead of "makan" "dimakan" is used. Also note that "by", the first CAU-arc from "dia", is not uttered, although sometimes *oleh* = by is mentioned. The implication by di- is very strong. "dijual" just means "sold", a rather normal way of speech, like "deal!", where the exclamation mark should be noted. Instead of F we might have used the symbol "!". In a similar way, the symbol "?" might be attached by a PAR-arc to a knowledge graph frame. In Chinese this is actually expressed in language by the word "ma".

5 Conclusion

Different to that of English and Chinese, the formation of Indonesian word often involves the use of prefix or suffix to some basic words. However, knowledge graph can still represent the formation of Indonesian word perfectly. Hence we conclude that knowledge graphs can be used to represent various characteristic features of many different languages effectively.

References

1. Bondy J. A. and Murty, U. S. R.: *Graph Theory with Applications*, The McMillan Press, London and Basingstoke, SBN 333-17791-6 (1976)
2. Willem, M.: *Chemistry of Language*; PhD Thesis, University of Twente, Enschede, The Netherlands, ISBN 90-9005672-6 (1993)
3. Berg, H. van den: *Knowledge Graphs and Logic: One of Two Kinds*, PhD Thesis University of Twente, Enschede, The Netherlands, ISBN 90-3651835-0 (1993)
4. Liu, X.: *The Chemistry of Chinese Language*, PhD Thesis, University of Twente, Enschede, The Netherlands, ISBN 90-3651834-2 (2002)
5. Zhang, L.: *Knowledge Graph Theory and Structural Parsing*, PhD Thesis, University of Twente, Enschede, The Netherlands, ISBN 90-3651835-0 (2002)
6. Hoede, C.: On the ontology of Knowledge Graphs, in: *Conceptual Structures: Application, Implementation and Theory*, (G. Ellis, R. Levinson, W. Rich and J. Sowa, eds.) Springer Lecture Notes in Artificial Intelligence 954, 308-322 (1995)
7. Hoede C. and Li, X.: Word Graphs : The First Set, in: *Conceptual Structures. Knowledge Representation as Interlingua*, Aux. Proc. of the Fourth International Conference on Conceptual Structures, (P.W. Eklund, G. Ellis and G. Mann, eds.), Bondi Beach, Sydney, Australia, 81-93 (1996)
8. Hoede, C. and Liu, X.: Word Graphs: The Second Set, in: *Conceptual Structures: Theory, Tools and Application*, Proceedings of the 6th. International Conference on Conceptual Structures, Montpellier, ICCS'98 (M.-L. Mugnier, M. Chein, eds.) Springer Lecture Notes in Artificial Intelligence 1453, 375-389 (1998)
9. Hoede C. and Zhang, L.: Word Graphs : The Third Set, in: Conceptual Structures: *Broadening the Base*, Proc. of the 9th International Conference on Conceptual Structures, (H.S. Delugach and G. Stumme, eds.), CA, USA, Lecture Notes in Artificial Intelligence Nr. 2010, 15-28 (2001)
10. Zhang, L. and Hoede, C.: Utterance Paths, in: *Using Conceptual Structures*, Contributions to ICCS 2003 (B. Ganter and A. de Moor, eds.), 15-28 (2003)
11. Sowa, J.F.: *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley, Reading (1984)
12. Ganter, B. and Wille, R.: *Formal Concept Analysis: Mathematical Foundations*, Springer Verlag New York, Inc. Secaucus, NJ, USA (1997)
13. Peirce, C.S.: On the algebra of logic, *American Journal of Mathematics*, 7 (1885)
14. Hoede, C.: On the Ontology of Knowledge Graphs, in: *Conceptual Structures: Applications, Implementation and Theory*, Springer Lecture Note in Artificial Intelligence 954, 308-322 (1995)
15. Hoede, C.: *Knowledge Graph Analysis of Particles in Japanese*, Memorandum 1746, University of Twente, Enschede, The Netherlands (2005)