Memorandum 1871 (June 2008). ISSN 1874-4850. Available from: http://www.math.utwente.nl/publications
Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

On Word Graphs and Structural Parsing

Cornelis Hoede[1] and Sri Nurdiati [2]

Abstract

In this paper we discuss the formation of a lexicon in knowledge graph theory and the way sentences can be mapped on sentence graphs using such a lexicon. That procedure is called structural parsing and the use of chunks in that procedure will be investigated for example sentences in Bahasa Indonesia.

Keywords: knowledge graph, word graph, sentence graph, lexicon, structural parsing.

1. Introduction

For most of the background of knowledge graph theory, we refer to our recent paper on specific features of Indonesian language [1]. In this paper, we first recall the knowledge integration and structuring system (KISS), developed about 25 years ago.

1. Given a set of texts, of different authors, we can first represent these *n* texts by so-called *author graphs* $AG_1, AG_2, \ldots, AG_n$. We will discuss how this is done in a later part of this paper. This is the phase of *text analysis.*

2. Having obtained the author graphs we may now combine these graphs into one *compiled graph* by identifying vertices and links. This already forms a first representation of the knowledge in the *n* texts, that are supposed to be about the same subject.

3. The compiled graph can now be subject to the procedures *construct analysis* and *link integration* in any order and repeatedly. These two procedures form the central part of KISS. We will discuss them in somewhat more detail.

    a. Construct analysis:

    In the compiled graph the first obvious thing usually is the occurrence of synonyms or concepts that are very similar. This will become clear by looking at the direct neighborhoods of vertices. The associations made with other concepts by two concepts may show high similarity, for which several measures are known. This allows an automatic identification of similar concepts. Once such similar concepts are found they may be identified, which simplifies the compiled graph.

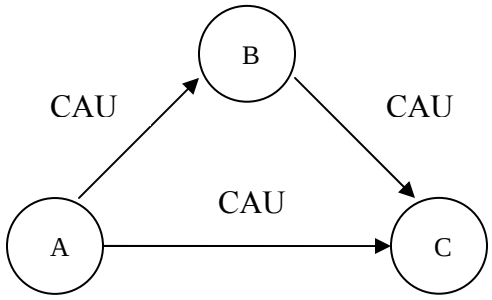
[1] Department of Applied Mathematics, University of Twente, Enschede, The Netherlands
[2] Department of Mathematics, Institut Pertanian Bogor, Bogor, Indonesia

A more advanced variation on this theme is to look at certain subgraphs of the compiled graphs, which may also consist of a set of unconnected vertices, and discover that the subgraph acts more or less as a unit. This is what construct analysis is about. The subgraph can be replaced by one vertex, possibly with a new name. We refer to Bakker [2] for more technical details. A warning must be given against homonyms. Two vertices might represent concepts with the same name, but have different meaning, the difference being in the local structure of the neighbors of the two vertices.

b. Link integration:

This procedure derives knowledge from the knowledge represented by the compiled graps or the graph resulting from the compiled graph after construct analysis. We give as example the concatenation of two causal links. If $A \xrightarrow{CAU} B$ and $B \xrightarrow{CAU} C$, then a "multiplication" rule might be chosen that implies $A \xrightarrow{CAU} C$. A rule for deriving knowledge in this way may be given for combination of all 8 types of links of the ontology of knowledge graphs. It may also occur that between two vertices two or more types of links occur. We then need "summation" rules for choosing one, final, type of link between the two vertices.

Multiplication rules and summation rules form a *path algebra* on the set of links. The derived links will, by the multiplication, generate more links and the graph may not be easily interpreted anymore. It is therefore good to note that links implied by two other links may also be deleted. An example: if three authors generate author graphs containing $A \xrightarrow{CAU} B$, $B \xrightarrow{CAU} C$ and $A \xrightarrow{CAU} C$ separately, in the compiled graph the subgraph



will be formed, from which the link $A \xrightarrow{CAU} C$ may be deleted, because implied by the other two links and the path algebra.

4. After repeated application of these central procedures we arrive at an *integrated graph*, giving a qualitative structural model of the knowledge contained in the texts. Usually, variables will occur, that may, for a certain object, have specific values. Such a graph is called *object graph*. As an example, that has been realized in the project MEDES (Medical Expert System) [3], the texts may contain

medical knowledge and the integrated graph is then a, huge, qualitative, model of the human body. The values of variables, like body temperatures, will vary from person to person. The person in this case is the "object".

2. Application of KISS to definitions.

In this section we will illustrate the system by considering 18 definitions of the concept "Tsunami". A lexicon of word graphs should contain one word graph for the word tsunami. As we have 18 at our disposal we must analyze them in order to come to one definition. The starting point consists of 18 texts, namely these definitions. After text analysis we get, small, author graphs, that might also be called *definition graphs*. The analysis will not use the full ontology of knowledge graph theory, as we only want to make some of the features of KISS clear.

The 18 definitions were taken from various sources, mainly from internet. No references to these sources will be made. The 18 "definitions" read:

1. a cataclysm resulting from a destructive sea wave caused by an earthquake or volcanic eruption.
2. a large, often destructive, sea wave, usually caused by an earthquake under the sea.
3. a very large ocean wave caused by an underwater earthquake or volcanic eruption.
4. a great sea wave produced by a submarine earthquake, volcanic eruption, or large landslide.
5. a sea wave of local or distant origin that results from large-scale seafloor displacements associated with large earthquakes, major submarine slides, or exploding volcanic islands.
6. one or a series of huge sea waves caused by earthquakes or other large-scale disturbance of the ocean floor.
7. a sea wave produced by any large-scale disturbance of the sea floor, principally by a submarine earthquake or by submarine earth movement.
8. vast sea waves caused by the sudden dropping or rising of a section of the sea floor following an earthquake.
9. large seismic waves produced by earthquakes, volcanic eruptions or underwater landslides.
10. a long-period gravity wave generated by a submarine earthquake or volcanic event.
11. a seismic sea wave that is potentially the most catastrophic of all ocean waves.
12. wave produced in a large body of water (ocean or lake) by an earthquake, volcanic eruption, or underwater landslide.
13. a wave generated by seismic activity.
14. a seismic sea wave usually caused by a submarine earthquake.
15. a natural phenomenon consisting of a series of waves generated when water in a lake or the sea is rapidly displaced on a massive scale.
16. caused by underwater earthquake or landslide, can rise to great heights and cause catastrophic damage near coasts.
17. a large and often destructive wave caused by a submarine earthquake.
18. a giant deadly wave caused by an earthquake.

The first remark to make is that definition 1 is the only one not focusing on the wave, but on the "cataclysm" caused by the wave. Number 16 is no definition at all, although it is clear what is meant. For this reason we skipped definitions 1 and 16.

2.1 Constructing the graphs

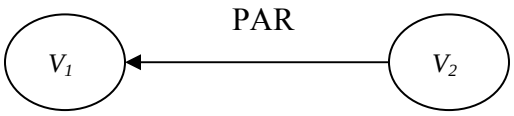
According to KISS we should now construct 16 definition graphs by structural parsing. We will illustrate that procedure later in some detail. We choose to apply construct analysis first, in its simplest form, by identifying synonyms or very similar concepts.

We made a list of nouns, 35, occurring in the 16 definitions. Also adjectives derived from a noun, like "seismic" derived from the Greek "seism(os)"= shake, were taken into account. A similar adjective is "marine", derived from "mare"= sea. There were 16 nouns occurring in more than one definition. Most frequently mentioned were the concepts : wave, earth, quake and sea, not particularly astonishing. We now determined sets of similar concepts like {sea, ocean, water, mare, lake} or {earth, land, floor}, {quake, seism} and {slide, disturbance} to obtain the following 9 concepts to focus upon and to be represented by labeled vertices:

v_1 = quake	v_6 = sea
v_2 = earth	v_7 = slide
v_3 = wave	v_8 = series
v_4 = volcano	v_9 = event
v_5 = eruption	

Replacing concepts by the one chosen as the representative of the set of similar concepts to which they belong, we could now construct definition graphs using only these 9 vertices. It turns out that only 9 different graphs remained after removal of those definition graphs that were contained in other definition graphs as a subgraph. In such a case, the information of the subgraph, the definition represented by it, is just part of the information of the definition represented by the larger graph.

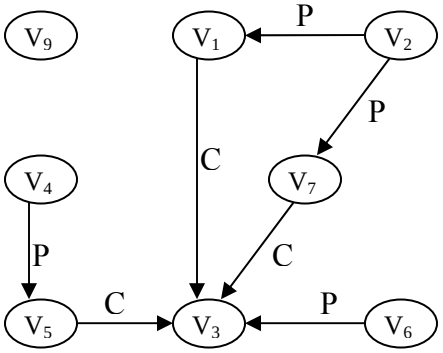
Before giving the 9 definition graphs we should mention that only three types of links between tokens are used: CAU, ORD and PAR, the latter in such combinations of two nouns like "earth quake", where "quake" is seen as the central concept and "earth" as attributed to (associated with) it. So the representation is



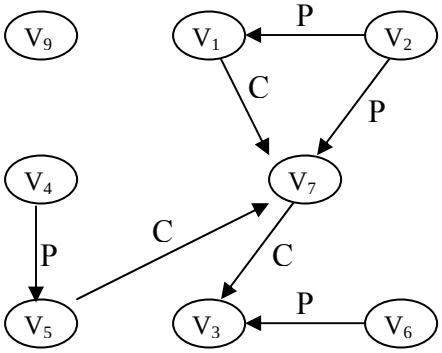
The ORD-link only occurs where locations are compared, like when "submarine" is to be represented. The results of the analysis are the graphs G_4, G_5, G_7, G_8, G_9, G_{10}, G_{13}, G_{14} and G_{15}, where the indices correspond to the numbers of the definitions.

The resulting graphs are drawn without v_8 : series, but always containing the other 8 vertices. In this way the differences, in the links, become clear. The graphs are:

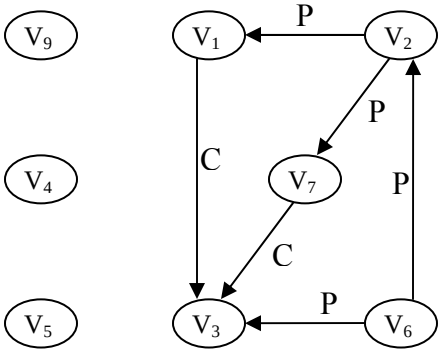
G_4:



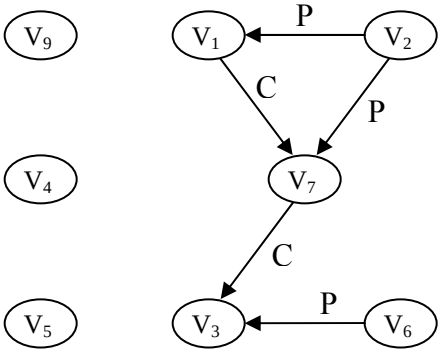
G_5:



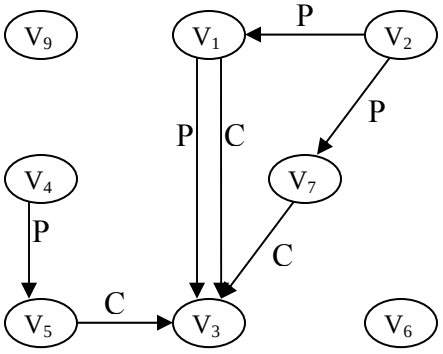
G_7:



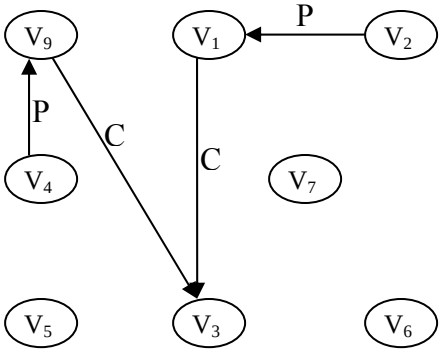
G_8:

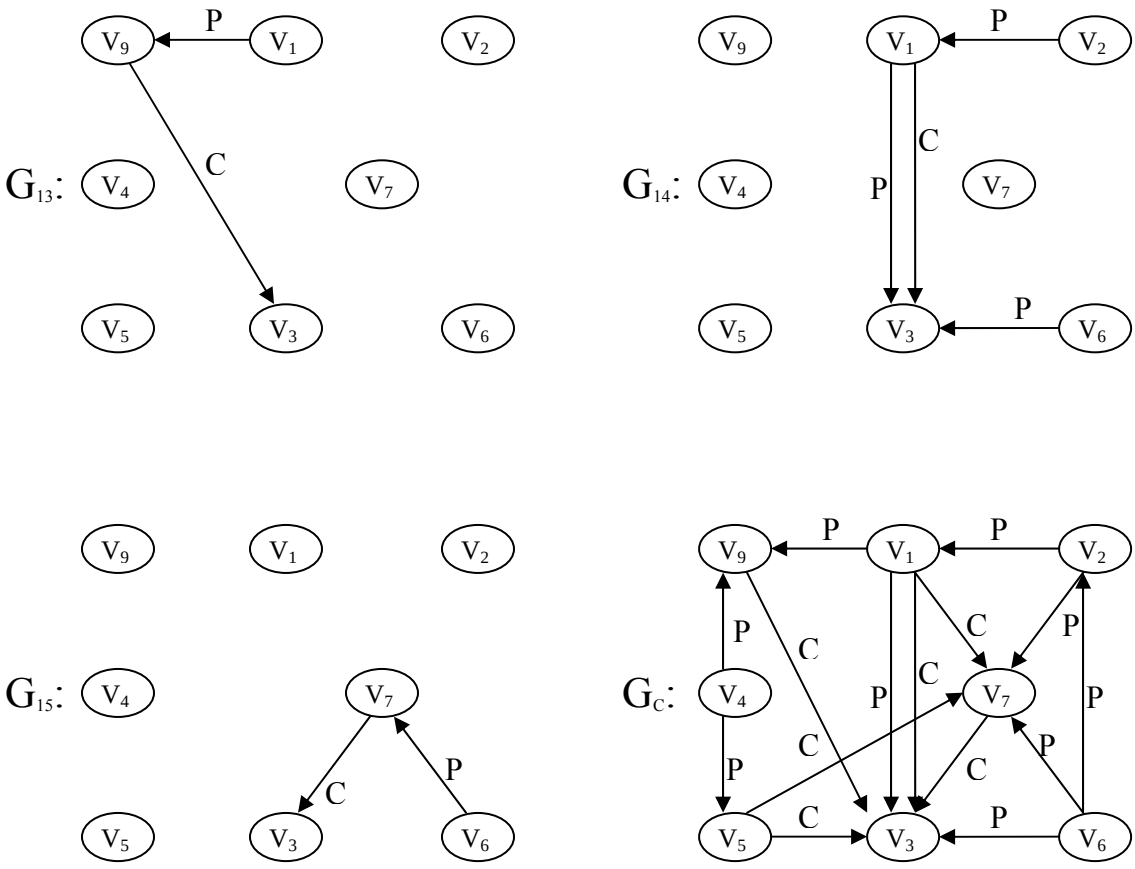


G_9:



G_{10}:





We have left out the tokens referring to locations and the ORD-links between them. G_c is the combination of the 9 definition graphs, so what we call the compiled graph, that is the starting point for further application of KISS.

The reader should remark the big differences in complexity of the definition graphs.

2.2 Link integration

The graph G_c is now used to illustrate the procedure link integration, involving the path algebra. There are only two types of links, CAU (C) and PAR (P) in G_c, not 8 like in the full ontology. However, these two suffice to explain the path algebra. We consider the following two tables,

+	P	C
P	P	C
C	C	C

X	P	C
P	{ }	{ }
C	C	C

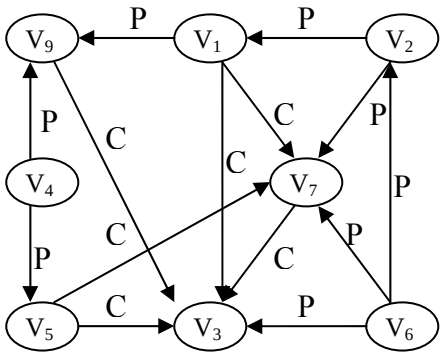
The "addition" table states that whenever two links, possibly identical, are present they can be replaced by one according to the table. Here it is assumed that the links are oriented in the same way, otherwise both are left unchanged.

In G_c we see a CAU-link and PAR-link in parallel from v_1 to v_3, so we replace this pair by a CAU-link from v_1 to v_3.

The "multiplication" table states that a PAR-link followed by another PAR-link or a CAU-link does not imply an extra link (the symbol { } was used to indicate this). A CAU-link followed by another CAU-link or a PAR-link implies an extra CAU-link. Note that another path algebra might have been chosen by the reader. That depends on the view on deriving knowledge.

No extra CAU-links are generated by the path algebra. We obtain an integrated graph

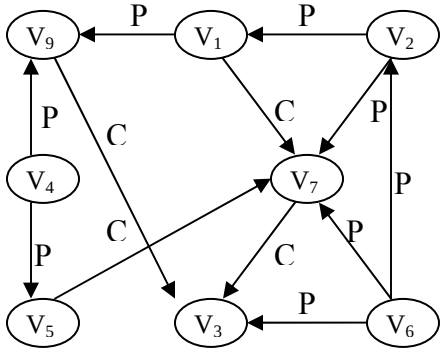
IG_1:



.

A particularly interesting aspect of the path algebra is the possibility to reduce the graph. A link that can be generated by two other links via the path algebra can be left out. In particular in the example of CAU-links implying a third CAU-link, that third CAU-link can be left out. This not only simplifies that graph but may also reveal an interesting aspect of the qualitative model. In IG_1 the described situation occurs for the vertex triple $\{v_1, v_7, v_3\}$, where (v_1, v_3) is the CAU-link that can be left out, and for the vertex triple $\{v_5, v_7, v_3\}$, where (v_5, v_3) can be removed. The integrated graph IG_1 is then reduced to

IG_2:



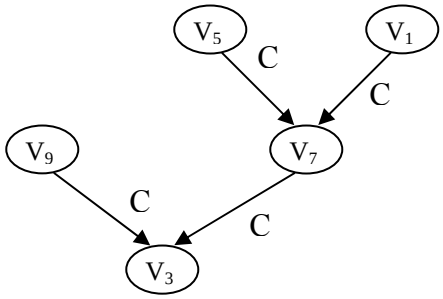
,

in which the direct cause of the "wave"= v_3 is "event"= v_9 or, more importantly, "slide"= v_7, the slide then due to a "quake"= v_1 or an "eruption"= v_5. The removed causations occur in G_4, G_7 and G_9, where the authors did not stress the important point that it is the slide or displacement of earth mass and consequent displacement of water mass that causes the wave. G_5 may be considered the most accurate of the definition graphs.

2.3 Construct analysis

We might consider IG_2 to be the final result of application of KISS. However, the procedure construct analysis may be applied too, now with respect to structural similarity, not just similarity of concepts, where we started with.

Let us focus on the CAU-links in IG_2 , that involve the vertices G_1, G_3 G_5, G_7 and G_9, or the subgraph induced by these CAU-links:



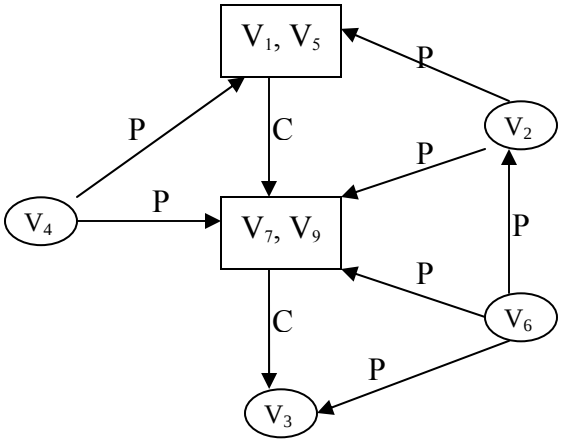
.

We see that v_1 and v_5 have a similar position. Both have an outgoing arc to v_7. v_7 and v_9, on their turn have a similar position with respect to V_3. We may therefore consider the constructs $\boxed{v_1,\ v_5}$ and $\boxed{v_7,\ v_9}$. (In computerized version KISS may produce such constructs and the user may decide to simplify or not, i.e. introduce the construct as unit or not).

$\boxed{v_1,\ v_5}$ takes together "quake" and "eruption", which makes sense. The construct might be described as "quake eruption". $\boxed{v_7,\ v_9}$ takes together "slide" and "event", which also makes sense, a slide being an event. Event is a genus of "event", as of "quake" and "eruption". In this case we might describe by the more differentiated concept, which is "slide".

Acceptance of the constructs now leads to the graph:

IG3:



The PAR-link from v_6 and v_2 stems from the word "sea floor" and therefore contains information of geometrical nature. It is a good example how accurate the analysis should really be, which boils down to using as sophisticated word graphs as possible.

The integrated graph, version IG_3, is rather satisfactory, the central process, the causation network, comes forward clearly. v_2, v_4 and v_6 are of secondary importance and indicate which other concepts are involved.

3. Structural parsing

We have shown how the system KISS may determine word graphs for a concept given a set of definitions that have been transformed into/ mapped on definition graphs. This latter process is called structural parsing. We will discuss this process, for which we give Zhang [4] as reference, for two sentences in Bahasa Indonesia, that also form a definition of the concept tsunami. The sentences are:

a. Tsunami adalah gelombang yang dipicu (oleh) kejadian alam.
b. Kejadian alam tersebut gempa di lepas pantai, letusan gunung di dekat permukaan laut, pergeseran lapisan di bawah laut atau hantaman meteor yang menimpa laut.

3.1. Chunking

A preprocessing of the text is by determining "chunks". It was remarked by Abney [5] that sentences are pronounced in pieces that he called chunks. Speakers tend to indicate separations between chunks by pauses in their speech. Zhang mentions the followings linguistic elements to look for:

1. Pairs of comma's, comma and period and other symbols indicating that the sentence is broken into pieces.
2. Auxiliary verbs, after which usually a subsentence is following that acts as a unit.
3. Reference words that separate what is referred to from the referring text.

4. “Jumps”; these occur in text when, for syntactic reasons, two consecutive words can not be linked. It is for this reason too that next to semantic word graphs, syntactic word graphs were introduced. As a simple example we consider “the cat” and “cat the”. In the second case, noun followed by determiner, a “jump” occurs.
5. Prepositions, that typically link other pieces of text.

We have added
6. Logical words like “and” or “or”, for the same reason as mentioned under 5.

Our first goal is to see whether these indicators for chunking also work in Bahasa. We will go back to English later on.

The numbers above words indicate why chunk separators were given on both sides of them, above signs in case these separate by their very meaning, we find:

a. Tsunami | adalah ^{3} | gelombang | yang ^{3} | dipicu | (oleh) ^{5} | kejadian alam.

b. Kejadian alam | tersebut ^{3} | gempa | di ^{5} | lepas ^{5} | pantai, | letusan gunung | di ^{5} | dekat ^{5} | permukaan laut,^{1} | pergeseran lapisan | di ^{5} | bawah ^{5} | laut | atau ^{6} | hantaman meteor | yang ^{3} | menimpa laut.

This chunking by indicators corresponded very well with the chunking by speech.

The next step in structural parsing would now be to combine the word graphs of the words occurring in a chunk, as available from a word graph lexicon, into chunk graphs. Combining these chunk graphs gives the two sentence graphs, from which the text graph can be constructed in the same way as the definition graphs were pasted together.

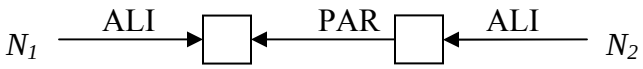
3.2. The lexicon

We will give a very simple set of word graphs. “Tsunami” is the concept to be defined. We will first discuss the nouns. They are simply, with translation, given by word graphs

gelombang	$\equiv$ wave $\xrightarrow{ALI}$ □	kejadian	$\equiv$ event $\xrightarrow{ALI}$ □
alam	$\equiv$ nature $\xrightarrow{ALI}$ □	gempa	$\equiv$ quake $\xrightarrow{ALI}$ □
pantai	$\equiv$ coast $\xrightarrow{ALI}$ □	letusan	$\equiv$ eruption $\xrightarrow{ALI}$ □
gunung	$\equiv$ mount $\xrightarrow{ALI}$ □	permukaan	$\equiv$ surface $\xrightarrow{ALI}$ □
laut	$\equiv$ sea $\xrightarrow{ALI}$ □	lapisan	$\equiv$ layer $\xrightarrow{ALI}$ □
pergeseran	$\equiv$ slide $\xrightarrow{ALI}$ □	hantaman	$\equiv$ hit $\xrightarrow{ALI}$ □
meteor	$\equiv$ meteor $\xrightarrow{ALI}$ □	menimpa	$\equiv$ fall $\xrightarrow{ALI}$ □ .

Each of these concepts might, of course, be given a much more elaborate word graph by describing the inner contents of the tokens □.

As the combination of two nouns: N_1 N_2 in Bahasa Indonesia means that N_1 is the focal noun, the representation is

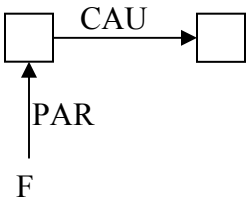


.

Herewith already several chunk graphs have been determined.

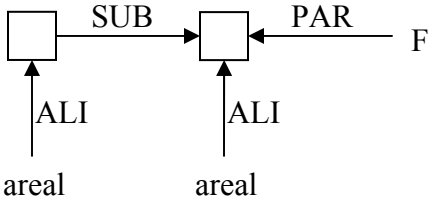
The structure of a knowledge graph is to a great extent determined by the prepositions, that form the "glue" of language. We have here "di" ≡ at, "oleh" ≡ by, "lepas" ≡ off, "dekat" ≡ close to, and "bawah" ≡ under.

The easiest word is "oleh" for which we can give the CAU-arc



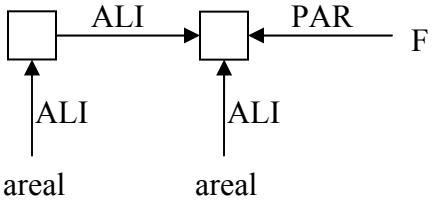
,

with focus on the token with which the arc is incident out. The other three need a more elaborate discussion. "di" can be described as "having location/ area that belongs to the location/ area on which the focus lies". Hence the word graph:



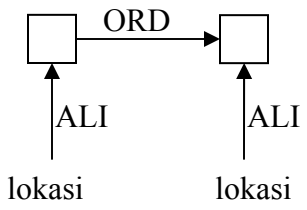
,

"dekat" involves a distance consideration. A simple version of the word graph is



.

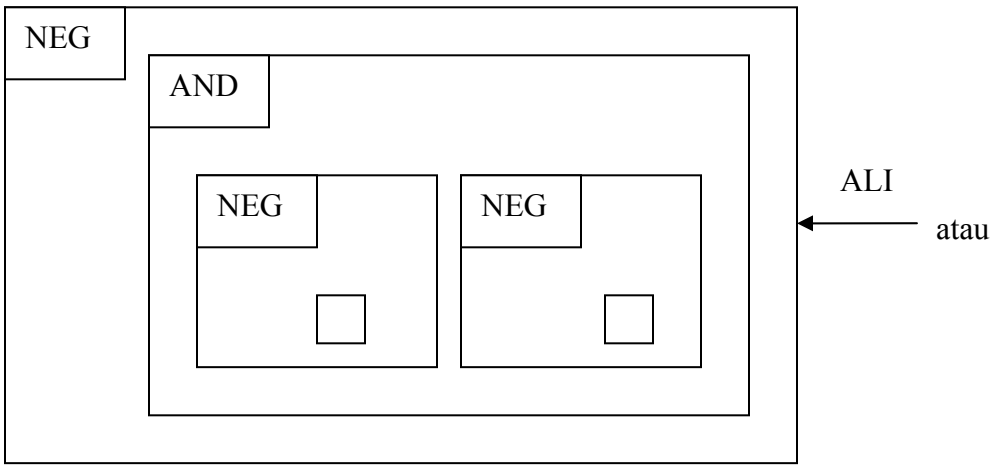
"lepas" and "bawah" involve geometrical information and imply reference points and coordinate systems. A precise description would lead to rather large word graphs in which the following kernel is present in both cases.



For “lepas” the focus would be on the first token, for “bawah” on the second. The location should be described as being considered with respect to a direction.

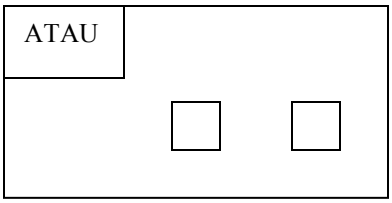
Prepositions are usually very short words, because used very often, but carry a large semantic load that can only be expressed by a complicated word graph, especially when space and time are involved.

“atau” ≡ or is a logical word. In logic the proposition $p \vee q$, *p OR q,* is logically equivalent to $\neg(\neg p \wedge \neg q)$, NOT (NOT p AND NOT q). As the word graph ontology contains an AND-frame and a NEG-frame, for negation, the word graph is in principle:



This, rather clumsy, structure is given for “atau” to show how logic can be incorporated in the knowledge graph formalism.

A more practical description would be to use the word “atau” itself, as in



Here we let the word refer to the more complicated word graph that should be given in principle.

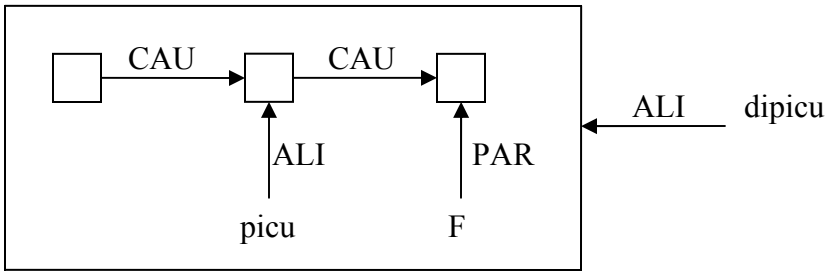
We are left with the reference words “adalah” ≡ is, “yang” ≡ that and “tersebut” ≡ mentioned, that are used very often. “adalah” is used when a definition is given. “tsunami adalah …” means “a tsunami is …”. The word graph that corresponds to it is the directed ALI-link that is used for typing a concept C, so

is to be read as “C adalah …” where the dots correspond to the inner structure of the token.

The word “yang” has the word graph

whereas “tersebut” refers to examples, so is related to , and can be expressed in “atau”. Tersebut may occur in combination with atau, in which case we only represent the word atau. Atau has a set character. Element-set relationship is expressed by a PAR-relationship.

Finally we give the word graph for the verb form di-picu (oleh) ≡ triggered (by). The use of this prefix was discussed in [1]. The basic verb is picu ≡ trigger. The prefix di- expresses that the passive form is considered. Consider the graph



The focus label is used to stress the passive form. The word oleh ≡ by is represented by the incoming CAU-arc for picu and often not stated in Bahasa.

3.2. The text graph

For constructing the knowledge graph of the text in the two sentences a) and b) it is now just a matter of superposing the chunk graphs in the right way. Usually the word order, so the syntactic aspect, suffices to do this.

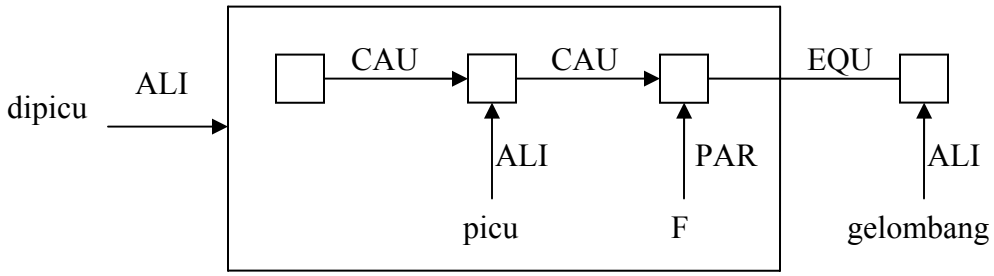
The two word chunks have the following graphs:

,

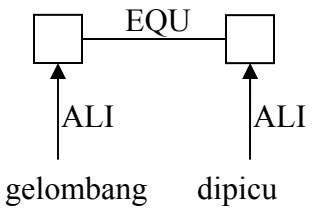
,

, .

We now leave out of consideration, for the time being, alam, gunung, laut, lapisan, meteor. We focus on the reference words and first fill in the concepts related by them.

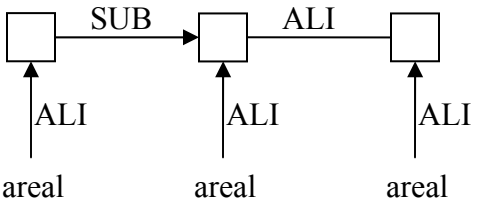


is the graph for the three consecutive chunks “gelombang | yang [3] | dipicu [5] |”. The token unspecified in “di-picu” is, of course, that of kejadian. Sentence b) specifies kejadian (= event) further. First we resolve “yang” as

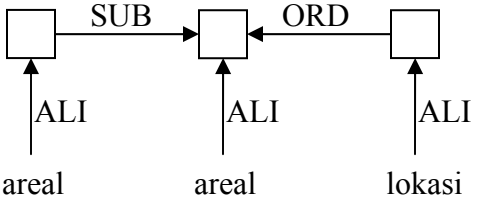


.

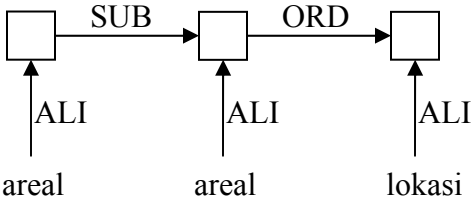
The prepositions come in pairs. All three pairs indicate relationships between locations or areas. We have, given the lexicon graphs,



for | di [5] | dekat [5] |,



for | di [5] | lepas [5] |, and

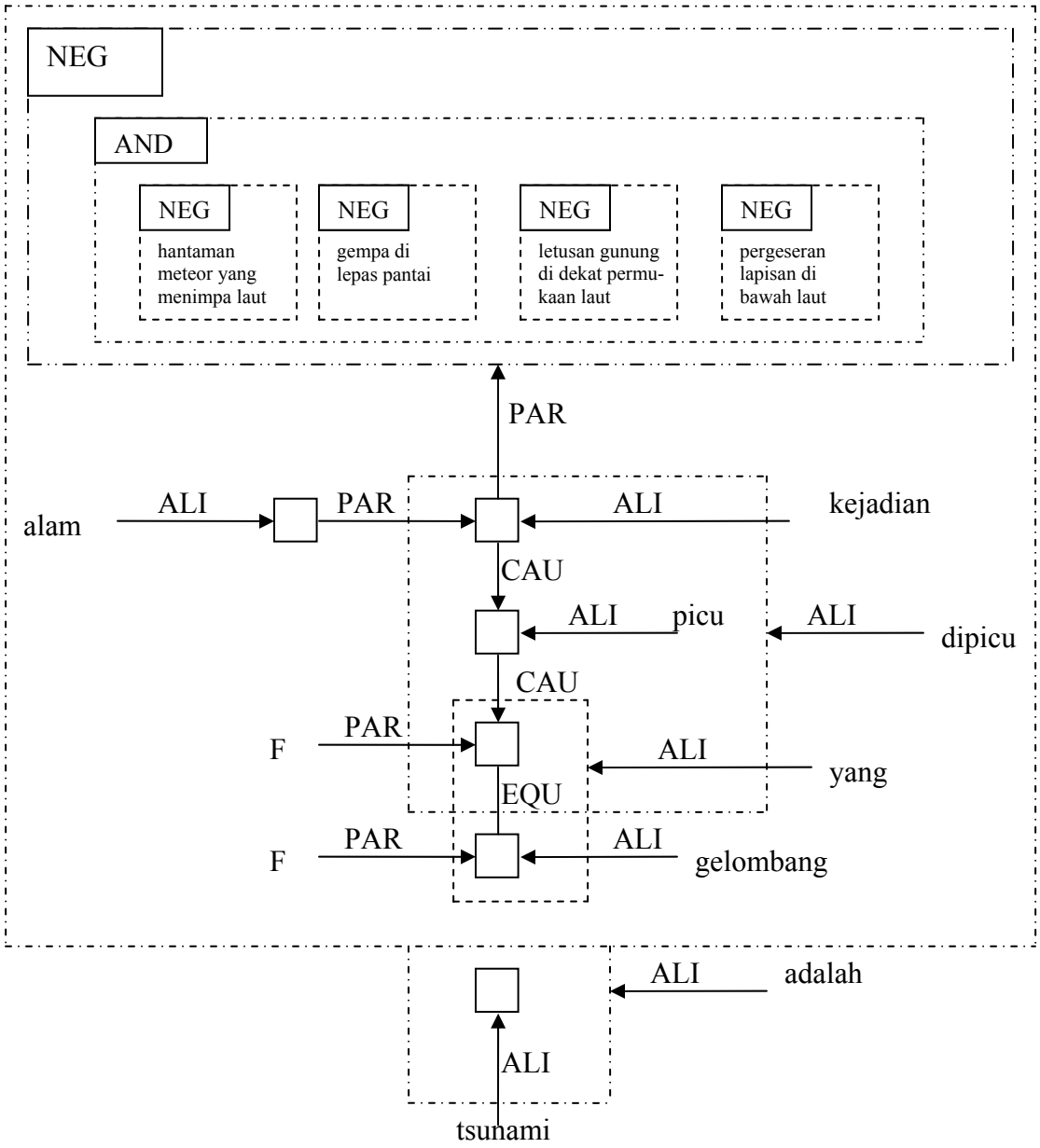


for | di [5] | bawah [5] |.

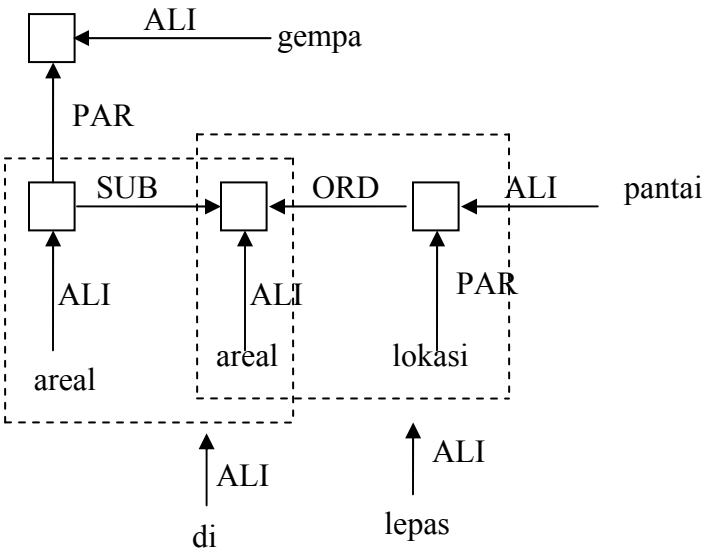
The next step is to indicate which concepts are described as far as their location is considered. For | di [5] | lepas [5] |, gempa and pantai attach to the outer tokens. For | di [5] | dekat [5] |, the concepts are letusan and permukaan. For | di [5] | bawah [5] |, the first concept is pergeseran and for the second the text gives laut. Here we meet a difficulty di bawah laut ≡ under (the) sea ≡ submarine, but what is the location of "sea"? In speech it is not clear whether the surface of the sea is meant, or the bottom, as reference point. For a boat the reference is the surface, but now it is indeed the sea floor.

The four alternative events can now be fillers of the tokens in atau, not seen in the exclusive or sense, and a frame can be drawn around the whole constructed graph. The token for gelombang gets a focus label and | adalah [3] | is giving rise to the typing relation between "tsunami" and the frame.

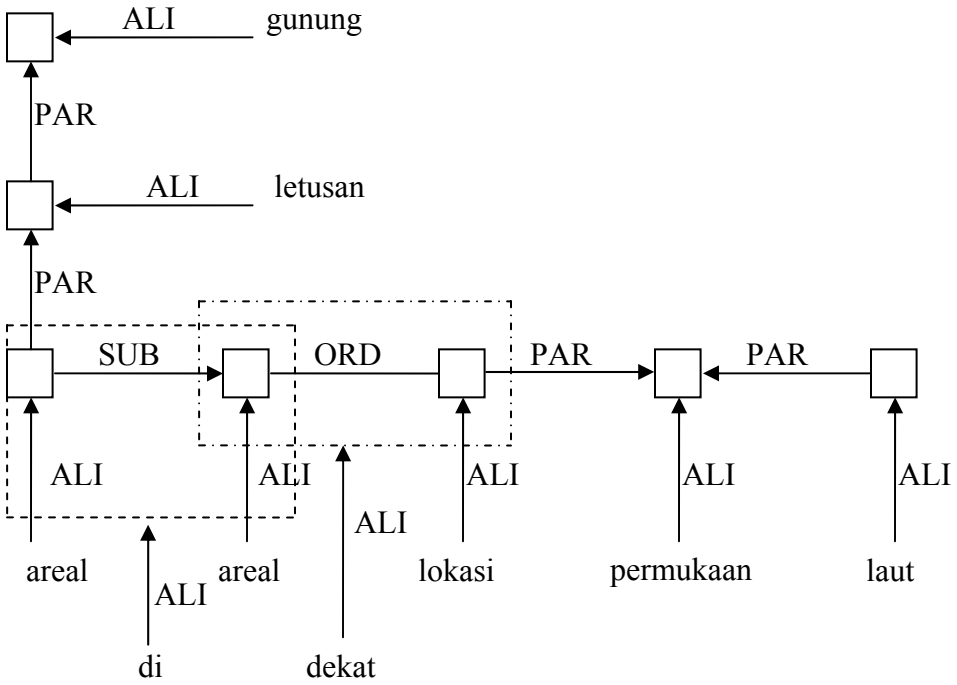
Our analysis thus gives the word graph for "tsunami" :



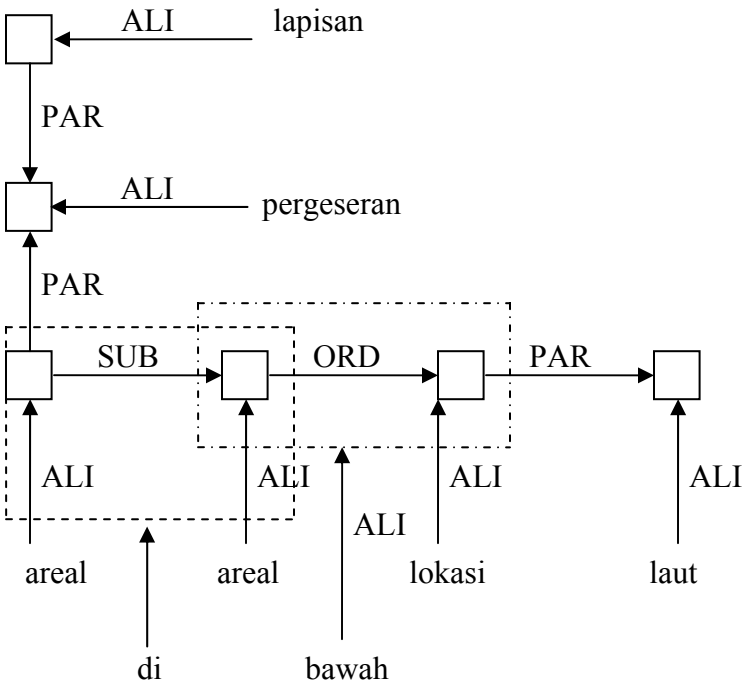
The graph of "gempa di lepas pantai":



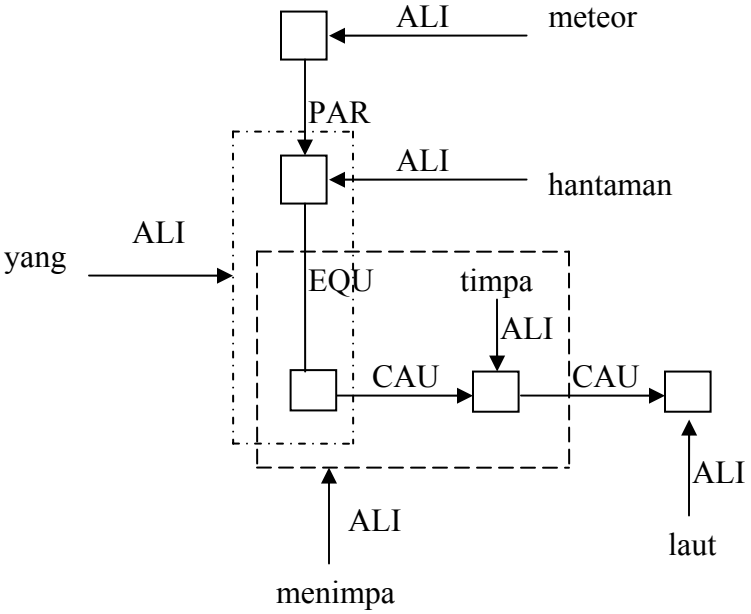
The graph of "letusan gunung di dekat permukaan laut":



The graph of "pergeseran lapisan di bawah laut":



The graph of "hantaman meteor yang menimpa laut":

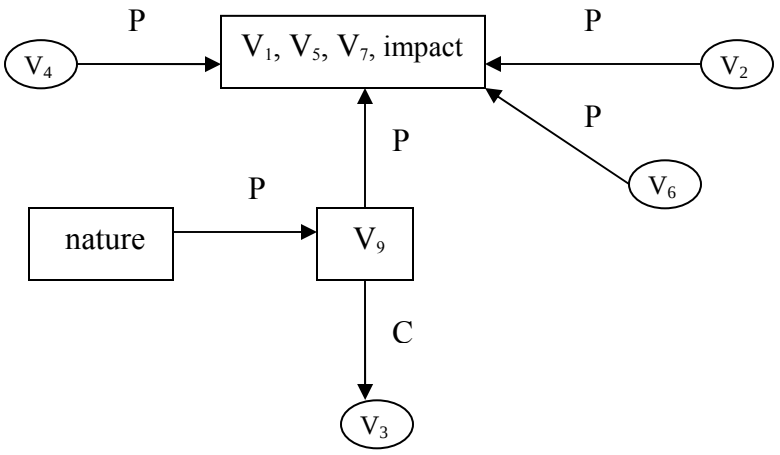


3.3. Comparison with the result from KISS

The main difference of the structural parsing result with that obtained in Section 2 lies in the precision with which the ontology has been used. We considered only CAU- and PAR- links in our personal interpretation of the 'English' definitions. Our main goal was to show how the system KISS worked and in particular which procedures could be distinguished.

The goal in Section 3 is to show how in a systematic way a sentence can be mapped on a sentence graph, the process that is called structural parsing, for the language of Bahasa Indonesia. It was not a priori clear that the procedures that work for English and Chinese also work for Bahasa. It turns out that chunking followed by constructing chunk graphs and combining chunk graphs can be carried out in an analogous way. It is interesting, but of secondary importance, to compare the results along both ways.

For this comparison we should stress two points. First, we choose English in both cases and second, as the first result was obtained by considering nouns only, in the result of Section 3 we also consider only nouns. The one verb form, dipicu, will be represented by a single CAU-arc. The result of the structural parsing is now



We recognise the causal axis of the outcome in Section 2. However, there are subtle differences. "event"= v_9, being any combination of the four specified events, is given as direct cause of the "wave"= v_3 here, whereas in Section 2 "slide"= v_7 came forward as direct cause of the wave. Here, "meteor *impact* on sea" is mentioned as event. No slide is necessarily involved in that case.

References

[1] Hoede C. and S. Nurdiati, *A Graph Theoretical Analysis of Certain Aspects of Bahasa Indonesia*, University of Twente, Memorandum 1870, (2008).

[2] Bakker, R R., *Knowledge Graphs : Representation and Structuring of Scientific Knowledge*, PhD Thesis, University of Twente, Enschede, The Netherlands, ISBN 90-9001963-4, (1987).

[3] De Vries Robbé, P.F., W.P.A. Beckers and P.E. Zanstra, *MEDES : Het Prototype*, UHG/MIDS/E00, ISBN: 90-367-0086-8, (1988).

[4] Zhang, L. *Knowledge Graph Theory and Structural Parsing*, PhD Thesis, University of Twente, Enschede, The Netherlands, ISBN 90-3651835-0, (2002).

[5] Abney, S. P. *Parsing by chunks.* In: Principle-Based Parsing (R. Berwick, S. Abney and C. Tenny, eds.), Kluwer Academic Publishers, (1991).