Graph Theoretical View on Text Understanding

Jure Zupan
National Institute of Chemistry, Ljubljana
E-mail: jure.zupan@ki.si

Keywords: graph theory, cyclic-connected graph, topological distance, network text analysis, information content

Received: October 6, 2017

The system STAVEK-02 described in the contribution is concentrated on yielding supplemental information (besides parsing/tagging of words) for text understanding through the clustering of nouns and/or verbs according to their meanings and common features. The system consists of two word processing blocks. The first block is a vocabulary of 149,000 Slovenian word-roots and 3,100 endings and assigns the grammatical feature to the words by the grammatical rules without any link to pre-tagged lexical corpora. The second block is a Network of meanings of Slovenian words which in principle is a graph connecting 45,000 and 15,000 noun and verb lexemes, respectively, all of them hierarchically clustered into larger and larger groups having /exhibiting specific features and/or common properties of the words encompassed Such formations are in a similar lexical systems usually called synsets. Due to the complete connectivity between the synsets (groups) in the graph it is possible to find all possible property/feature paths between any pair of two words (nouns and/or verbs) in the network. Because clustering of words according to their meanings is made during the parsing of one, a pair, or several consecutive sentences, the features and properties that appear on the closest path between the particular words within the sentence are quite informative for their interpretation of the text. Clustering of the words according to their meanings during the parsing of text is a novel concept of the text interpretation. Ob the basis of a simple example of parsing a sentence and clustering of the nouns within it the concept using the network of meanings in the program STAVEK-02 is described and discussed.

Povzetek: Opisani sistem STAVEK-02 je orientiran na širše izločanje informacij iz slovenskih besedil, kot je samo besedna analiza in označevanje besed. Osnova sta dva programska dela. Prvega sestavljata podatkovna baza (149.000 korenov besed in 3.100 končnic), drugega pa 45.000 samostalnikov in 15.000 glagolov, ki so s skupinami teh besed grupirani po različnih skupnih značilnostih v ciklični graf (connected cyclic graph). Prvi del izvrši slovnično označevanje besed v tekstu, drugi pa med posameznimi besedami, ali v grafu hierarhično povezanih skupin besed (synsets) s podobnimi lastnostmi in značilnostmi izračuna topološke razdalje in nariše shemo povezovanja skupin samostalnikov ali glagolov. Izkazalo se je, da topološko izračunana razdalja med besedami dobro predstavi pomensko razliko/sličnost med njimi. Obe besedni zbirki skupaj vsebujeta in obdelujeta pretežni del najpogostejših slovenskih besed (cca 149.000 slovenskih besed). V prispevku so razložene nekatere pasti slovenščine pri obvladovanju več-smiselnosti besedila. Opisana je tudi struktura cikličnega grafa besed (samostalnikov in glagolov) in način izračuna topološke razdalje med besedami Poudarjena je dvosmernost poti in sprehodov (paths and walks) v omenjenem grafu besed. Dodan je kratek primer analize stavka, ki se konča z matriko topoloških razdalj med besedami stavka in drevesom podobnosti. Na koncu so omenjene nekatere možnosti razvoja sistema STAVEK-02 in hierarhične mreže za določanje pomenov slovenskih besed.

1 Introduction

The parsing or tagging of words in the sentence provides the user with all relevant grammatical features of each word, which itself is a very hard task to implement either by the computer or by hand alone. The fact hat most of the modern parsing programs today rely on large corpora of previously parsed data does not mean that the efforts and programs solving the tagging of sentences by hand are either unnecessary or outmoded. Even if one forget that the testing of parsing-algorithms based on previously parsed corpora first relay on the hand-made parsing, the ab-initio, i.e., parsing by exclusively using grammatical rules will always be necessary. It should not be forgotten that statistical solutions mostly ignore the occurrences of rare specific cases. Such problems can be solved easier by considering and combining both methods (corpora driven and rule-based tagging) consecutively and/or iteratively. For example: the problem of the words having two or more clearly different meanings of which at least two can have grammatically correct but for any kind of machine parsing or rule-based tagging completely indistinguishable forms. Unfortunately, in Slavic languages with a much higher degree of flexibility of words than in English the problems of the word senses begin already on the parsing level. In the case of a grammatically correct sentence with two completely different interpretations of word senses it is possible that no parsing can correctly identify even the word classes of the constituent words, not to talk about the senses. The possible solution of such problems is to list all possible meanings or senses of each word and leave this information for further consideration when the context of the following sentences allow to single-out the actual meaning. For example, neither the sentence To je dobro za vas nor the title of the well-known Slovenian story Martin Krpan can be tagged correctly by the computer. In the first case the word vas can be interpreted either as for you or, alternatively, as the village, hence, the sentence can mean either: This is good for you, or This is good for the village. In the second example, the title of the well known Slovenian story Matin Krpan introduces the name of the main character. However, the title has, unfortunately, a second grammatically correct meaning of the word Martin, not as a noun (name Martin) but as the adjective meaning belonging to female Marta, which implies that man of the name Krpan is a husband of Marta or at least involved with Marta. Of course, the machine interpretation based on the pre-tagged corpora will always yield grammatically ‘correct’, i.e., the most often used variation, but at the same time always omit the less probable, but grammatically correct possibilities, witch nevertheless can appear in the spoken or written communication, and should therefore be at least considered. Such cases are handled better by the rule-based tagging compared to the statistical ones.

In order to bring attention to such possibilities and to provide the tool for helping the developers of man-machine dialog to handle such cases the program STAVEK-02 with options of showing *all* grammatical possibilities and additionally provide the user with clusters of various word meanings at each sentence (or group of sentences) was developed and is described in this paper.

2 Related work

The most closely related system to the PMSB (*Pomenska mreža slovenskih besed* [1], (Engl. Network of Meanings of Slovenian Words) used by the program STAVEK-02 is the well-known WordNet [2,3] lexical collection developed by the Princeton University with its graphic visualization VisuWords [4] based on the Thinkmap, data visualization technology. In order to handle the difficulties in the cross-language differences in the meanings of lexical words the Universal Word Net (UWN) Project was launched [5,6]. According to the UWN suggestions and guidelines specific versions for close to 200 different languages are now under development. Similar to the other Slavic languages (see Polish [7], or Bulgarian [8], for example) the Slovenian version named sloWNet [9] is as well progressing. At the moment the version described in the present paper is not included into sloWNet. There are several features of the PMSB that are similar to the WordNet but some of them are not. The organization of synsets for nouns in the hipo- hypero-, mero-, and holonym groups (the word A is a *meronym* of B if A is a part of B; the nose is a part of head, while head is a *holonym* of nose) is very similar, while the verbs in PMSB follow closely the six branch division (*to exist, to have, to move, to do/to, to think/to create, and to sense/to*) as suggested by Vidovič Muha [10] is quite different. The way the distances between the word senses in PMSB are calculated compared to the similarity evaluation between two synsets in WordNet is practically the same: it calculates the length of the shortest path between two nodes in the graph. It is worthwhile to mention that the distance measure used in our case is the length of the shortest path between two nodes (synsets) in a graph. This graph theoretical path distance is not related to the distances between objects (words) represented by the multi-dimensional distributed representations of word vectors as obtained by the word2vector software [11] developed by Thomas Mikolev at Google. The number of words and meanings (synsets), 60,000 and 110,000, respectively, in PMSB is already large enough to cover a large variety of texts.

A considerable difference with WordNet is in the design of our network STAVEK-02. Although the PMSB can act as a stand-alone program in the role of a sort of thesaurus of Slovenian language, its is actually designed as a subroutine to support the system STAVEK-02 which goal is to enhance and/or to improve the machine-man dialog, by pinpointing and/or explaining the *meanings* of specific words.

The mentioned goal can be clearly seen through the selection of hyper- and hyponym groups of the PMSB network which is described in the following paragraph more in detail.

3 Hierarchical Network of Meanings of Slovenian Words (PMSB)

The solution to the discussed information enhancing problem seems to be the organization of words into network of words linked according to the common features or some other commonly present or absent property(ies). Therefore, the links (branches) between nodes in the graph must contain meaningful information about the relation between the nodes they connect. For example: if one node is labeled *tool* and the other one *object* (*man-made*) the link between them must exhibit the property that the first node (synset) labeled *tools* is a part of the second node labeled *all man-made object*) and not *vice versa*. At the same time these two nodes should occupy positions in the work much closer to each other than they have to the synset labeled *insect*, for example. Either individual words or clusters of words could simultaneously be members of several groups (synsets with larger number of meanings) what makes the network to contain cyclic paths (circular paths between clusters) in the structure (Figure 1).

	VERBS (24,626)
Verbs of existing (3,405)	to exist on a specific way (542), verbs to sustain living (1,427), to end existence (299), emission verbs (949), weather verbs (187)
Verbs of having (1,339))	to posses (154), to obtain/take (333), to use possession (288), to negotiate possession (461), to spend possession (102)
Verbs of moving (3,129)	to move (general) (804), to move (specific way) (692), to move (body/parts) (629), to arrive/leave (676), to change movement (206), to do while moving (121)
Verbs of doing (9,663)	to put (2,416), to do (general) (669), to assemble/disassemble (1,340), to change (2,164), to use force/influence (1,322), to do complex tasks (1,751)
Verbs of thinking/creating (1,583)	to create (intellectually) (550), to think (general) (145), to think (specific) (407), to expressing thoughts with symbols (480),
Verbs of communication (5,507)	to exchange of information (2,770), verbs of perception (322), to have/response to feelings (883), verbs of social contact (1,531),
	NOUNS (86,799)
nature (31,988)	**nature (non-living)**(3,130) is divided into: nature (general) (10), nature (phenomenon) (521), nature (physical parameter) (151), nature (space) (82), matter (general) (1,359), matter (Earth) (933), matter (outer-space) (84) **nature (living) (28,847)** is divided into: nature (general/broader) (4,218), nature (plant kingdom) (3,111), nature (animal kingdom) (3,431), nature (human) (18,087)
product (19,222)	**product (origin) (552)** divided into: product (origin (human)) (40), product (origin (nature)) (53), product (origin (plant)) (258), product (origin (animal)) (201) **product (human) (18,670)** divided into: product (human (material)) (13,190), product (human (intellectual)) (5,352) product (human (commodity)) (29), creation (general) (5), creation (limitation) (94)
concept (35.589)	**activity (11,645)** is divided into: activity (general) (101), activity (to do something) (3,507), activity (society) (3,045), activity (emotion) (76), activity (sense) (15), activity (existence) (1,068), activity (movement) (1,240), activity (communication) (1,912), activity (possession) (582), activity (mind) (97) **property (5,943)** is divided into: property (action) (323), property (animal) (45), property (broader meaning) (357), property (company) (17), property (device) (90), property (form) (62), property (general) (37), property (human) (2,774), property (mind) (128), property (matter) (267), property (nation) (35), property (number) (13), property (object) (482), property (phenomenon) (42), property (plant) (34), property (procedure) (390), property (religion) (15), property (ruling) (52), property (society) (111), property (sound) (39), property (space) (309), property (status) (159) property (word/speech) (123), group of properties (38), and 8 other groups: **event (1,208), form (3,169), group (1,958), phenomenon (526), procedure (992), result (5,342), space (1,532), state (2,910).**

Table 1. The first two levels of verbs (upper part of the Table 1) and nouns (lower part of the table) are shown according to their common features. In the parentheses the number of words in each group is given. Because individual word can have several meanings or senses it is listed in as many groups (synsets) as there are meanings. Therefore, the sum of words given in parenthesis is larger than the number of meanings in the network. The largest groups are printed bold.

The PMSB Network consists of 45,000 noun and 15,000 verb dictionary lexemes (words) forming 85,000 and 25,000 different entries of noun and verb meanings, respectively. For example, if *'kon*j' (Engl. *horse*) is one of the 45,000 lexemes the four senses of the word '*horse*' in Slovenian language (*horse* – an *animal*, *horse* – a *clumsy man*, *hors*e – a *chess-piece*, and *horse* – a gymnastic equipment, *paddle-horse*) are four of 85,000 noun meanings or senses.

Using the above kind of reasoning, a graph of about their meanings and properties containing close to 4,500 clusters of words (nodes) was generated [1]. The closest collection to our database is the Levine's collection of verb classes [12] and Dornseiff's Wortschatz [13]. There are various Internet versions like WordNet [2,3]) and for the Slovenian language the sloWNet [9]. What the size, i.e. the number of words is concerned; only the Dornseiff's [13] collection has about the same number of verbs (14,000) as our collection. The part of our network containing verbs is based on six main groups [10] and is already well described in the literature [14,15] and is accessible on the web [16]. The complete structure of verb hierarchy in English language (16,000 verbs and 1000 groups) is given in [17]. The basic division of nouns has three groups: the *product*, the *nature*, and the *concept*. It can be seen from second part of Table 1. The clusters of verbs and nouns in all levels of hierarchy are of very different sizes (Table 1).

On the contrast to the English language, the Slovenian lexical forms of verbs can be well distinguished from those of nouns, however, due to high flexibility of Slovenian declination and conjugation (approximately 20 per each noun, verb, adjective, pronoun, and numeral) there are numerous cases where two or even three word types mix. For example the sentence *To je lepo padalo* has two meanings: a) *This is a nice parachute* and b) *It was falling nicely*. In the first case the word *padalo* is a noun (*parachute*) while in the second case it is the verb (*to fall*). To have all words together in one network (graph) both word types are linked in the network on the highest node.

It is worthwhile to mention that the same word in different languages has different synsets of meaning. This is the reason why such a hierarchy cannot be 'blue-printed' from one to another language. The effect of 'lost with translation' is unavoidable: each translated word could be connected to completely different clusters of words. For example, the English word *plant* in its botanical meaning can be linked with Slovenian counterpart *rastlina*, or German *Pflanze*, but has no connection to the second sense of a production place like Slovenian *tovarna* or German *Fabrik*).

4 Semantic distance measure

Mathematically, the network is a connected cyclic bi-directional graph. Vertices or nodes represent single words, meanings and/or or clusters of words with similar properties/features (synsets). The connected graph enables a continuous walk, described as a sequence of connected nodes (path), between any two nodes. The graph is cyclic if it contains closed paths (cycles), i.e., paths that starts and ends on the same node) with all nodes on that path different (with exception of the closing node). Hierarchical graph has one special node called top node Ntop or root, distinguished from the other ones by defining the orientation of the graph and walk directions within it. All valid paths between nodes must have one of the two directions: either towards the Ntop (up) or backwards from (down). Therefore, each node must have two lists for connections, to up and to down connected neighbors, respectively. Similar to the Ntop which is the last node of all up-paths, so at the end of any down-paths is always a node called terminal, having no down directions. The terminal nodes are individual words or senses if the word has only one sense (meaning).

The fact that the walk path is not allowed to change direction assures that from any node one can always reach either a terminal node or the N_{top}. Thus no walk with the constant direction could be captured in a cycle and thus end in an infinite loop. In the case of update of new words or relocation of nodes the described hierarchy prevents updates to generate infinite loops and self-referencing nodes. All the explained features of our graph offer the advantage of calculation the topological distance between the nodes. The topological distance D_{ij} between two nodes N_i and N_j has all four properties classifying it as a standard metric distance:

1) $D_{ij} > 0$ for all $i \neq j$

2) $D_{ij} = 0$ only for $i = j$

3) $D_{ij} = D_{ji}$, the distance is symmetrical, and

4) $D_{ij} \leq D_{ik} + D_{kj}$ triangle rule for any node k

To evaluate all topological distance D_{ij} between two arbitrary nodes N_i and N_j in the graph, one needs a complete connectivity matrix of order ($N_i \times N_j$). For a graph containing approximately 10^5 nodes this means storing and handling the matrix of about 0.5×10^{10} distances. Fortunately, instead of keeping this large connectivity and/or distance matrix, only two connectivity tables one for keeping all *up* and the other one keeping all *down* connections from each node to neighboring nodes are needed. Using these two connectivity tables it is straightforward to determine topological distance between any two nodes N_i and N_j or words *i* and *j*, respectively. The procedure is as follows:

1. Find the complete set $\{P_i\ (N_i,N_{top})\}$ of n_i paths from the node N_i to the node N_{top}.
2. Find the complete set $\{P_j\ (N_i,N_{top})\}$ of n_j paths from the node N_i to the node N_{top}.
3. Compare *k* pairs of paths from *both* sets $\{P_i\ (N_i,N_{top}),\ P_j\ (N_j,N_{top})\}$, $k = 1...n_i(n_j - 1)/2$ and for each pair determine the common node C_k
4. Determine the length l_k of the path from node N_i to the node N_j passing node C_k for each pair *k*.
5. Keep the shortest path.

To summarize: the distance D_{ij} between two nodes N_i and N_j is the length l_k of the shortest path from node N_i to the node N_j through the common node C_k, from which both nodes N_i and N_j have access to N_{top}:

$$D_{ij} = \min\{\ l_k\ \} \text{ of } \{P_k(N_i,\mathrm{C},N_j)\ \},\ \ k = 1...n_i(n_j - 1)/2 \qquad /1/$$

where n_i is the number of *different* paths from the node N_i to the top node N_{top}; and $P(A,\mathrm{C},B)$ is the path from node *A* to node *B* passing node *C*.

5 The Case Study

The described system STAVEK-02 can serve as a model how to use the PMSB hierarchy of word meanings and synsets for enhancing the information in free text. The system can handle individual sentences input by the keyboard or text files of any size. The system handles sentences one by one, hence, the information are reported

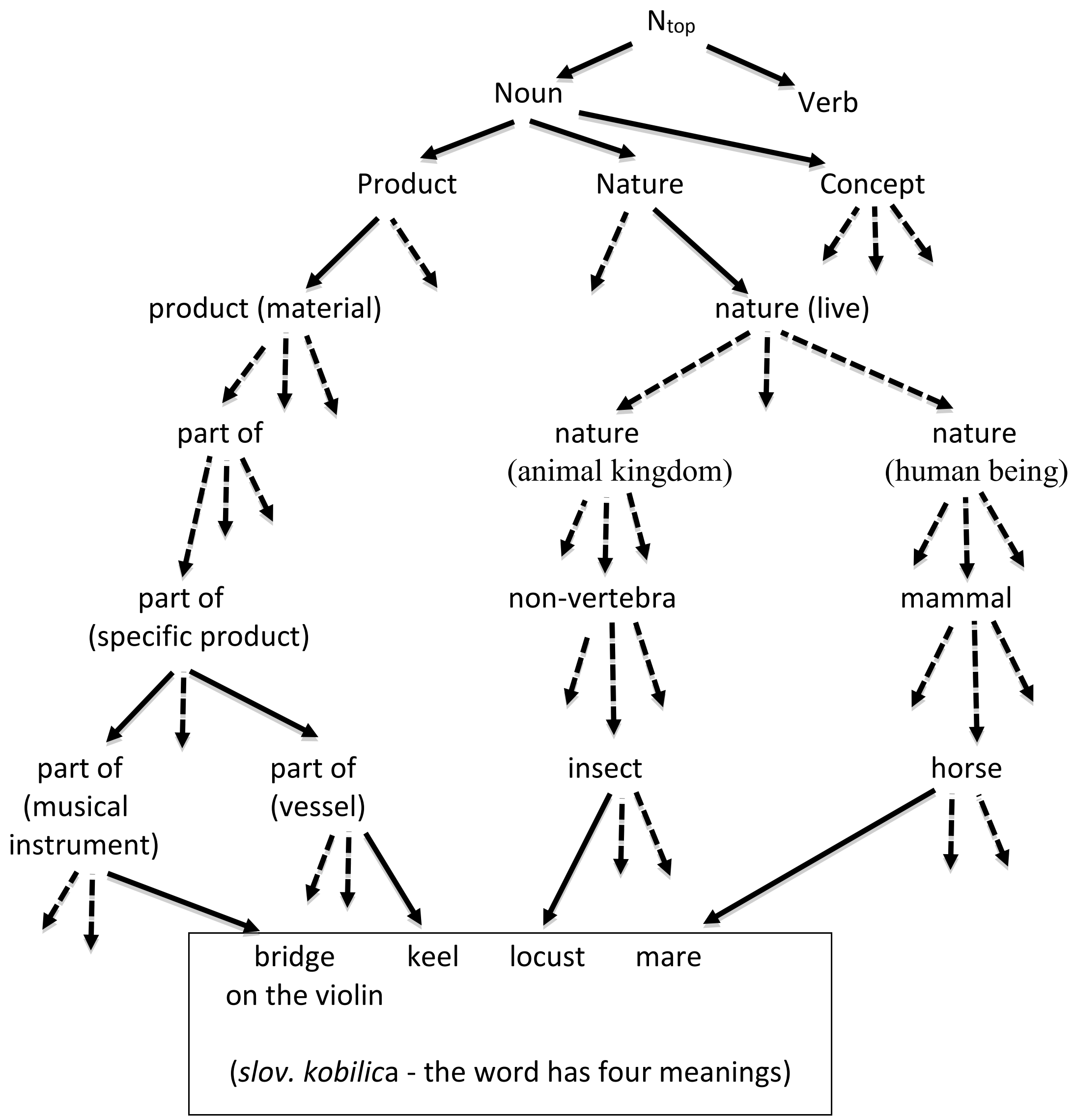

Figure 1. A simplified part of the discussed network of words showing essential features of a cyclic bi-directional graph. Each label represents a node (synset). A cycle is a path that starts and ends on the same node. From the word *kobilica* having 4 meanings in Slovenian language, six cycles can be drawn to calculate six distances between all four meanings. Because the graph is 2-directional, only the paths in *up* or *down* to the N_{top} (opposite to arrows) or to the terminal nodes (words, along arrows), respectively, are allowed. The cycles are detected *via* the common nodes C_k on the paths.

at the end of each sentence. First, all grammatical information for all words in the sentence, are reported (Part A in Figure 2). This part, of the tagged text is similar in the content, but quite different in the form to the output provided by the public Slovenian parser [18] available on the Slovenian ZRC portal. All tasks performed by the parser are executed *ab initio*, i.e., by the grammatical rules without considering any corpus or web connection. For highly flexible language like Slovenian, the parser is very important, because for tagging of words it must convert the words into appropriate lexemes, what is mandatory form for further search in the PMSB network. Our tagging system uses 149,000 Slovenian word-roots for all types of words and combines them with about 3,100 different endings in order to obtain the correct lemma for any Slovenian word in SSKJ [19].

Second, for each noun and each verb appearing in the input sentence(s) the system searches through PMSB network and lists all the meanings/senses associated with any noun and verb together with the complete list of synsets from the lemmas to the top of the network. The list of senses for all words together with full paths through the PMSB is output after each sentence (part B in Figure 2).

Third, the program draws a clustering scheme of up to 500 nouns or verbs together with all their senses. The cluster-tree of senses is built into the system as an option and can be performed a) after each sentence, b) after a certain number of sentences providing the number of

```
Part A
===================================
 1  Na; na; predlog, (4,5);
 2  violini; violina; sam., f(d/4)(e/5);
 3  je; biti; pom.gl., (3os./f/e) pretekli čas
 4  počila; počiti; gl., nedov. nepov. (3os./f/e) pretekli čas
 5  struna; struna; sam., f(e/1);
 6  . ločilo, pika;
-----
 1  Kaj; kaj; prislov, (vzročnostni);
 2  je; biti; pom.gl., (3os./m/e) pretekli čas
 3  bil; biti; gl, nedov. nepov. (3os./m/e) pretekli čas
 4  vzrok; vzrok; sam., m(e/1)(e/4);
 5  ? ločilo, vprašaj;
-----
 1  Kobilica; kobilica; sam., f(e/1);
 2  je; biti; pom.gl, (3os./f/e) pretekli čas
 3  bila; biti; gl, nedov. nepov. (3os./f/e) pretekli čas
 4  poškodovana; poškodovan; pridevnik, (m/d/1)(m/d/4)(f/e/1)
 5  . ločilo, pika;
===================================

Part B
===================================
/001/01: violina (violin); strings; instrument (musical (specific)), instrument (musical); product (sound emitting);
         product (communication); product (material); product/creation; Noun; Ntop.
/002/01: struna (string): part of (musical instrument)), instrument (musical); product (sound emitting); product
         (communication); product (material); product/creation; Noun; Ntop.
/002/02: struna (string); product (sound emitting); product (communication); product (material); product/creation;
         Noun; Ntop.
/003/01: vzrok (cause); factor; measure (specific); creation (measure/unit); creation (intellectual); product/creation;
         Noun; Ntop.
/004/01/ kobilica (violin’s bridge): part of (musical instrument); instrument (musical); product (sound emitting);
         product (communication); product (material); product/creation; Noun; Ntop.
/004/02/ kobilica (keel): part of (vessel); part of (specific device); product (machine/device); product (general part);
         product (material); product/creation; Noun; Ntop.
/004/03/ kobilica (locust): insect; insect (pterygota); insect (arthropoda); insect (general); antropoda; non-vertebra;
         nature (animal taxonomy); nature (animal kingdom); nature; Noun; Ntop.
/004/04/ kobilica (locust): insect; insect (pterygota); arthropoda; polimeria; animal (common name); nature (animal
         kingdom); nature; Noun; Ntop.
/004/05/ kobilica: mare; horse (animal (general)); horse (animal); animal (domestic); animal (property); nature
         (animal kingdom); nature; Noun, Ntop.
/004/06/ kobilica: mare; horse (animal (general)); horse (animal); odd-toed ungulate; mammal; vertebra; chordata,
         nature (animal-taxonomy); nature (animal kingdom); nature; Noun, Ntop.
===================================
```

Figure 2: Output of the program STAVEK-02 after the input of three sentences representing a short dialog. *Na violini je počila struna. Kaj je bil vzrok? Kobilica je bila poškodovana.* (Eng.: *The string on the violin broke. What was the cause? The bridge was damaged.* The word types are nouns (sam.), verbs (gl.), adverbs (prislov), adjective (pridevnik), the letters *m, f, os, e,* and *d* stand for (masculine, feminine, person, singular, and dual), respectively; the numbers mark the falls. Part B shows ten chains of nodes (synsets) of words and meanings from the PMSB network as used for the distance matrix ***D*** and dendrogram calculations (see Figure 3). N_{top} is the top node of the PMSB hierarchy of meanings. In the actual output of program STAVEK-02 the synsets assigned to words of one sentence are printed immediately after one of the main three punctuation marks (full stop, question mark, or exclamation mark) is encountered.

words does not exceed 500, or c) at the end of parsing a text file after the user can selects up-to 500 nouns or verbs from the list of the most frequent word types of the scanned text.

Finally, at the end of each session (either for one sentence or for the text file) the program yields a) statistics of the input text with respect to the word frequencies of all word types and separators, b) the distribution of word-lengths (in characters) of each word type, and c) the frequency is of 2000 most frequently used nouns, adjectives, verbs, and adverbs.

In order to show the entire procedure more in detail the output as given by the system STAVEK-02 for three short consecutive sentences is worked out and discussed more in detail. The three sentences in English translation are: The string on the violin broke. What was the cause? The bridge was damaged. (slov. Struna na violini je počila. Kaj je bil vzrok? Kobilica je bila poškodovana.) (Figure 2, parts A and B). This particular example using the word kobilica in two separate sentences was chosen deliberately to show how the graph-theoretical distances (Figure 2. and Figure 3) as obtained by the PMSB network could correctly determine the sense of a word. Similar to English the word bridge having several senses, the Slovenian word kobilica has been coded by six synsets in PMSB. It has four (4) main senses (locust, keel, mare, and the bridge on the violin) of which both animal senses have two synset paths for showing the relevant taxonomies of both species. (Figure 2, part B).

Each chain is a sequence of labels of nodes (synsets) encountered during the walk between the word and the N_{top}. The search algorithm finds all possible walks from any encountered noun or verb to the N_{top}. The reader can verify this part of the search engine in real time on-line on the link given in [20]. Mostly, the labels are organized in self-explanatory manner using structure of keywords in which each keyword is itself a cluster label with the link to the particular cluster in the network. For example, the node labeled *property* (*human*) contains words each of which marks a property of a human' (*intelligence, beauty, greed, innocence*, etc.). On the other hand, the words in the cluster with the same two keywords, but ordered differently e.g., *human* (*properties*) describe a human being with a particular property, *genius* and *liar* are in the synsets *human* (*property* (*intelligence*)) and *human* (*property* (*bad*)), respectively. Additionally, both words *human* and *property* are labels of other clusters. The cluster *property*, for example, contains 5,964 nouns with 14 sub-clusters named property (*keyword$_i$*), i = 1,…14. Each keyword of these clusters: *property* (*animal*), *property* (*human*), *property* (*number*), … *property* (*object*), contains again cluster descriptors with keywords. Take for example the sub cluster *property* (*object*): *property* (*object* (*color*)), *property* (*object* (*form*)), *property* (*object* (*price*)). At the end each *keyword$_i$* represents a cluster with a smaller set of words.

Table 3 shows the topological distance matrix ***D*** of 45 distances between the ten meanings. All distances reflect the relation between the similarities of meanings of the words concerned very reasonable. The two main groups, the upper one representing material products (*violin, string, bridge on the violin*) and the lower one representing *locust* and *mare*: have two descriptions each, respectively. In the middle of both groups is the word *vzrok* (*cause*), representing the concept of non-material products. In the group of *material objects* the string /002/01/ *(part of the violin*) and *kobilica* /004/01/ (*part of the violin*) are joined at the lowest level. The pair goes together with the second meaning of the string /002/02/ as a sound emitting device and then three join

There is not much to say about tagging shown as part A in Figure 2), however, the tagging the second word violini as singular locative (e5) is a good example showing how the statistical approach ignores the possibility that the word violin has in the dual the same form (for example: 'Pozabil sem na violini' Engl. I forgot about two violins) of accusative in dual (d/4). STAVEK-02 tags both possibilities (d/4) and (e/5). Additionally, the rule-based tagging is considerably faster compared to the statistical pre-tagged-corpora-based one. The public Slovenian parser [18] can tag on the average 8 sentences per second, while the parser of the system STAVEK-02 managed to tag 400 sentences per second. By additionally searching for all noun and verb meanings through the database of close to 110,000 synsets makes the rule-based parser almost two orders of magnitude faster then the public one. Part B shows all the synset paths for the nouns in the sentences. In the actual output the synset paths for verbs are also given. In the print option, the paths are listed after each sentence.
together with the fourth sense *violin* combining all four into a reasonable synset *musical instrument*. As said above, the last four meanings represent the animal synsets (*animal living beings*). To this group of four meanings (*horse* (*domestic animal*)), *horse* (*taxonomy*), *locust* (*insect*), and *locust* (*taxonomy*), there is no counterparts of meanings from the rest of the considered three sentences, hence, one can safely assume that the four meanings of the word *kobilica* do not apply in this context.

It is interesting to see that the remaining two words *kobilica* /004/03/ (*keel* as a part of a vessel) and *vzrok* (*cause*) fit well between the two larger group. The sense *keel* and *violin* are linked together relatively high in the dendrogram because there are both material objects, however, the level of the link between the concept *cause* and the material object *keel* shows that there is still a lot of space for improvements of the procedure for distance evaluation.

This results help us to argue that as much the meanings of single word is important, the distance between the words is important as well. This in turn requires two things; first each word should be represented in unique and uniform way based on various kind of properties and second, the words should bi organized in a system that allows definition of a metrics.

6 Conclusion

The discussed example and hierarchical network of words PMSB present only a very simple and small part of the general solution that can be accomplished by the use of an exhaustive and therefore much more complex network of word meanings. Neither the presented network, nor the presented model for extracting broader information from the text, is the final product. Still a lot of improvements can be implemented.

Although the present network links together slightly more than 60,000 words (nouns and verbs) forming about 110,000 meanings (synsets) of various sizes, it is not the number of words that is a limiting factor, but rather more factors like the absolute number of synsets (clusters of words with different features), the number of links to which each synset is connected, and least but not last the ability of algorithms for distance calculation to reflect the actual distinction between the meanings of word. These are the issues that should be of first concern. One should add not only more clusters presenting larger variety and number of properties, features, and/or meanings, but as well clusters of words pointing to rare, dangerous, or by any other criterion extreme features that the words represent, for example synsets containing words like non-poisonous plants, extremely hard or non combustible material, etc. The constant updating and enhancement of the networks of meaning require much more man-power and/or machine-supported feature selection efforts for addition of new groups than it has been spent for the present variations of WordNets on varieties of languages. However, for each specific language the native speakers are responsible for the growth and complexity of their specific meaning networks and no automatic procedure could completely replace their manual work and decisions. The presented PSMB network of meanings was put together by hand what requires approximately eight man-years to reach the present size. Some critics are afraid that such knowledge bases has arbitrary structure, because the meanings of the words are subjective and no objective criteria exist how to link or cluster words according to their meanings. The described example has shown the potential of such network to help understanding the context of the communication. As a matter of fact it is true, that such a hierarchy of meanings will always be subjective, but so is human mind.

7 Acknowledgement

The author wishes to thank National Institute of Chemistry for providing him with the facilities to work at the Institute as a research emeritus.

8 References

[1] Zupan, Jure; Koncept mrežnega pomenskega slovarja slovenskih besed, Jezik in slovstvo, 54, (3-4), 2009, pp. 139-151.

[2] Miller, George A, WordNet: A Lexical Database for English. Communications of the ACM. 1995, Vol. 38 (11), 39-41.

[3] Fellbaum, Christiane; WordNet: An Electronic Lexical Database, Editor, 1998, Cambridge, MA: MIT Press.

[4] Visuword™, On-line graphical dictionary and thesaurus, https://visuwords

[5] Towards a Universal Multilingual WorldNet - D5: Databases and Information Systems, Max-Planck-Institut für Informatik; Mpi-inf.mpg.de; 2011-08-14.

[6] Vossen, Piek, EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Editor, 1998, Kluwer, Dordrecht, The Netherlands.

[7] Maziarz M., Szpakowicz S., Piasecki M., Semantic Relations among Adjectives in Polish WordNet 2.0: A New Relation Set, Discussion and Evaluation, Cognitive Studies / Études Cognitives, t. 12, s. 149–179, 2012.

[8] Koeva, S., G. Totkov and A. Genov. Towards Bulgarian WordNet. Romanian Journal of Information Science and Technology, Vol. 7, No. 1-2, 45-61, 2004.

[9] Fišer, Darja, Novak, Jernej. Visualizing sloWNet. Proceedings of the conference on Electronic lexicography in the 21st century: New applications for new users (eLEX2011). Bled, Slovenia, 9-12 November 2011.

[10] Vidovič Muha, Ada, *Slovensko leksikalno pomenoslovje*. Ljubljana: Znanstveni inštitut Filozofske fakultete, 2000:

[11] Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S.; Dean, Jeff; Distributed representa-tions of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 2013.

[12] Levin, Beth; English Verb Classes and Alternations, The University of Chicago Press, Chicago, 1993.

[13] F. Dorenseiff, der deutsche Wortschatz nach Sach-gruppen, 8. Edition, Ed. U. Quasthoff, W. de Gruyter, Berlin, 2004.

[14] Zupan, Jure; Problemi in nekaj rešitev računalniških obdelav slovenskih besedil, Slav. revija, 47 (3), 1999, 277-296.

[15] Zupan, Jure; Hierarhična mreža slovenskih glagolov, v Obdobja 30, Interdisciplinarity in Slovene Studies, Filozofska Fakulteta, Ljubljana 2011, pp. 551-557.

[16] Zupan, Jure; Lajovic, Andrej; PMSG – Network of Slovenian verbs, web address: http://pmsg.zrc-sazu.si.

[17] Zupan, Jure; Pomenska mreža slovenskih glagolov, Založba ZRC SAZU, 2013, pp. 31-51,

[18] Oblikoslovni označevalnik za slovenski jezik, Amebis, d.o.o. Kamnik, Inštitut Jožef Stefan, Univerza v Ljubljani, ZRC SAZU, Trojina, Zavod za uporabno slovenistiko, 2008-2013, konzorcij projekta Sporazumevanje v slovenskem jeziku: link to the network: http//www. oznacevalnik.slovenscina.eu

[19] Slovar Slovenskega knjižnega jezika (SSKJ), Bajec, Anton, et al., Eds., Državna založba Slovenije, DZS, Ljubljana, 1995.

[20] J. Zupan, A. Lajovic; PMSB, Pomenska mreža slovesnkih besed, link to the network of meanings of Slovenian words: http://mreza.andrej.ad-vega.si.

Distance matrix between ten meanings of four words. The distances are the numbers of nodes (synsets) between two meanings in the network PSMB evaluated according to the procedure and equation /1/.

	1	2	3	4	5	6	7	8	9	10
1 violina (violine) /001/01	0	6	6	13	6	12	19	19	18	21
2 struna (string, violin's part)/002/01		0	4	12	3	11	18	18	17	20
3 struna (string, sound emitter)/002/02			0	10	5	9	16	16	15	18
4 vzrok (cause) /003/01				0	12	12	17	17	16	19
5 kobilica (violin's part)/004/01					0	10	17	17	16	19
6 kobilica (keel)/004/02						0	17	17	16	19
7 kobilica (locust) /004/03							0	3	15	16
8 kobilica (locust-taxonomy)/004/04								0	15	18
9 kobilica (mare)/004/05									0	3
10 kobilica (horse-taxonomy)/004/06										0

```
                  Strategy: Ward method
                  D(link)*100 / D(max)
100---90---80---70---60---50---40---30---20---10---00
                                             .______  violina /001/01
                                             |   .__  struna /002/01
                                             | ._|__  kobilica /004/01
                                  .__________|_|____  struna /002/02
                                  |    .____________  vzrok /003/01
  ._______________________________|____|____________  kobilica /004/02
  |                                              .__  kobilica /004/03
  |                  .___________________________|__  kobilica /004/04
  |                  |                           .__  kobilica /004/05
 _|__________________|___________________________|__  kobilica /004/06
```

Figure 3. The distance matrix ***D*** between ten different senses of four words (*violin, string, cause and kobilica*). The word string has two meanings a) part *of the violin* and b) *sound-emitting device*. The word *kobilica* has four meanings and six synset paths from the meanings to the top of the network (see Figure 1). The distances between individual meanings are calculated using the procedure and equation /1/. The dendrograms based on the distance matrix ***D*** can be output optionally after any number of tagged sentences providing there is no more than 500 nouns or verbs.