Taking Differences between Turkish and English Languages into Account in Internal Representations

Tassadit Amghar[2], Bernard Levrat[1,2], Sultan Turhan[1], and Burak Parlak[1]

[1] LERIA, Universit dAngers, Angers, France
[2] Galatasaray University, Department of Computer Engineering, Ciragan Cad. No:36
34357 Ortakoy, Istanbul, Turkey
amghar@info.univ-angers.fr

Abstract. It is generally assumed that the representation of the meaning of sentences in a knowledge representation language does not depend of the natural language in which this meaning is initially expressed. We argue here that, despite the fact that the translation of a sentence from one language to another one is always possible, this rests mainly on the fact that the two languages are natural languages. Using online translations systems (e.g. Google, Yandex translators) make it clear that structural differences between languages gives rise to more or less faithful translations depending on the proximity of the implied languages and there is no doubt that effect of the differences between languages are more crucial if one of the language is a knowledge representation language. Our purpose is illustrated through numerous examples of sentences in Turkish and their translation in English, emphasizing differences between these languages which belong to two different natural language families. As knowledge representations languages we use the first order predicate logic (FOPP) and the conceptual graph (CG) language and its associated logical semantics. We show that important Turkish constructions like gerunds, action names and differences in focus lead to representations corresponding to the reification of verbal predicates and to favor CG as semantic network representation language, whereas English seems more suited to the traditional predicates centered representation schema. We conclude that this first study give rise toideas to be considered as new inspirations in the area of knowledge representation of linguistics data and its uses in natural language translation systems.

Keywords: Knowledge representation, Effect of natural languages differences, Turkish, Reification, Conceptual graph.

1 Introduction

It is generally assumed that the representation of the meaning of sentences in a knowledge representation language does not depend of the natural language in which this meaning is initially expressed. We argue here that, despite the fact that the translation of a sentence from one language to another one is always possible, this rests mainly on the fact that the two languages are natural languages. Using translations systems

D. Nagamalai et al. (Eds.): *Adv. in Comput. Sci., Eng. & Inf. Technol.*, AISC 225, pp. 231–241.
DOI: 10.1007/978-3-319-00951-3_23

(e.g. Google, Yandex translators) makes it clear that structural differences between languages gives rise to more or less faithful translations depending on the proximity of the implied languages and there is no doubt that effect of the differences between languages are more crucial if one of the language is a knowledge representation language.

Therefore, this study aims at showing through the representations of some illustrative examples of sentences in Turkish and their corresponding translation into English, emphasizing differences between these languages, that the Turkish and English version corresponds to different structures each of them rendering the way each language has to depict the word.

As knowledge representation language we use the first order predicate logic (FOPP) and the conceptual graph language (CG) introduced by J. Sowa [REF] and its associated logical semantics. We show that important Turkish constructions like gerunds, action names and differences in focus lead to representations corresponding to the reification of verbal predicates and to favor CG as semantic network representation language. For their part, English seems more suited to the traditional representations where the logical structure of formulae renders the predicates centered representation schema of these natural languages.

The first paragraph is devoted to the representation languages and since FOPL language is commonly well known we only give a sketchy presentation of it and a more detailed one for the conceptual graphs language (J. Sowa 84).

The following paragraphs each tackles some important Turkish structures whose from which corresponding translation in English differs greatly and the consequences on the representation in each language. We will in turn analyze the representations of gerunds, and action names constructions, structures induced by the lack of auxiliary to be and to have and finally expression of moods, and other feelings.

We conclude that this first study furnishes interesting ideas to be considered as new inspirations in the area of knowledge representation of linguistics data and its uses in natural language processing. This a primary work in this direction in the semantic field but we are currently studying other aspects of Turkish language (morphology, syntactic construction) which also deserves specific treatments.

2 Methodology

Knowledge representation languages are classically classified into two families: The logical languages on one hand and the labeled graph languages on the other hand.

- **The logic based language family:** This language family contains standard and non standard logic languages, but often, if soundness of reasoning is crucial for the application, classical logical languages are preferred due to a well grounded semantics and rather good tractability properties.
- **The graph modeling language family:** Also known as semantic nets [Findler, 1970] they are mainly inspired by psychological works [Collins&Quillian,1969]. If their expressive power is greater than the preceding representation languages, they loose in counterpart some formal properties in the definition of their semantic.

Nevertheless, to characterize this last one, a logical semantic in terms of first order predicate formulae is generally associated with them when it is possible.

2.1 First Order Logic

In this paper, we started to adapt a first order logical formalism and its reification to create a model the tuples in Turkish and English sentences. The relationship between the agent and the object was designed through the verbs. In an sentence like X is doing Y, the predicative first order expression will be written as do(X,Y) and X Y işini yapıyor as yapmak(X,Y), respectively. This relationship is generally represented as It exists R, R(x1,x2) similarly for both languages. On the other hand, the reification procedure re-expresses the same sentence in a distributive way.

2.2 Conceptual Graph Formalism

CGs are a knowledge representation language well known for its ability to cope with natural language data due to the direct mapping to language they permit [Sowa, 2000]. They currently conveniently serve as an intermediate language in applications between computer-oriented formalisms and natural language. We only focus here on the component of CGs necessary to I ⋃@the presentation of the main ideas underlying our study

Definition 1: A conceptual graph is a finite, connected, bipartite graph with nodes of a first kind called *concepts* and nodes of the second kind called *conceptual relations*.

Definition 2: Every conceptual relation has one or more arcs, each of which must be attached to a concept. If the relation has n arcs, it is said to be n-adic and its arcs are labeled 1,2,...,n

In the graphical representation, concepts are rectangles and relations are circles

Axiome 1: There is a set T of type labels and a function *type*, which maps concepts and conceptual relations into T. T is a partially ordered by a subsumption relation ≤. Practically T is often taken as being a complete lattice in case of simple inheritance assumption between types. If a is a concept, *type*(a)=ta if ta denotes the lowest type a belongs to.

Axiome 2: There is a set I={i1, i2, …, in} of individual makers and a *referent* application from the concept set to I∪{*} where *referent*(a)=ij denotes an a particular individual ij. In this case *a* is said to be an *individual concept*. If referent(a)=* a is said to be a generic concept.

There is different notation of GCs, a linear and a graphical notation. More over there is a logical semantic associated with GCs. Each conceptual graph is associated with a formula in FOPL. Concept types correspond to unary predicates. For example, [**human** :*] *has exist (x) human(x)*, as associated semantic and [**human :"Burak"**], has human(Burak) as associated formulae. (individual markers correspond to logical constants, generic marker correspond to existentially quantified variables. *n*-ary relation correspond to *n*-ary predicates for example : [**human** :*] <---(owns)---> [**car** :*] *has exist(x), exist (y) (human(x) & car(y) & owns(x,y)* as logical semantic.

GCs permits to define lambda abstraction a way to define types from existing ones. Generally new types defined by lambda abstractions use the Aristotelian way to define types. They can be used to abstract CGs to have a general view or to expand an existing CGs by replacing a defined type by its associated lambda abstraction.

Two operations expansion and contraction consist in replacing a defined type in a graph by the body of its definition in its associated lambda abstraction for expansion and replacing the body of the definition of a defined type by its name as it is defined in its lambda abstraction.

For example, if we have a type person in the type set we can define the type Turkish learner as a person which learns Turkish by the following lambda abstraction:

Define type Turkish_learner (x) is
[**human** :*]←(agt)→[**to learn** :*] ← (obj) → [language : "**Turkish**"] ;

3 Results

In this section, we try to focus on the examples to demonstrate what the theoretical methods cited on the previous sections mean in the sentence and how these representation forms convert to an analysis. We examine our methods predicate logic and conceptual graph on Turkish and English sentences.

Our supports are linguistic differences between Turkish and English languages and the hypothesis that it would be more accurate to use different methods while identifying the computer based representation form of these two languages.

In a typical Turkish sentence, the verb includes all the information about the subject, time and other qualities. For this reason, a representation model which describes the entities with process logic should be chosen. When we evaluate the verb in the sentence concept, all this information set must be handled with the reification approach in order to concretize this discrete concept with qualitative properties as well as to cite all the on-going attributes (e.g. subject, complement, time and location) hidden in this verb.

For example; when we evaluate the English sentence "John eats an apple" with predicate logic, the representation model will be:

$$eat(John,apple)$$

If we evaluate the same sentence with reification method, with the basic logical operators, the representation model become as follows:

$$\exists m\ (fact\_of\_eating(m) \wedge agent(m,John) \wedge object(m,apple))$$

m: A specific event → The fact that Jean eats

If we evaluate the Turkish translation of the same sentence, “John bir elma yiyor”, with the same method, the output will be

yemek(John,elma) (*)

As in this example, the linguistic differences between Turkish and English are not very recognizable; the results are similar to each other. However, when we take into consideration the dominant role of the verb in Turkish sentence, it is obvious that Turkish is closer to be evaluating with reification method. The verbal “X”-me/ma belonging to the verb “X”-mek/mak can be done in English with the transformation: the verb “to X” → the fact of “X”-ing.

Thanks to the existence of these verbals in Turkish language, we obtained the following form (**) if we passed from the previous example (*) to a reified representation format:

$\exists m\ (yeme(m) \wedge agent(m,Jean) \wedge object(m,elma))\ (**)$

The analysis realized on the previous example demonstrate that both predicate and reification representation method can be used for Turkish and English.

However, it may be argued that Turkish is more advantageous in terms of reification comparing to English, especially in the following cases:

- The sentences where the subject is changed directly or indirectly
- Turkish sentence where the subject is indefinite while the subject is very explicit in English translation of the same sentence.

3.1 Gerunds

Gerunds are verbal constructions which in Turkish are often used in order to replace the relative and completive proposition. These gerunds may mention more or less complete form of the verbal construction such as person. They can have several grammatical functions and play the roles of nouns, adjectives or adverbs.

The following examples of gerunds introduce briefly our systematic approach by comparing their respective translations into English. The logical and reified logical versions were given as follows;

Araba **sürdüğünde** cebi çaldı. While he was driving his mobile phone rang

∃p1,p2 (fact_of_ringing(p1)∧agent(p1,mobile)∧time(p,t1)) ∧

(fact_of_driving(p2)∧agent(p2,he)∧time(p2,t1)): fact_of_ringing (mobile,time) ∧

fact_of_driving (he,car)

∃p1,p2 (çalma(p1)∧agent(p1,mobile)∧zaman(p,t1)) ∧

(sürme(p2)∧agent(p2,he)∧time(p2,t1)): çalma (cep,zaman) ∧ sürme (o,araba)

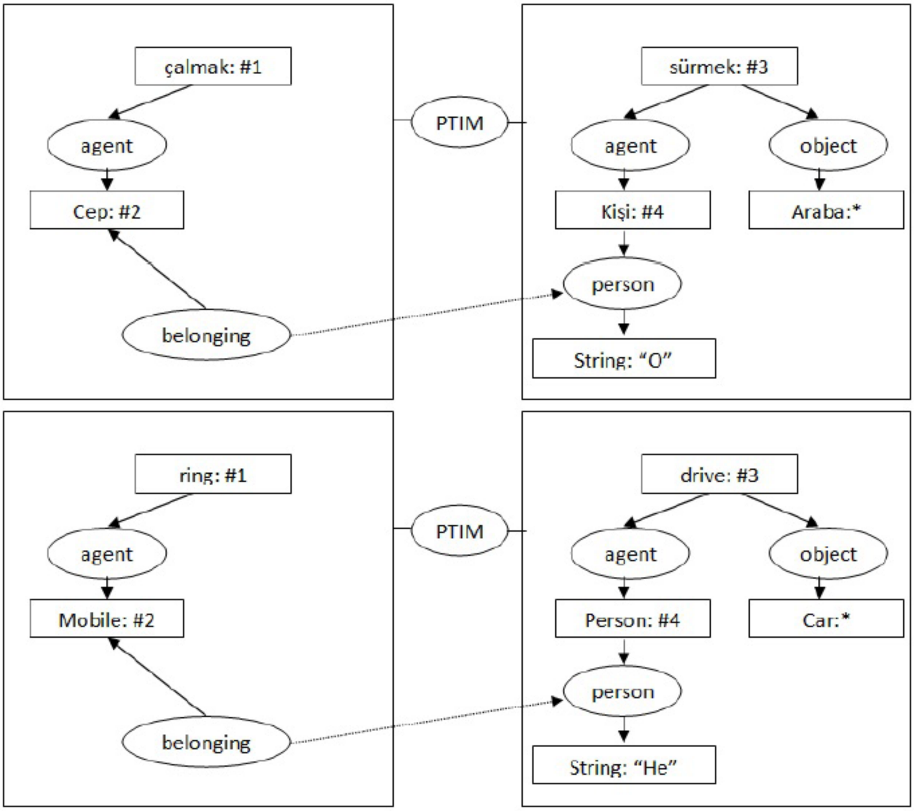

Fig. 1. Representations of *Arabasürdüğündecebi çaldı. While he was driving his mobile phone rang* with conceptual graphs where PTIM represents the concept of time.

Geldiğin zaman çayımı içiyordum. When you arrived I was drinking tea

∃p1,p2 (fact_of_arriving(p1)∧agent(p1,you)∧time(p,t1)) ∧ (fact_of_drinking(p2)∧agent(p2,I)∧object(p2,tea)): fact_of_arriving (you,time) ∧ fact_of_drinking (I,tea)

∃p1,p2 (gelme(p1)∧agent(p1,sen)∧zaman(p,t1)) ∧ (içme(p2)∧agent(p2,ben)∧object(p2,çay)): gelme (sen,zaman) ∧ içme (ben,çay)

3.2 Verbal Nouns and Noun Clauses and Moods

As gerunds these verbal nouns may play the role of relative and completive propositions and can also occur as epithets. The following examples illustrate the comparison between Turkish and English using the similar methodology.

Due to the lack of auxiliary (to be, to have) Turkish language uses noun clauses with the particles (var/yok≈there is/not)

1 Usage of indefinite subject: (The possession association)

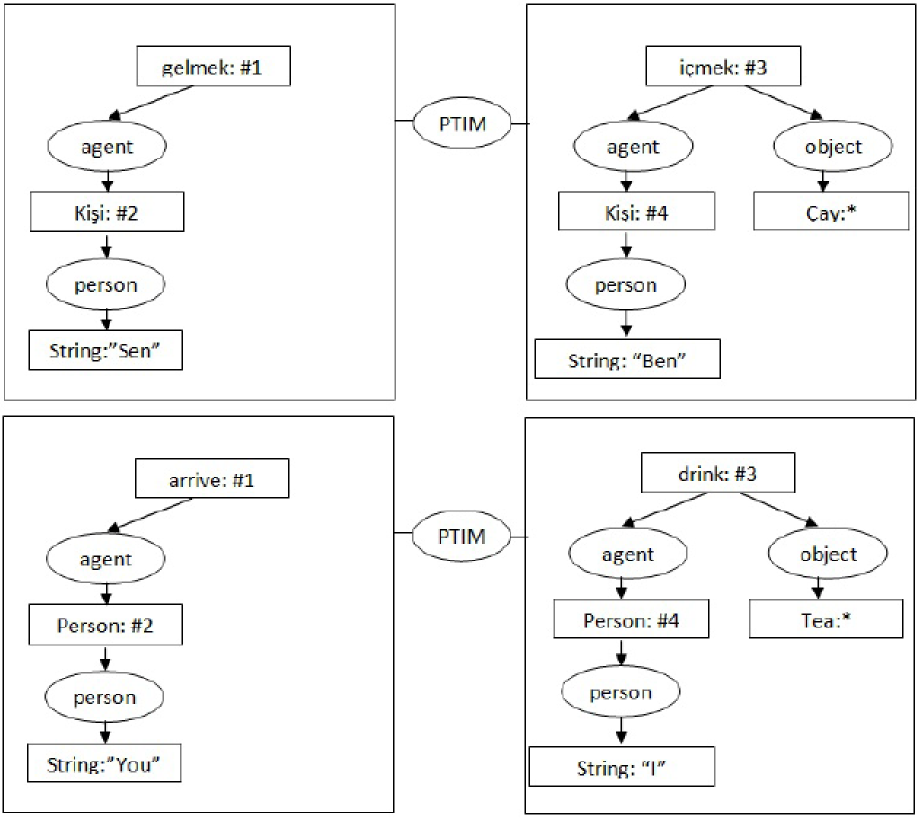

Fig. 2. Representations of ***Geldiğin zaman*** *çayımı içiyordum. When you arrived I was drinking tea* with conceptual graphs where PTIM represents the concept of time.

Mehmet'in arabası var → *Mehmet has a car (or There is a car of Mehmet)*

∃p(fact_of_having(p)∧agent(p,Mehmet)∧objet(p,car)) : fact_of_having (Mehmet,car)

∃p(sahip_olma(p)∧agent(p,Mehmet)∧ object(p,araba)) : sahip_olma (Mehmet,araba)

2 Usage of subject transformation: English and Turkish thematization greatly differs and this results in having different ways in expliciting agents of actions.
Canım yapmak ist(em)iyor ⇔ *I do not want to do*

∃ p (fact_of_wanting(p) ∧ agent(p,je) ∧ action(p,faire)) : fact_of_wanting (me,do)

∃ p (isteme(p) ∧ agent(p,can) ∧ action(p,yapmak) ∧ belonging(can,ben)): wanting (can,yapmak) ∧ belonging(can,ben)

Başım ağrıyor ⇔ *I have a headache*

∃ p (fact_of_having_ache(p) ∧ agent(p,me) ∧ location(p,head)) : fact_of_having_ache (me, head)

∃ p (ağrıma(p) ∧ agent(p, baş) ∧ belonging(baş,ben)) : aching (baş)∧ belonging (baş,ben)

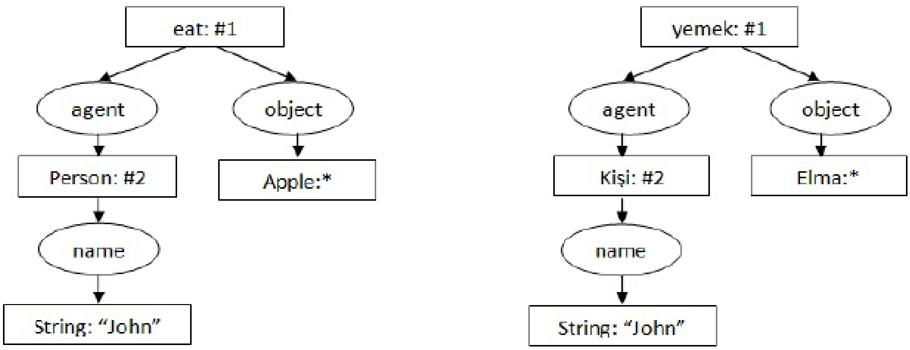

Fig. 3. Representations of *John eats an apple-Jean bir elma yiyor* with conceptual graphs

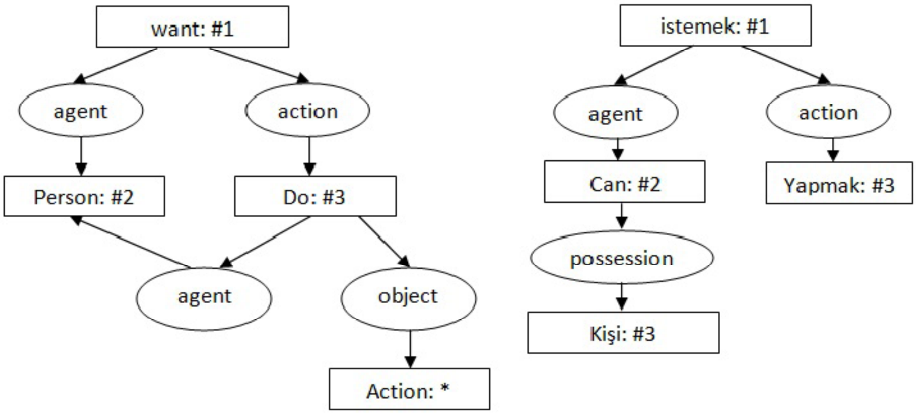

Fig. 4. Representations of *I want to do-Canım yapmak istiyor* with conceptual graphs

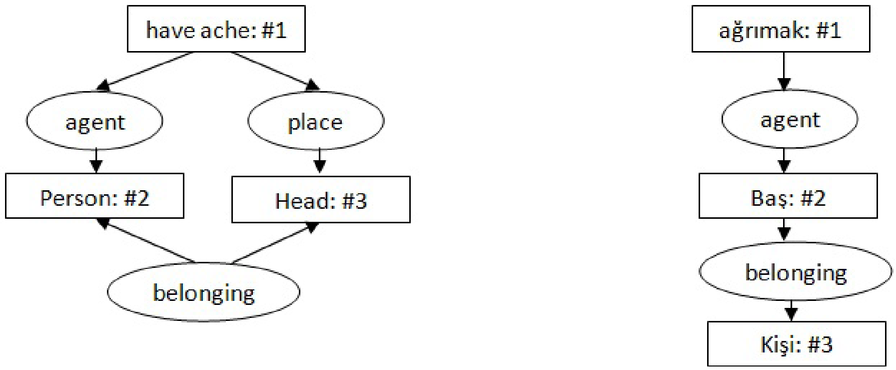

Fig. 5. Representations of *I have an headache - Başım ağrıyor* with conceptual graphs

Geç **kalanlar** sınıfa girdiler. The persons who arrived in late entered into the class-room

∃p1,p2 (fact_of_entering(p1)∧agent(p1,persons)∧location(p1,classroom)) ∧ (fact_of_arriving (p2)∧agent(p2,persons)): fact_of_entering (persons,classroom) ∧ fact_of_arriving (they)

∃p1,p2 (girme(p1)∧agent(p1,kişiler)∧location(p1,sınıf)) ∧ (geç_kalma(p2)∧agent(p2,kişiler)): girme (kişiler,sınıf) ∧ geç_kalma(kişiler)

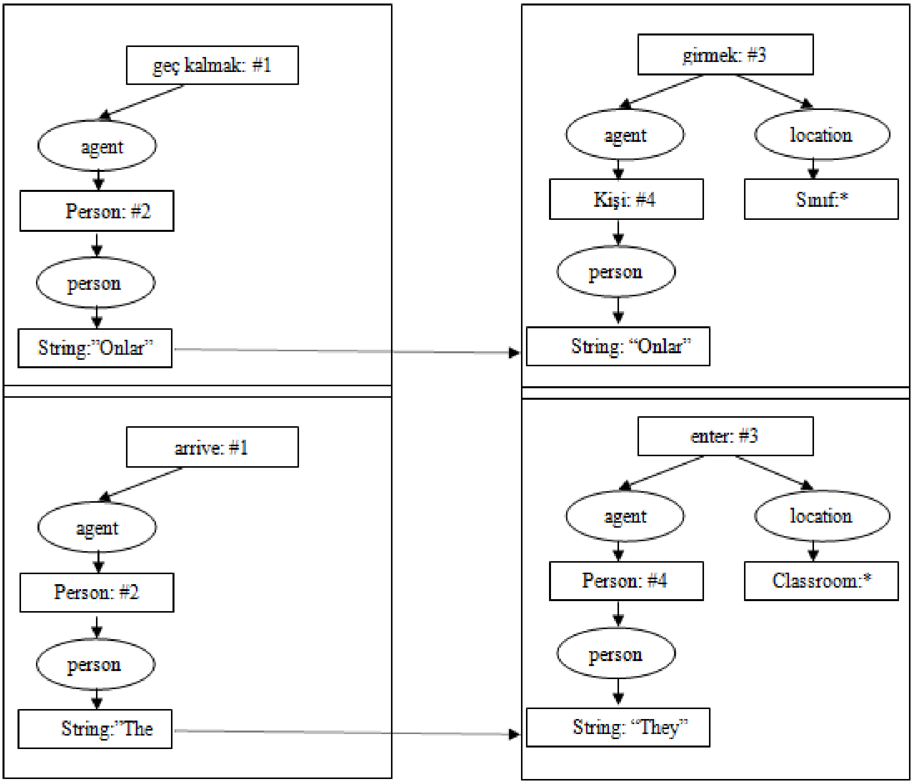

Fig. 6. Representations of Geç **kalanlar** sınıfa girdiler. The persons who arrived in late entered into the classroom with conceptual graphs.

4 Conclusion

In this study, we try to develop a new approach if the representation methods used in linguistic and NLP domains are suitable for Turkish and English which may be generally accepted as equivalent in terms of language representation, in an automatic analysis environment.

We concretize the aim of the study with the utilization differences of the subject in Turkish and English sentences. In an English sentence, the subject which is found at the beginning of the sentence is emphasized significantly, whereas in the Turkish sentence which has the same meaning, the subject will be hidden, and may either be a null subject or may be expressed with an indefinite person.

This property becomes more explicit when a predicate logic representation for English and reification representation for Turkish are preferred and transferred to automatic analysis.

Finally, we indicate how the texts can be transformed to automatic analysis for NLP by applying conceptual graphs to our examples. This graph based representation method barely known may reify naturally the sentence and so the text and render the analysis possible.

References

Sowa, J.F.: Conceptual Structures: Information processing in mind and machine. Addison-Wesley (1984)

Sowa, J.F.: Knowledge representation: logical. MIT Press (2000)

Findler, N.V.: Associative networks: The representation and use of knowledge by computers. Academic Press (1979)

Collins, A.M., Quillian, M.R.: Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior 8, 240–247 (1969)