Kunnskap for en bedre verden

Faculty of Natural Science
Department of Chemical Engineering

TKP4555 - Process Systems Engineering
Advanced Process Simulation

Final Report: Introduction to Ontology Application on Linguistic Studies

Written By:

Huang (Patrick) HUANG

Supervisors:

Prof. Heinz Preisig
Dr. John Morud

Abstract

This report presents a result of Ontology Technology on Linguistic Studies, the specific topic from the Advanced Process Simulation module of TKP4555 - Process Systems Engineering, Specialization Course. In this report, some essential background knowledge, including the description of involved disciplines, a review of former studies will be mentioned. A brief analysis is based on those knowledge, in order to explain the application. Besides, there is a further discussion with technical difficulties of the studies, some criticism of language improvements and a brief overview of the future perspective. Generally speaking, the ontology analysis provides a different viewpoint of linguistic studies. The greater use of ontology technology leads to the linguistic subject more scientific (apart from humanities), which is also beneficial for a more quantified and statistical linguistic studies.

1. Introduction

The topic of this report is "Ontology Application on Linguistic Studies", which is highly interdisciplinary, and this makes the description very complex. For instance, the term "ontology" has meaning in philosophy and information science, but this report will use both of them. Besides, the modern linguistic studies has several different sub-disciplines, and they have a huge different from each other, and the key idea of some other subjects, such as statistics, semiotic and cryptology may be used for linguistic analysis. In addition, some previous researches are already in an interdisciplinary background, this also increases the complexity, but some relevant information will be mentioned in the below content.

On the other hand, more and more interdisciplinary topics are merged nowadays. This topic becomes more potential for scholars with different backgrounds to cooperate, as a topic without full development, it is relatively easier to get more research finding.

Hence, this report aims to provide a comprehensive background knowledge, with putting forward some hypothesis to explain, as well as some weakness point which could be improved. This may help for further researches.

2. Description of Ontology

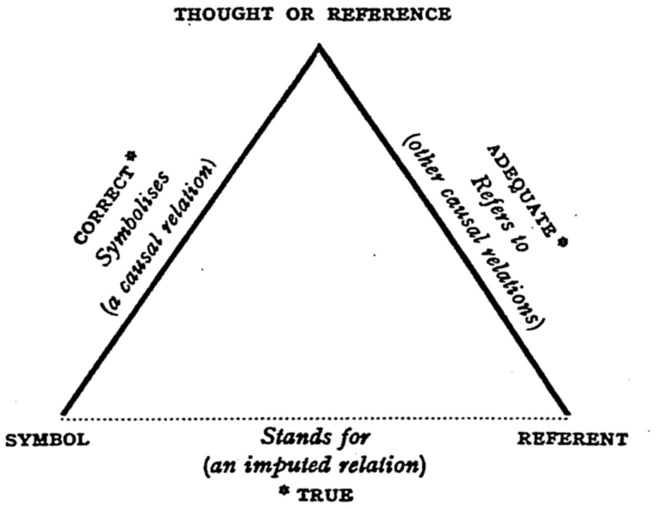
2.1. Ontology in Philosophy

Ontology is originally a philosophical term, and especially been widely used in analytical philosophy (Smith, 1998), it concern about the nature of being, existence and reality. In this report, one of the key focus is, whether a term is referring to something really exists, or just representing a concept.

This acts as a guiding ideology of the information science aspect and the relevant linguistic research. The most persuasive explanation is made by Russell called Theory of Descriptions. This inspires the further research of ontology database, and will be discussed more in section 2.3.

2.2. Ontology in Information Science

In information science, the term ontology stands for the specification of a conceptualization (Gruber, 1993). This accurate definition is stating that, the field should be specific, and there is an accurate formal representation of the relationship between each concepts. For the purpose of data analysis in computer science, the concepts and their relations should be unique, at least with less ambiguity for an accurate analysis.

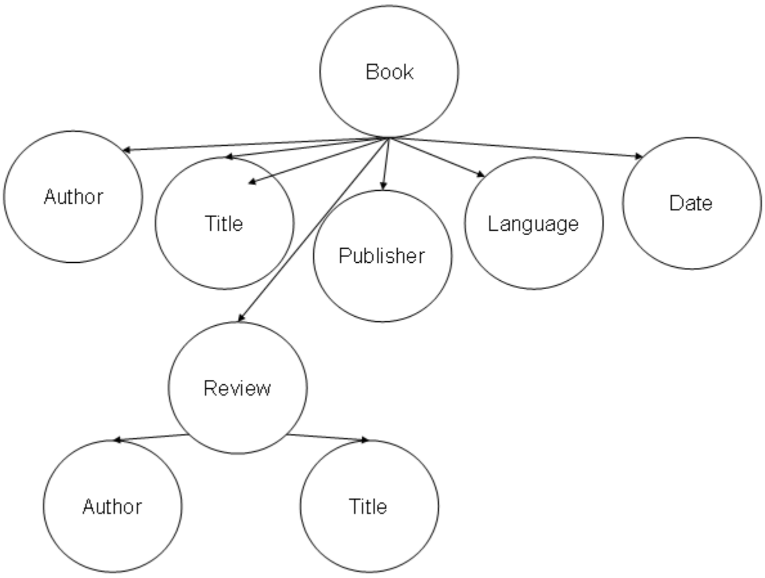


(Figure 1. Triangle of Reference)

Figure 1 called the Triangle of Reference (Ogden & Richards, 1923), which shows the relationship between symbol, reference and referent. It is the simple model which can be used directly in linguistic studies, to describe the word used and the actual substance or concept that the word mentioned about. This figure will compared with the Saussure Linguistics in section 3.1.

2.3. Ontology Database

For simplify of further research, an ontology database will be necessary. There are various possible species of database construction, but here will only discuss some recommended types.



(Figure 2. A simple ontology database)

Figure 2 shows a simple instance of ontology database (Pharo, 2014). It can be either visualize as what the figure shows, or presented by text in a systematic manner. This type of database is widely used, it can clearly show the logical relationship between each elements, but weak of presenting the meaning of each term, in other words, lack of explaining linguistics proposition.

Another form is the lexical referencing system, such as WordNet (Miller, 1995). It is similar to a dictionary, but under the Gruber's description of ontology (specification of a conceptualization), this type of database gives more details of each word, and with a systematic and constant way to describe, in order to make it easier to be "read" by computers. This type of database is helpful for the development of automatic translation between languages, furthermore, it is even useful for providing a database for construction of artificial intelligent model.

The WordNet structure is very suitable for this topic, because it is just designed for linguistic application. There are some very useful "synsets" of synonyms for each category (nouns, verbs etc.), each synset represents a basic concept and these sets are connected by different relationships. But a bit modification will be better for analysis in this report. It is, we can add several "tags" for description of each term, it is much easier to find out the similarities of different words, here is a simple instance:

[English: The - German: Der (Masculine); Die (Feminine); Das (Neuter)]

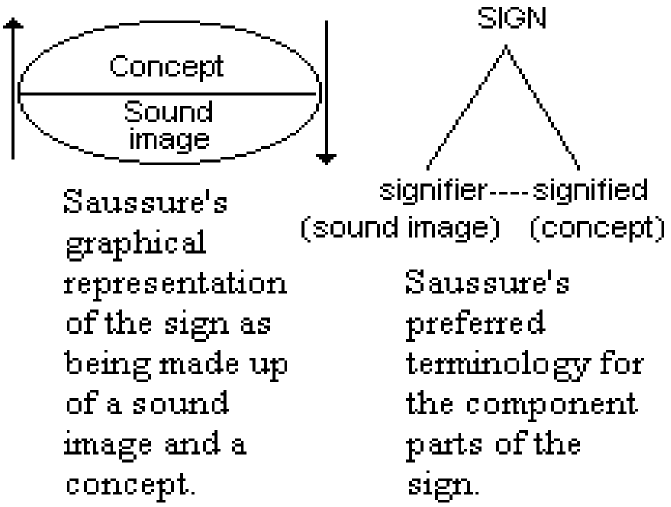
This is about the definite article translation between English and German. There is only "The" used in English, which means any definite article in German will become "The" when translating into English. But it is more complex when translate English to German. The word in bracket shows the basic tags in German. Those articles will be used in different situations (Masculine, Feminine and Neuter). Also we may put the one of those three tags in each German nouns. To determine which articles should be used for a specific noun, simply compare the tags between the article and noun. Sometimes one tag is not accuracy enough, and we need to add more tags to get uniqueness. With adding different tags in both nouns and articles, this could be a simple ontology database for definite article using in English to German translation.

3. Linguistic Studies

3.1. Saussure Linguistics

Ferdinand de Saussure is a Swiss linguist. Before his period, linguistic was mainly be regarded as a humanities subject, but Saussure invented the analytic methods for linguistic studies. His core idea is, linguistics is a type of science that based on meaning and symbol, which is also called semiotics or semiology (Nöth, 1995).

The main characteristic of Saussure semiotics is, the "sign" can be analysed as signifier and signified. Signifier stands for the sound or spelling, and signified stands for the concept and meaning (Chandler, 2007). The relationship between signifier and signified is arbitrariness, which means no relation at all.


Concept
Sound image
Saussure's graphical representation of the sign as being made up of a sound image and a concept.
SIGN
signifier----signified
(sound image) (concept)
Saussure's preferred terminology for the component parts of the sign.

(Figure 3. Relationship of component parts of sign)

Figure 3 mainly shows the relationship between signifier and signified (Chazelle, n.d.). The main idea of this triangular graph is very similar to the Triangle of Reference in Figure 1. This shows the high similarities and compatibility between semiotic and ontology analysis.

3.2. Chomsky's Universal Grammar Theory

Noam Chomsky is an American linguist. He is the main supporter of the Universal Grammar Theory. Combined with linguistic and cognitive science researches, Chomsky and DiNozzi (1972) believe that, the studying process of different languages are common. In other words, children will follow the same rule to study different languages.

This also comes to an interesting inference, since people will follow the same rule for studying different languages, therefore it is sufficient that to study language just by the universal grammar, without touching with the specific cultural background. This idea also leads to several arguments with other linguists. In this report, the result and the idea of ontology database conforms Chomsky's theory. With enough descriptions or "tags" in ontology databases, it is possible to distinguish any slightly difference use of language due to cultural differences. Such as quantified by different situations and degrees.

3.3. Quantitative Linguistics

Quantitative linguistics is a branch of linguistic, it mainly uses mathematical statistics to describe and research the nature language. Because of this, databases are frequently used in this field. In addition, it is highly possible to explore more by quantitative linguistic methods, after some ontology databases are fully built.

Martin's Law is one of the representative result in quantitative linguistic researches (Altmann, 1993). It show that, there are several lexical chains obtained by looking up the definition of a word in dictionary, then looking up the definition of the definition obtained and so on. Those lexical chains are with different "levels", which forms a hierarchy of more and more general meanings. The level will decrease with increasing generality of the lexical chain. Interestingly, the ontology analysis can reasonably explain this law. Further discussions are in section 5.3.

4. Philosophical Background

4.1. Kant's Identity Issue

Identification is the thing been discussed in Aristotle's period. Some typical considerations are proposed by philosophers, the most famous one is Descartes' "Cogito ergo sum" (which means I think, therefore I exist). But Descartes did not discuss further about the identity issue.

Later on, the German philosopher Immanuel Kant had put forward the issue of identity (Kitcher, 1982). In this report, the expanded definition will be used. The original definition is about self-identification, but the main body could be something else, such as words discussed in ontology analysis.

For example, "Kant is Kant" is an analytic proposition, it is logically true, but also meaningless.

On the contrary, "Kant is the author of 'Critique of Pure Reason'" is a synthetic proposition. It is not possible to judge its authenticity by logic, therefore relevant knowledge or experience need to be used to determine. We are mainly concern about answering these types of synthetic proposition.

4.2. Russell's Theory of Descriptions

Kant's Identity Issue makes some problems of the Law of Identity. As an example, "Kant is the author of 'Critique of Pure Reason'", which means "Kant" and "the author of 'Critique of Pure Reason'" has the same identity, and they can replace each other without changing the meaning. However, "Kant is Kant" becomes something meaningless. This problem was finally solved by the British philosopher Bertrand Russell.

Russell's Theory of Description considered that, the identity and description are something separated. In the above case, "Kant" is the identity and "the author of 'Critique of Pure Reason'" is the description. With further subdivision, the descriptions could be divided into definite description (pointed to the only identity, usually started with "the") and indefinite description (not pointed to only identity, usually started with "a" or "an"). This gives a possible solution to the Law of Identity problem.

5. Analysis of Ontology Databases

5.1. Use of Russell's Theory

Combined with the application of ontology database and Russell's Theory of Descriptions, the "tags" for word explanation can be series of descriptions. This may bring several advantages: with the shorter descriptions, the database can be simplified and the computer is easier to catch key words. Also, this helps the automatic translator much easier to replace synonyms.

Russell's theory is highly systematic, therefore a database following the rule of descriptions will be useful for maintenance and improvement of database. Since Russell's theory is about analytical philosophy, the database may be useful in this field as well.

5.2. Hypothesis of Initial State

There are very limited information about the proto-language, but since all languages are developed by communication of different language users, and the vocabulary is increasing and grammar is more systematic, it is reasonable to simulate proto-language by pidgin and creole language.

Pidgin language is the very beginning state of a new language between two or more groups without common language (Kaye & Tosco, 2001), the main characteristic of pidgin language is lack of vocabulary and the grammar regulation are reduced to almost minimum. Pidgin language will develop and become a creole language, while their offspring studies the pidgin language as a mother tongue. After further development such as social isolation, creole language is possible to "develop" from the usual oral language to a formal written language.

Although there is no recent creole language presents more formally than its parent language, but it is reasonable to suppose the ancient language had that process, such as Spanish was originally mixed by Latin language and the Iberian local language, it is possible that Spanish had the pidgin and creole period, then finally become a formal and even literature language.

5.3. Hypothesis of Language Development

With the development of pidgin language, there is something worth mentioning. Because of the lack of vocabulary, the language user can only create a new word, or borrow more words from another language, or using existing words to explain some new substances. Pidgin user will mainly use the third method at the beginning. Here will pick the pidgin language in Papua New Guinea called Tok Pisin as an example.

According to The Guardian (2012), the Tok Pisin language called accordion as "liklik box you pull him he cry you push him he cry", beard is called "grass belong face", and a very thin person is called "bone nothing".

It is obvious that in some words, such as "liklik box you pull him he cry you push him he cry" for representing accordion is definitely too long, and will be probably be simplified after this word was used more frequently. But "bone nothing" for thin person is quite accurate and with a suitable length, it can possibly become a formal word. Also, sometimes the word may not be used that frequently, but becomes a twister or proverb (with some specific connotation) or a slang (becomes very informal). Generally, those words are possibly developed in different directions, for a more simplified and clear use.

5.4. Hypothesis of Description Error

In another case, some words may have big ambiguity, such as "grass belong face". With analysis of ontology and theory of descriptions, it is easy to tidy up this following database:

[Grass: green coloured, strip shape, will grow, organism, etc.]
[Unknown substance (beard): black coloured, strip shape, will grow, etc.]
[Face includes: "eye", "nose", "mouth", etc.]

According to the database, grass and beard are similar in some aspects, and their similarities is higher than grass and eye, grass and nose etc. So the word "grass belong face" can be used is because of the listener can basically analysed what is the signified of the word, without mistakenly think of eye or nose or something else instead.

However, this term still has some ambiguity. For example, "grass belong face" can be eyebrow also. But this ambiguity is not that big, so it can still be basically used. This phenomenon comes to this following hypothesis.

With explaining a new substance, the only two methods are either introducing a new word (no matter creating a new one or borrow one from another language) or represent as a description by existing words. Think of human being as a computer, introducing new words occupies more RAM and the language user needs to use more time to remember new terms, using existing words occupies more CPU because of the necessity of analysis, also some errors will occur due to the difference between identity and component of description.

We can assume an extreme case: there is a pure artificial language, and every substance requires a unique word to describe, it will be highly accurate but unrealistic since the user has to remember unlimited words. In contrast, with least word and represent all signified by those words, this leads to the maximum of error, in other words, unrealistic because of the huge ambiguity. So the ideal case of language development should be keeping the balance in the middle. Finding shorter words to present identities (such as "liklik box" instead of "liklik box you pull him he cry you push him he cry"), and searching for better combination for a less error are recommended also (use "grass belong face" but not "something belong face").

6. Discussions

6.1. Technical Problems

There are several technical problems found during report writing, solving these problem could help a lot for further studies.

Firstly, lack of database. There are a lot of useful data from traditional linguistic research, but it is not managed in a computer science manner, so it is not readable by computers for mathematical statistics analysis. However, the database for computer analysis is not enough, so many logical result cannot be proved by data and therefore can be only considered as hypothesis.

Another strange thing is, there are several ontology databases for linguistics, but most of them are deadlink. The reason is still unexplained, especially some databases are developed by world-class universities.

Also, language can be viewed as a system but quite special. It cannot be quantified or linearized (possibly in the future), this requires a detailed descriptive database, and we cannot find any equations or functions for modeling. As a study with simulation basis, some simulations of language is easy to cause unexpected error, such as using pidgin language to "simulate" proto-language. We had few knowledge about proto-language, so it is currently not possible to compare how similar they are (although they should be similar logically).

6.2. *Simplification of Languages*

Natural languages are always developing, many of them are still in a complex state. Therefore, with some improvement and simplification, the language can be easier to studied and more convenient for propagation. Such as, the Hypothesis of Description Error could be used to analyse English language.

There is a main character of English: because of the historical reason, there are a lot of borrowed words from different cultures. In this time, Chinese language will be given as a reference point. The below is a brief example:

English: June, Chinese: 六月 (six month).
English: Pork, Chinese: 猪肉 (pork meat).
English: Diabetes, Chinese: 糖尿病 (sugar urine disease).

There are a lot of similar cases. I think this is one of the main disadvantage of English. In both English and Chinese, "six" and "month" are basic words, and even an English speaker can understand the term "six month", because it is simple and almost without ambiguity. However, people knows "six" and "month" cannot understand what is "June", and they need remember the extra word "June" with the corresponding meaning: the sixth month. From a practical point of view, those extra words are highly useless and can be simplified by some improvement.

6.3. *Artificial Languages*

This report had assumed a "pure artificial language" that describe all substances in different words. This ideal concept is only set for a reference point of a extreme situation. Although with careful consideration, the "real" artificial language such as Esperanto is not discussed in this report. It is valuable to consider, but since the situation are different, it is probably without effective result when comparing with natural language. Setting another artificial language for further simulation may be an useful idea.

7. Conclusion and Perspectives

In conclusion, with studying an interdisciplinary topic, this report focused on keeping balance of introducing background knowledge, analysis, hypothesis and suggestion. This future research could be very flexible, since this topic can be developed freely in different directions. There are several perspectives mentioned below, and they will be possible after the built of ontology databases.

For example, some basic definitions may be re-explained in an ontology view, such as the "literature level" of a language. Some languages are rich in literature but some are purely oral and there is no mature poetry or other literature form developed. This may partly quantified by the number of "useless words", since more identities and more thesaurus give chance to more delicate expression. Besides, additional quantified "indicators" may be found by future studies, which is similar to quantitative linguistic research.

With the time passing, immigration and communication with other ethnic group, a language family had changes chronologically and gradually change with longitude and latitude. Such as Germanic languages, because of the geographic difference, there may be some regular changes from Switzerland to Netherland, and it can be measured by some detailed databases.

According to the hypothesis, all languages are started from a young state and develop and become "mature". The way of development varies, but some "immature" component was abandoned and some "mature" components are developed. Since these are possible to be found and quantified, we can assess each language's "degree of maturity", and also gives suggestions for the language's further improvement direction.

Finally in one sentence: with the development of ontology, the view of linguistic studies could be completely changed.

Reference

Altmann, G. (1993). *Science and linguistics. In Contributions to quantitative linguistics (pp. 3-10)*. Springer Netherlands.

Chandler, D. (2007). *Semiotics: the basics*. Routledge. Chicago

Chazelle, B. (n.d.). *Introductory models & basic concepts: semiotics*. [ONLINE] Available at: http://www.cs.princeton.edu/~chazelle/courses/BIB/semio1.html.

Chomsky, N., R, DiNozzi. (1972). *Language and mind*. Harcourt Brace Jovanovich. New York

Cooke, H.P. (1983) *Aristotle Categories Vol. 1*. Harvard University Press. Cambridge.

George A. Miller (1995). *WordNet: A Lexical Database for English*. Communications of the ACM Vol. 38, No. 11: 39-41.

Gruber, T. R. (1993). *A translation approach to portable ontology specifications*. Knowledge acquisition, 5(2), 199-220.

Kaye, A. S., & Tosco, M. (2001). P*idgin and creole languages: A basic introduction (Vol. 5)*. Lincom Europa. Chicago

Kitcher, P. (1982). *Kant on self-identity*. The philosophical review, 41-72. Chicago

Nöth, W. (1995). *Handbook of semiotics*. Indiana University Press. Bloomington

Ogden, C. K., & Richards, I. A. (1923). *The meaning of meaning*. New York.

Pharo, N. 2014. *The semantic web*. [ONLINE] Available at: http://www.jbi.hio.no/bibin/dig_korg/sem_web.htm.

Reicher, M. (2010). *Nonexistent objects*. [ONLINE] Available at:

http://stanford.library.usyd.edu.au/entries/nonexistent-objects/

Smith, B. (1998). *An introduction to ontology*.

The Guardian. (2012). *Prince Charles in Papua New Guinea: how to speak pidgin English like a royal*. [ONLINE] Available at: http://www.theguardian.com/uk/shortcuts/2012/nov/05/prince-charles-papua-new-guinea