「最近のオントロジーについての考察」

□先日「言語処理学会」から届いた「自然言語処理」2008 年9 月号Volume15 Number4 に特集として掲載されていたオントロジー（Ontology）の構築・連携・利用について…一言。

冒頭の巻頭言を簡単にまとめると、

①辞書や辞典といわれているものは、形態素単位で順序列に整理されたもので、

②シソーラスは形態素単位で上位/下位概念体系でまとめられ、

③WordNet やEuroNetなどは形態素単位で意味的概念の関係でドメイン別にまとめられているものである。

④そして、オントロジーとは、形態素単位で階層的意味概念の関係でドメイン別共通項としてまとめられているが、

これだけでは文や文書などの意味解析を十分に処理できない…のは皆さんも判っている通りである。

その理由は、形態素だけをクラスタリングし、同義語や類義語、関連語あるいは分野別に分けたり、また階層化してシソーラス辞書のようなものを作っても、「文の構造」によって意味は大きく変化するため…である。形態素には意味素や祖語と呼ばれる形態素が複数含まれており、それらがその形態素を構成し、その上、それらの要素（基底）自体に曖昧性を含んでいる。国語辞書を考えれば判る。見出し語と呼ばれている単語を語釈文と言われている文で解説されているのが国語辞書だ。その語釈文の中には、また形態素が含まれており、その形態素を見出し語で調べると、また語釈文で解説がなされている。すなわち、形態素は、ある形態素で構成されているので、その構造を明確にすることが必要である。もうひとつの理由は、形態素と形態素を結ぶ形態素（特に助詞や助動詞などの付属語）により、すなわち係り受け関係により形態素どうしの意味的な変化が大きく成されるから文や文書の不確定さや曖昧さも発生し、文書全体としての文脈的な意味の変化に伴う一意性の問題などが生じ、オントロジーと言われている形態素を主体とした体系だけでは大きな問題が残ってしまう。だからといって…不必要であるとは思ってませんが、コーパスとしては使えます。

すなわち、言語の形態素とは、基底と云われるある形態素で構成され、その意味的な輪郭も曖昧で、数学的にいうと位相空間上の開集合で構成されているn次元位相多様体や特異複体の単体のようなものである。もちろん、その最小単位は文字そのもので、その文字の集合族が位相になり、位相である開集合で構成されているのが多様体や圏になる。多様体や圏だけでは意味解析処理するのに不便であるので、剰余類や商群により群化して計算し易くしたり、多様体に向き付けたり、茎や芽としての主部（主語）/述部（述語）+修飾部（修飾語）を抽出して、意味解析をし易くする。もちろん、文だけではなく、文章や文書によっても意味は変化するので、文字⊆形態素⊆文節⊆連文節⊆文⊆文章⊆文書という階層化された圏と層を単位とする列も定義できる。圏の定義には、射という係り受けの単位オブジェクトどうしの関係も自然に導かれる。あとは列の関係をホモロジーや双対関係であるコホモロジーで要約などを処理することも可能である。これらを数学的に厳密に体系的に定義する必要があり、新版として執筆中である。（第7版）

⇒ *cTag*＞意味位相空間ページへ