# Ontological and Terminological Commitments and the Discourse of Specialist Communities

**Khurshid Ahmad[1], Maria Teresa Musacchio[2], Giuseppe Palumbo[3]**

[1]School of Computing and Statistics
Trinity College Dublin
Dublin 2, Ireland
kahmad@cs.tcd.ie
[2]Dipartimento di Lingue e Letterature AngloGermaniche e Slave, Università di Padova
Via Beldomandi 1, 35137 Padova, Italy
mt.musacchio@unipd.it
[3]Dipartimento di Scienze del Linguaggio e della Cultura, Università di Modena e Reggio Emilia
Largo Sant'Eufemia 19, 41100 Modena, Italy
palumbo.giuseppe@unimo.it

**Abstract**

The paper presents a corpus-based study aimed at an analysis of ontological and terminological commitments in the discourse of specialist communities. The analyzed corpus contains the lectures delivered by the Nobel Prize winners in Physics and Economics. The analysis focuses on (a) the collocational use of automatically identified domain-specific terms and (b) a description of meta-discourse in the lectures. Candidate terms are extracted based on the *z-score* of frequency and weirdness. Compounds comprising these candidate terms are then identified using the ontology representation system *Protégé*. This method is then replicated to complete analysis by including an investigation of metadiscourse markers signalling how writers project themselves into their work.

## 1. Introduction

Discourse analysis now plays a major role in the research, teaching and learning of the language of specialist domains. There is increasing evidence that discourse analysts are moving from the traditional intuitive and hermeneutical analysis, based on hand-selected key-words and sentences in carefully selected texts (see, for example, Sinclair and Coulthard 1975) to a more empirical corpus-based study of spoken (Sinclair 1992) and written (Stubbs 1996) discourse. The discourse of an author, and its interpretation by the readers of the discourse, is regarded as an 'ideological question': what were the intellectual commitments of the author and those of the readers? What is it that the author does not quite believe in and will reinforce/contradict the beliefs of the readers? These questions are asked frequently in literary criticism and recently have been asked in applied linguistics literature: 'how writers project themselves into their work to manage their communicative intent' (Hyland 2000). Researchers interested in the question of intellectual commitment of specialists are increasingly to be found in the various areas of computing including ontological engineering, semantic web and information extraction (Sowa 2000, Maedche 2002).

In this paper we explore ways of finding (a) the ontological commitment, primarily through a discourse analysis of the specialist terms coined and used by specialists, and (b) how this commitment may or may not be projected in texts through metadiscourse analysis of the so-called discourse 'markers' that help an author to foreground/background his or her claims, and cite sympathetic members of his or her own community. This intellectual commitment is an ontological commitment to use domain-specific vocabulary or 'terminology' that is consistent with the theory specified by an ontology (Gruber 1993). In any specific domain, the objects that can be represented are referred to as universe of discourse. Thus, those who subscribe to a given ontology also make a terminological commitment to agree on the meaning of any term in that ontology (Kashyap 2004). In terminology a list of concepts constituting a domain is established, concepts are related logically and ontologically to one another, and a designation – that is, a term – is assigned to each concept in the domain (Cabré 1998). Finally, in discourse analysis a universe of discourse is a set of discourse elements that pertain to the beliefs, conventions, and knowledge shared by members of a sociolinguistic community. As can be seen, there is a clear overlap in terms and concepts derived from philosophy in ontological engineering, terminology and discourse analysis. We use this interdisciplinary common ground to analyse ontological and terminological commitments in the discourse of specialist communities.

## 2. Method and motivation

LSP studies – and it has to be said many a computational studies – of terminology are rooted deeply in the so-called Platonist tradition. For a Platonist, intellectual commitment in one sense relates to a belief in a set of abstract objects – whatever exists in the world of physics, economics, chemistry and so on, exists in the abstract and is transcendental to human sensory perception. Specialist knowledge comprises a set of sentences written by specialists reflecting the 'reality of [the] abstract objects' (Orenstein 1977). In discourse analysis the focus is on analysing a set of sample sentences from one or more texts, even corpora, and examining the communicative intent of the author(s). In computing, the focus again is on identifying a set of *true* sentences, based usually on the intuition of the computing researchers and true in the sense of Ornstein above, and the subsequent conversion of the true sentences into statements of logic. But there are criticisms of the Platonist approach in both LSP studies in particular and in scientific texts in general by philosophers like Quine (1969).

Our method can be summed up as follows:

*Choice of texts:* Research is conducted generally by independently-minded, self-motivated individuals. The texts produced by the researchers are usually peer-reviewed and successfully published papers have the *proxy* approval of the specialist community. We have chosen to use texts that have been approved by the widest possible section of the specialist community – the Nobel Lectures in Physics and Nobel Foundation-sponsored Bank of Sweden Prize in Economics. We have analysed all Nobel Lectures since the inception of the prize in Physics (1902-2004, 969515 tokens in 157 texts) and in Economics (1969-2004, 727658 tokens in 55 texts). The majority of texts in Economics are in American English; texts in Physics are in British and American English.

*Method of Analysis:* Our intention is to be as minimally subjective as possible. To this end, our text analysis is based on a contrast between the distribution of words in a corpus of specialist texts and in a *representative* general language corpus (e.g. the British National Corpus). The relative frequency of every word is computed in the specialist and general language corpus ($f_{Sp}$(word), $f_{Ge}$(word)) and the *weirdness* ratio is calculated (Ahmad 1995) using System Quirk. Subsequently, the *z-score* of frequency and weirdness is computed: all tokens with z-score greater than zero for both metrices are regarded as candidate terms. The candidates are used extensively to make compound terms. The candidate compound terms are then selected on the basis of statistics related to co-occurrence probabilities of the members of the compound. The relationships between the constituents of the various compound terms that are based on a single candidate term are *animated* using the ontology representation system *Protégé.*

*Results:* We have divided the lectures, where possible, in clusters of lectures given over a 10-year period; for Economics our last period only contains Lectures for five years (99-04). For each period between 10-20 candidate single terms were found that satisfied the z-score criterion mentioned above. For the sake of brevity we will only show results for three terms per period for Economics and show that similar ontological patterns are found in Physics as well. (The candidates were found after excluding proper nouns, symbols for units and dimensions and abbreviations).

## 3. Ontology and discourse analysis

The subject domain of economics is developing and is a relatively new comer to the Nobel Foundation. The term *theory* (comprising 0.34% of the 55 lectures) is a significant single term – and appears as the head in compounds like *equilibrium theory* and *rationality theory* in 1969-78 Lectures; then there is *endogenous growth theory* in the 1980's followed by *game theory* and *shock theory*. Table 1 shows the candidate terms and each candidate reflects its dominance during that decade.

Figure 1 shows the compounds formed by the single terms in Table 1. Each interval shows the dominance of a given theoretical paradigm in Economics – from the notions of *bounded rationality* in the 70's through to *endogenous growth theory* in the 1980's and onto *rational* and *Bayesian players* and *supergame dilemmas in* the 1990's. Then there is the turbulence of the 21st century where the concerns are about creating a framework to study *shocks* to the market that cause *daily variance* and the *volatility* in the market place.

Most of the terms you see in Figure 1 were coined and elaborated in the texts by the individual authors; some are yet to make it into a dictionary of economics and this is especially true of multiword terms.

Much the same is true of compound terms in Physics Lectures: here the ontological commitment is changing: first there was doubt *electrons* existed at all (there is reference to the *so-called electron theory*) in 1901-10 Lectures; in the next decade there was mention of *electron emission*, *electron conduction*, and the *outer electrons* that were projected to move around the nucleus (c. 1911-1920). In the next 50 years *electrons* dominate Physics Lectures – there are *electron beams* and *electron energy* (1961-70). But in the last decade of the 20th century, the frequency distribution of the single term *electron* is such that it fails our *z-score* test: a concept that has been consolidated perhaps? The answer appears to be positive in that *electrons* are now regarded to belong to the family of the so-called 'light particles' – the *leptons*. Figure 2 shows this changing commitment together with the emergence of new concepts like *laser*, the elusive *neutrino* (c.1991-99) that is a far cry from worrying about the questions about the existence of matter in *ray* form (c. 1901-10).

As far as patterning is concerned, we have focused on the three fundamental concepts in physics, *force*, *mass* and *energy*. The word "energy" is the key key-word of the corpus, both in absolute terms (it is the most frequent open-class word in the lectures) and when measured against a word list taken from a general language corpus such as the BNC. In terms of distribution, all three terms are evenly spread across the whole corpus. What is more interesting is the patterns these three words enter into and the way these patterns develop over time. The structures thus identified reflect the ontological commitments of physicists and, viewed in the context of quantitative data gleaned from the corpus, allow one to hypothesise and analyse how the subject is being shaped. Over time, and sometimes within single decades, we observe an ongoing process such that from any one of three terms in question and, in turn, from their derivatives, terms of increasing syntactic complexity are coined. The term "mass", for instance, can be seen in the course of one decade (1970-1980) to give rise to increasingly longer compounds such as *electromagnetic mass* and *quantum-mechanical electromagnetic mass*. Likewise, the term "force" can be seen to extend to *force field* and then to *cohesive force field* over the ten years going from 1961 to 1970. But perhaps (and not surprisingly, given its centrality in the field) the most productive of the three terms in question is, in this respect, the term "energy". Table 2 shows the persistence of some of its derivatives over the decades; it also indicates how these derivatives gradually give rise to a whole series of complex compounds, pointing to the conceptual refinement carried on by physicist on one of the crucial concepts in their field.

| 1969-1978 ($N_{69-78}$=115,115 ; $T_{69-78}$=15 ) | | 1979-1988 ($N_{79-88}$= 95,866; $T_{79-88}$=11 ) | | 1989-98 ($N_{89-98}$=153,480 ; $T_{89-98}$= 17) | | 1999-2004 ($N_{99-04}$=195,224 ; $T_{99-04}$=12 ) | |
|---|---|---|---|---|---|---|---|
| Token | Rel Freq (W) | Token | Rel Freq (W) | Token | Rel Freq (W) | Token | Rel Freq (W) |
| equilibrium | 0.10% (58) | growth | 0.35% (27) | games | 0.12% (19) | shocks | 0.04% (103) |
| rationality | 0.04% (71) | income | 0.32% (26) | player | 0.12% (22) | variance | 0.03% (45) |
| underdeveloped | 0.06% (543) | saving | 0.20% (83) | supergame | 0.04% (39113) | volatility | 0.07% (205) |

Table 1: Selection of candidate terms in Economics Lectures (*N* denotes total number of tokens in *T* texts for the period; *W* denotes the weirdness ratio and is given in parantheses).

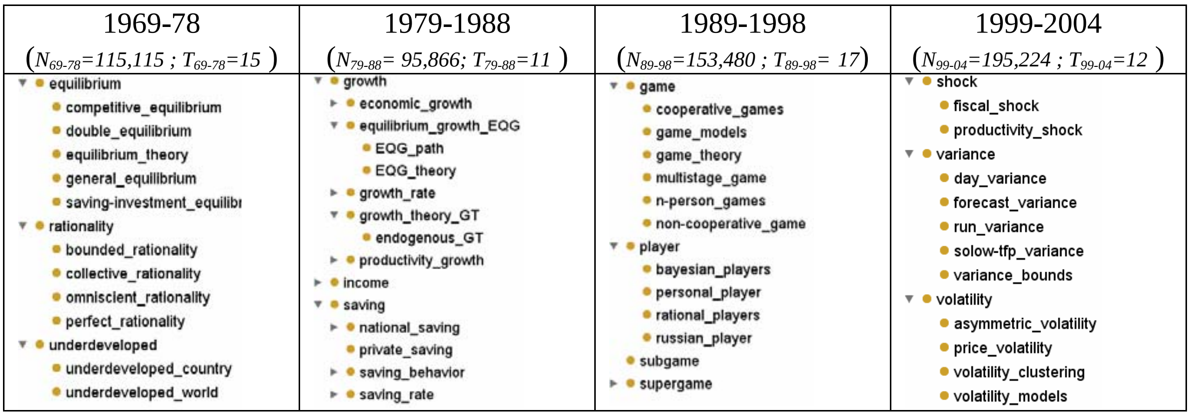


Figure 1: A hierarchy of terms that are amongst the most frequent significant collocates during a given time period.

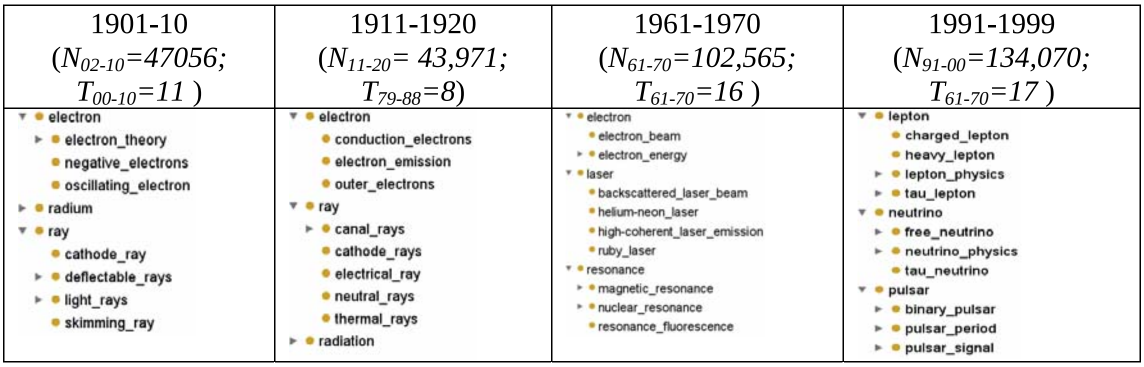


Figure 2: Protégé-generated diagrams for significant collocates of single terms in the Physics Lectures.

| Decade | Patterns |
|---|---|
| 1911-1920 | kinetic energy |
| 1921-1930 | kinetic energy; energy levels |
| 1931-1940 | kinetic energy |
| 1951-1960 | kinetic energy; energy levels, Bohr's energy levels, Zeeman energy levels |
| 1971-1980 | kinetic energy, total kinetic energy, negative kinetic energy, average kinetic energy; energy levels, Mulliken's energy levels, populated energy levels, atomic energy levels, molecular energy levels, spin-split energy levels |

Table 2: occurrence of some derivatives of the term *energy* and their gradual extension into compounds over the decades.

A similar process of terminologization can be seen to operate with some of the proper names contained in our Nobel Physics lectures corpus. The 'star' name in this respect is that of Enrico Fermi. Thanks both to direct quotes and to the large frequency of eponymous terms in which it appears, the name Fermi is by far the most frequent in the corpus with 359 occurrences. Among the frequent compounds in which it is used are *fermi liquid* and *fermi energy*, which in turn give rise to such extended compounds as *fermi liquid interactions* and *metal fermi energy*.

Lexical-pattern identification in corpora can, in combination with the identification of discoursal patterns, give a more complete picture of the mechanism whereby language reflects how specialists shape a given subject. In particular, whereas discoursal patterns can be said to realize the *constitutive* function of language (in Hylands' [2000] terms), lexical patterns centred around terms and candidate terms can be taken to reflect the *representational* function of texts. In other words, they point to the concepts, and to the ways of elaborating those concepts, that at a given moment in time have more currency between the specialists of a field. Together with measures of frequency distribution, patterns can help to trace the lexical and conceptual 'infrastructure' of a given subject domain.

## 4. Metadiscourse analysis

We next apply our method of analysis to investigate how commitment may or may not be projected in texts through metadiscourse analysis of the so-called discourse 'markers' that help an author to foreground/background his or her claims, and cite sympathetic members of his or her own community. Our working hypothesis is that the Nobel Prize gives scientists such 'visibility' that their position within their discourse community should change and they should be not be so guarded in expressing their ideas. We follow Hyland (2000) in identifying two types of metadiscourse markers, textual markers and interpersonal markers. Textual markers are used to organise texts in a convincing way for the audience or readership and pre-empt attempts at criticism. Interpersonal markers show the author's perspective and his/her attitude towards the audience or readership. Textual markers fall into four categories: logical connectives (*and*, *but*, …), frame markers (*next*, *finally*, …), code glosses (*for example*, *in fact*), endophoric markers (*discussed below/above*) and evidentials (*cite*, *said*, *says*, …). Interpersonal markers include attitude markers (*admittedly*, *hopefully…*), hedges (*almost*, *might…*), emphatics or boosters (*definitely*, *undoubtedly…*), relational markers (*incidentally*, *consider…*), and person markers (*I*, *we…*). In this paper we show the importance of terminological resources and of thesauri of metadiscourse markers for the study of the writers' intellectual commitment.

Again, our analysis is based on comparing and contrasting the distribution of words in a corpus of specialist texts and in a *representative* general language corpus (the British National Corpus). The overall relative frequency of textual and interpersonal markers is calculated in the Nobel Lectures in physics and economics. We then compare our data with Hyland's findings. Nobel Lectures – both in physics and in economics – are still analysed in clusters of 10-year periods with the exception of our last period in Economics lectures which only covers five years. For each period we look at the most frequent markers. For the sake of brevity we will only show results for 9 logical connectives and indicate how consistent their weirdness is in Physics and Economics compared to our representative general language corpus.

Table 3 shows metadiscourse in Nobel Lectures in Economics and Physics compared and contrasted with Hyland's findings as to subjects with similar frequencies and overall with reference to eight subjects (biology, physics, mechanical engineering, electronic engineering, marketing, sociology and philosophy). The prevalence of textual markers over intertextual ones points to a strategy for guiding the reader in the hybrid text type of the Nobel lectures whose language is more transparent than academic journal articles when they are delivered but is then aimed at the peer community as the only audience in a position to understand fully the Nobel Prize winner's contribution to his/her discipline. The contrast in text composition is quite clear in the subject we can compare directly, that is physics. Nobel lectures in physics have 46.27 textual markers and 39.04 interpersonal markers per 1,000 words, physics textbooks have 40.6 and 19.2 respectively. Hyland (2000) has looked at metadiscourse in science and engineering textbooks and has found a 'cline' from philosophy, with high use of textual and interpersonal markers, to electronic engineering, with low use of textual and interpersonal markers. As can be seen from Table 3, Nobel Lectures are closer to philosophy than physics or electronic engineering.

In Table 4 we investigate the frequency of a type of textual markers, logical connectives, and their weirdness if compared with the British National Corpus – used as representative of contemporary English. Frequency and weirdness vary over the four 10-year periods we consider.

| **Marker** | Economics (Nobel Lectures) | Physics (Nobel Lectures) | Physics | Philosophy | Electronic Engineering | Hyland (8 subjects: Bio; Phys; Mech.Eng.; Elec.Eng; Mkt; AL; Soc; Phil) |
|---|---|---|---|---|---|---|
| Textual | 50.72 | 46.27 | 40.6 | 41.5 | 37.4 | 59.4 |
| Interpersonal | 37.61 | 39.04 | 19.2 | 51.9 | 18.7 | 40.6 |

Table 3: Textual and interpersonal markers in Nobel Lectures and Hyland's analysis of textbooks (per 1,000 words).

| **Token** | **1969-1978** | | **1979-1988** | | **1989-1998** | | **1999-2004** | |
|---|---|---|---|---|---|---|---|---|
| | *Frequency* | *Weirdness* | *Frequency* | *Weirdness* | *Frequency* | *Weirdness* | *Frequency* | *Weirdness* |
| hence | 0,039% | 9.06 | 0,012% | 2.51 | 0,028% | 5.84 | 0,023% | 4.88 |
| moreover | 0,011% | 2.56 | 0,006% | 1.39 | 0,034% | 7.92 | 0,012% | 2.82 |
| consequently | 0,003% | 1,00 | 0,001% | 0.39 | 0,011% | 4.16 | 0,007% | 2.87 |
| thereby | 0,009% | 3.50 | 0,003% | 1.12 | 0,008% | 3.02 | 0,002% | 0.91 |
| similarly | 0,009% | 1.86 | 0,010% | 2.20 | 0,006% | 1.36 | 0,013% | 2.78 |
| and | 2,447% | 0.91 | 2,763% | 1.03 | 2,832% | 1.05 | 2,983% | 1.11 |
| because | 0,077% | 0.74 | 0,075% | 0.73 | 0,088% | 0.85 | 0,094% | 0.91 |
| nevertheless | 0,011% | 1.53 | 0,005% | 0.69 | 0,006% | 0.86 | 0,003% | 0.47 |
| but | 0,404% | 0.88 | 0,336% | 0.73 | 0,267% | 0.58 | 0,293% | 0.64 |

Table 4: Frequency and weirdness of logical connectives in the Nobel Economics corpus (1969-2004) to see how the Nobel Lectures are different from a representative sample of British English (the British National Corpus).

| **Token** | **1902-1910** | **1911-1920** | **1921-1930** | **1932-1939** | **1943-1950** | **1951-1960** | **1961-1970** | **1971-1980** | **1981-1990** | **1991-2000** | **2001-2004** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| hence | 9.81 | 1.88 | 5.17 | 2.07 | 0.00 | 7.44 | 7.26 | 2.51 | 1.47 | 3.28 | 3.40 |
| moreover | 0.00 | 0.00 | 3.71 | 2.30 | 7.51 | 1.94 | 0.43 | 0.97 | 0.81 | 1.36 | 1.77 |
| consequently | 0.00 | 3.55 | 4.58 | 1.30 | 0.00 | 7.83 | 1.84 | 0.94 | 2.54 | 1.54 | 0.75 |
| thereby | 3.13 | 4.22 | 2.72 | 0.00 | 1.51 | 1.56 | 1.40 | 1.80 | 1.75 | 1.96 | 3.58 |
| similarly | 0.91 | 0.99 | 2.55 | 2.91 | 0.88 | 2.07 | 1.85 | 0.25 | 1.16 | 0.71 | 0.84 |
| and | 0.78 | 0.70 | 0.70 | 0.78 | 0.73 | 0.79 | 0.86 | 0.90 | 0.99 | 0.83 | 0.88 |
| because | 0.30 | 0.39 | 0.55 | 0.48 | 0.86 | 0.80 | 1.06 | 0.85 | 0.68 | 0.94 | 0.76 |
| nevertheless | 2.03 | 5.96 | 1.82 | 2.76 | 1.12 | 1.31 | 1.17 | 0.41 | 1.47 | 2.09 | 1.19 |
| but | 0.67 | 0.56 | 0.62 | 0.66 | 0.40 | 0.49 | 0.68 | 0.59 | 0.40 | 0.58 | 0.72 |

Table 5: Weirdness of logical connectives in the Nobel Physics corpus (1902-2004) compared to the British National Corpus.

The consistently weird logical connective is *hence* as its score is never lower than 2.51; *moreover* (7.92 in 1989-1998), *consequently* (4.16 in 1989-1998), *thereby* (3.50 in 1969-1978) and *similarly* (2.78 in 1999-2004) exhibit 'peaks' of weirdness in some 10-year periods, but not all. Table 5 shows similar results of weirdness in the Nobel Physics corpus. These data – if confirmed by further analysis of textual and interpersonal metadiscourse markers – may point to peculiar features of Nobel Lectures as a text type.

## 5. Conclusions

We attempted to explore ways of tracing the ontological and hence terminological commitment in the discourse of specialist communities through an analysis of specialist terms and we then extended our investigation to the possible projection of commitment in texts through an analysis of metadiscourse markers. We chose to analyse a single text type – Nobel Lectures in two disciplines, economics and physics – as this enabled us to build two quite representative special-domain corpora where all lectures were included that were delivered and published. Our method was based on analysis of the distribution of words in special-domain corpora and comparison with a general language corpus through calculation of frequency and weirdness (*z-score*) to identify keywords. Tokens with a z-score higher than zero were used as candidates to analyse collocating strings and obtain compounds. Results were then placed in an ontology editor to find ontological patterns. Single and compound terms show the Nobel Prize winners' terminological commitment, collocational patterns reflect their ontological commitment. We then used our method based on frequency and weirdness to investigate the Nobel Prize winners' commitment to the discourse of their specialist community as outlined by their use of textual and interpersonal markers. Our aim was to show how terminology and discourse analysis can be combined to investigate the ideological question of intellectual commitment in specialist communities.

## 6. References


Ahmad, K. (1995). Pragmatics of Specialist Terms and Terminology Management. In P. Steffens (ed.) *Machine Translation and the Lexicon. 3rd Int. EAMT Workshop, Heidelberg, April 26-28,1993*. Heidelberg: Springer, pp.51-76.

Ahmad, K. and Gillam, L. (2005). Automatic Ontology Extraction from Unstructured Texts. In R. Meersman and Z. Tari (eds) *On the Move to Meaningful Internet Systems - ODBASE 2005*. Heidelberg: Springer, pp. 1330-1346.

Cabré M.T. (1998). *Terminology. Theory, Methods and Applications*. Amsterdam/Philadelphia: Benjamins.

Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), June, pp. 199-220.

Hyland, K. (2000). *Disciplinary Discourses: Social Interactions in Academic Writing.* Harlow: Pearson Education.

Kashyap V. (2004). Information Modeling on the Web: The Role of Metadata, Semantics and Ontologies. In M.P. Singh (ed.) *Practical Handbook of Internet Computing*. London: CRC Press, http://www.lsdis.cs.uga.edu/~kashyap/publications/CRC-chapter.pdf.

Maedche, A. (2002). *Ontology Learning for the Semantic Web*. Dordrecht: Kluwer Academic Publishers.

Orenstein, A. (1977) *Willard Van Orman Quine*. Boston: Twayne Publishers (G.H. Hall and Co).

Quine W.V.O. (1969). *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Sinclair, J. and Coulthard, M.R. (1975). *Towards and Analysis of Discourse.* London: Oxford University Press.

Sinclair, J. (1992). Priorities in Discourse Analysis. In R.M. Coulthard (ed.) *Advances in Spoken Discourse Analysis.* London: Routledge, pp. 79-88.

Sowa, J.F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* Pacific Grove, CA: Brooks Cole Publishing Co.,.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer Assisted Studies of Language and Culture*. Oxford and Cambridge, Mass.: Blackwell.