Discovery of (New) Knowledge and the Analysis of Text Corpora

Khurshid Ahmad
Department of Computing, University of Surrey
Guildford GU2 5XH, United Kingdom
k.ahmad@surrey.ac.uk
Maria Teresa Musacchio
Dipartimento di Lingue e Letterature AngloGermaniche e Slave, Università di Padova
Via Beldomandi 1, 35137 Padova (Italy)
mt.musacchio@unipd.it

Abstract

This paper describes how methods and techniques developed in corpus linguistics can be used to compare and contrast samples of language use over time and across genres. A diachronic Italian corpus of nuclear physics texts belonging to different genres is collected, organised, and analysed to demonstrate the use of language in shaping one of the key sciences of the 20th century.

Introduction

As science and technology develop, new concepts are introduced and established ones are abandoned or altered. Change is often signalled by the emergence of terms designating the new concepts, the elaboration of already existing terms and the discarding of terms that are no longer useful or become controversial. We have been investigating issues related to knowledge discovery and the use of information extraction in this area with specific reference to the use of two key language resources: diachronically organised corpora of domain specific texts, special language/terminological dictionaries and general language dictionaries (Ahmad & Al-Thubaity, 2003; Ahmad & Musacchio, 2003).

In this paper we look at key historic developments in (nuclear) physics in Italy during 1925-50 and relate these developments to modern day writings in nuclear and elementary particle physics in Italy. The geographical confinement of our study is deliberate: given the dominance of English in the last quarter of the 20th century, it is impossible to envision a scientist or technologist working in any language other than English. There are geopolitical reasons for the dominance of English which can be analysed in a variety of ways. Ours is a historiographic approach to this fascinating phenomenon of language dominance in science and technology.

A group of Italian physicists working for and with the physicist Enrico Fermi had made major contributions to esoteric areas of physics which then had an impact on technologies as diverse as nuclear engineering and solid state physics. It is true to say that without Fermi's insight, the development of self-sustained nuclear chain reaction would have been delayed; without his insight current models of how electrons form in semiconductors would have been less elegant and it is Fermi's model that is used to fabricate semiconductors used in information and communication technologies. In this paper we will look at the use of Italian language in one of the major discoveries of the 20th century – the nuclear atom.

Method

To the historiographic study again. We have used methods and techniques developed in corpus linguistics to compare and contrast 'representative' samples of language use for studying the structure of language in use or changes in the language. We collect data according to a timeline that starts from the inception of the subject or thereabouts to the present day. The timeline is divided into three major zones: discovery, adaptation and utility. The discovery phase is characterized by writing in the formal genre, that of learned papers, the adaptation phase is where the knowledge has been researched to a good degree and has been adapted for teaching and learning about the discovery and, in addition to the learned text, instructional text is also created. The third utility phase is where the discovered knowledge is routinely used and this knowledge is reported in the less formal genre of newspaper reportage and popular science literature. What this amounts to is a comparison of subcorpora within a corpus over the timeline.

The method of analysis includes: (a) frequency distribution of open and closed class words – frequency of terms (mainly open-class words) reflects orientation in research and consequently in writing; (b) a diachronic comparison of the open class words to see how the use of these keywords changes over time and across genres as the relationship between terms and concepts changes over time and register; (c) patterns of usage; (d) the inclusion of foreign words into the Italian language of physics (Casadei 1994; Altieri Biagi, 1998). We have observed that the diachronic comparison is essentially that of keywords related to physics, namely the so-called building blocks of nature – the *particelle elementari* (elementary particles) that constitute the nuclei and the atom – and the fundamental concepts in physics – *forza (force)*, *massa*, and *energia (*energy) together with the derived concepts of *momento* (momentum) and *stato* (state). This diachronic comparison shows the *rise* and *fall* of concepts and objects to the (Italian) physicists.

Data Collection

Our corpus is divided into three subcorpora: the early period of research in nuclear physics, the period of adaptation of the concepts of nuclear physics for teaching and learning, and the period of utility – the general discussion of the concepts of nuclear (and particle) physics. There is, of course, a diachronic dimension to

our historiographic study; but – more importantly from the point of view of the study of discovery – we have demarcated the three phases through the most commonly used text type or genre in the sub-corpora of each of the three periods. For the discovery phase, we draw only from journal papers during 1925-51; for the adaptation stage we have sampled text books for university and secondary school students, and the description of syllabi (c. 1965-2001); and, for the period of utility we have mainly used a popular science monthly (*Le Scienze*), and the feature pages of quality newspapers – the weekly supplements *Tuttoscienze* (*La Stampa*) and *Cultura/Domenicale* (*Sole 24 Ore)* – (c.1980-2003). In Table 1 subcorpora are grouped by phase and text type or genre. Size of each component of the subcorpus and overall size are also given.

Phase	Genre	Authors/Source	Tokens	Docs
DISCOVERY	Learned	Fermi (1925-50)	70127	25
1925-50	Journals	Majorana (1928-42)	24527	10
		E Amaldi (1939-51)	27274	10
TOTAL			**121928**	**45**
ADAPTATION	Text-	Persico (1967)	22267	3
1967-2001	books	U. Amaldi, etc. (1997-98)	88913	3
		Monograph samples ; Web sites, syllabi (2000-01)	100597	13
TOTAL			**211777**	**19**
UTILITY	Popular	*Le Scienze* (1993-2003)	83311	18
1982-2003	Science	*Tuttoscienze* (1992-2000)	49928	64
		Sole (1983-2003)	25082	24
TOTAL			**158321**	**106**

Table 1: Components of our corpus of nuclear physics grouped into the discovery, adaptation and utility sub-corpora.

The corpus has 492026 tokens distributed amongst 170 texts. The average number of words per document varies from 2710 (discovery) to 11146 (adaptation) and to 1494 words per document (utility). Unusually for corpus studies, we have been able to collect all the published writings of some of the authors in our corpus. All the Italian nuclear-physics writings of the two pioneers of nuclear physics (Fermi and Majorana) have been included. This is perhaps important for the discovery phase as it is sometimes characterised by lower linguistic output when compared with the other phases. Our corpus is 'balanced' in terms of the size of the three sub-corpora.

Analysis

Frequency Analysis and Diachronic Distribution

The frequency analysis of the first 100 most frequent words in each of the discovery corpus shows that the closed class words (*di, e, a, la, che, il, un, è, per,* and *in*) dominate the corpus accounting for over 15% of the total corpus. The same is true of the adaptation and utility corpora. This is not surprising as these closed class words provide grammatical cohesion and give the text its 'basic' *Italian* texture. Overall, the first 100 words in each of the subcorpora account for 45% of the total and the number of open-class words is roughly the same: 27 (discovery), 29 (adaptation) and 31 (utility). These words account for 5.03%, 4.88% and 5.75% of the subcorpora. It is the open-class words that give the subcorpora the flavour of the Italian special language of physics.

The distribution of the words related to the concept of 'building blocks' does dominate each of the three subcorpora. The total frequency of occurrence of the 10 keywords remains constant – around 16,000 tokens per million words. But the distribution within the keywords changes over time In the discovery phase the keywords *neutrone/i* and *nucleo/i* comprise 61% of all the 10 building block words. This situation changes in the adaptation phase that focuses on *elettrone, nucleo/i* and *neutrone/i* comprising 50% of the building blocks. The progress in nuclear physics contributed to the elementary particle physics and the key building blocks of the late 20th century are the quarks. The interest in *atomo/i, nucleo/i* appears to decrease. The generic term *particella* appears to gain ground as the concept of nuclei has been 'accepted' but that of particles is still at the fore front.

	Discovery	Adaptation	Utility
atomo/i	**1279**	926	246
elettrone/i	1616	**2800**	1175
neutrone/i	**6742**	1823	0
nucleo/i	**3658**	**2635**	720
protone/i	517	1006	ell
neutrino/i	0	0	985
quark	0	756	**4781**
mesone/i	0	0	**1270**
fotone/i	0	**1520**	0
particella/e	3051	4358	**6443**
	16862	15823	16656

Table 2a. The distribution of the keywords related to the building blocks of physics given as frequency per million words of text. The frequency includes that of singular and plurals.

What of the basic concepts in physics over time? Table 2b shows the change here as well:

	Discovery	Adaptation	Utility
energia	1886	3031	1731
massa	1042	1572	2091
forza	0	954	1661
momento	1427	0	0
stato	968	661	1920
Total	5323	6218	7403

Table 2b. The distribution of the tokens related to the fundamental concepts in physics per million words.

The use of these terms is increasing over time and one can argue that frequency correlates with acceptance and that these concepts have been integrated within the belief systems related to the nature of matter. There is some

decrease in the use of the term *energia* otherwise the adaptation corpus shows increase in all tokens. It appears that theories of physics fascinate people more than the experiments as shown in our corpus (Table 2c):

	Discovery	Adaptation	Utility
teoria/e	1476	1435	2855
esperimenti	0	0	499

Table 2c. The distribution of the two nouns in the three subcorpora.

Patterns of Usage

The lexical level analysis shows the gross features of the special language. The usage of the lexical items throws more light on the goings-on in a science. We focus on one of the most frequent terms in our corpus – *particella/e.*

Particella or small part is a diminutive word formed from Latin *parte(m*) and as such has been in use in Italian since the 14th century. In general Italian dictionaries (elementary) particle is currently defined as "ogni costituente non divisibile della materia" (Zingarelli, 2003) or a constituent of matter that cannot be further divided. This general definition does not suggest that discovery has changed the referent of the term. A comparison with earlier general language dictionaries shows that the definition of the term has become more and more technical over the years. In Garzanti (1971) a particle was simply a "corpuscolo materiale o radiante: *particella elementare*, costituente elementare della materia e della radiazione", while Garzanti (1987) described it as a "costituente fondamentale della materia e della radiazione, individuato da quattro grandezze caratteristiche: massa, carica elettrica, spin e momento magnetico".

Particella/e is among the 100 most frequent words in all components of our corpus except the Amaldi component. The term is used with reference to elementary particles (called *particelle elementari* or – less frequently – *particelle fondamentali* in Italian). However, in one of his 1949 lectures during a visit to Italy Fermi (1950) warned that elementary particles refer to particles that in a given state of knowledge cannot be described as compound.[1] Therefore, elementary particles in Fermi, Amaldi, Majorana and Persico are components of the nucleus – alfa, beta particles, or particles described as heavy or neutral (*particelle pesanti*, *particelle neutre*). The picture becomes – linguistically – more complex in the contemporary components where focus is on nuclear or subatomic particles that have positive or negative charge (*particelle positive*, *particelle negative*), can be unstable (*particelle instabili*), heavy or light (*particelle pesanti*, *particelle leggere*), virtual or real (*particelle virtuali*, *particelle reali*) or antiparticles (*antiparticelle*). Compounding is not the only linguistic process at work here as reference is also made to the eponymous Higgs particles (*particelle di Higgs*) which in popular science are metaphorically described as *particelle di Dio* (God's particles) owing to their elusiveness. Table 3 below gives an idea of the most typical collocation patterns in the various components of our corpus and suggests that particles first referred to electrons, protons and neutrons, later to bosons, quarks, but now also have antiparticles.

LH co(n)text	Term	RH co(n)text	C
elettroni, le	particelle	α o nuclei di elio,	F
l'altra per	particelle	senza momento angolare	M
fra le coppìe di	particelle	pesanti protone-protone,	A
emissione di una	particella	ß+ sposta il nuclide	P
quark. Tutte le	particelle-	materia,come tutte le	UA
protoni, anche	particelle	prive di carica elettrica	Mo
trasformando una	particella	nella sua antiparticella.	*LS*
da Lederman come	"particella	di Dio",il bosone	*T*
bottom quark e	particelle	di tipo W.	*S*

Table 3: A concordance of particella/e showing typical collocates. [C(orpus component): F(ermi), M(ajorana), A(maldi), P(ersico), U. A(maldi etc.), Mo(nographs etc.), *L(e) S(cienze), T(uttoscienze), S(ole 24 Ore)*]

The concordance in Table 3 shows a change in the context in which the term is used and a different orientation based on text type or genre. In the discovery phase particles are described in terms of types (Fermi), of their properties under investigation – angular momentum (Majorana) – and their interactions (Amaldi). From a linguistic point of view Amaldi's *particelle pesanti protone-protone* is an early example of juxtaposition of units in Italian term formation (Dardano, 1993) which was to gain more and more ground in the 20th century owing to the influence of English. In the adaptation phase focus is on features of particles and experiments with them (*emissione di particelle β+*). Finally, in the utility stage classification and designation of particles are in the foreground.

Foreign Words in the Italian Special Language: A Case Study

Scissione-fissione

Scissione is the Italian equivalent of the English term *fission* in biology, which is defined as a "forma di riproduzione caratteristica degli organismi unicellulari, che avviene per divisione della cellula madre in due nuovi individui" (DISC, 1997), i.e. the division of a cell into new cells as a mode of reproduction. Fission designating the division of a cell or organism was borrowed by Hahn (Maltese, 2003) in nuclear physics to describe 'the splitting, either spontaneously or under the impact of another particle, of a heavy nucleus into two (very rarely three or more) approximately equal parts, with resulting release of large amounts of energy' (OED). In Italian dictionaries of science *fissione* (Garzanti 1991) is a synonym of *scissione* in biology while *fissione nucleare* is a term in nuclear physics. In general Italian dictionaries *fissione* has only one sense as a term of nuclear physics, while *scissione* – or rather *scissione nucleare* – is the 'scissione del nucleo atomico prodotta con bombardamento di neutroni. SIN. fissione' (Zingarelli, 2003) or splitting of an atomic nucleus induced by neutron bombardment since *scissione* is a general word to designate splitting in Italian. The origin of fission (1950)

[1] "In generale si potrebbe dire quindi che ad ogni stadio della scienza si chiamano elementari le particelle di cui non si conosce la struttura, e che pertanto si possono considerare come punti."

as a term taken from biology is thus completely blurred in general Italian. When writing in Italian Fermi used *scissione*, which he described as a discovery and a new phenomenon (*scissione del nucleo*, *scissione dell'uranio*). Amaldi described experiments of *scissione dell'uranio* and *scissione del torio*, while in the 1960s Persico termed the process *fissione*. In secondary school manuals in our corpus, however, *scissione* is given as a synonym of *fissione* or is used in the definition of the term. In popular science *scissione* is still used, but it has a very low frequency compared to *fissione*. The concordance in Table 4 shows how *scissione* was supplanted by *fissione* over the years.

LH co(n)text	Term	RH co(n)text	C
che all'atto della	scissione	del nucleo in due	F
sezione d'urto per	scissione	dell'uranio dall'energia	A
del tutto eccezionale	(fissione	dell'Uranio e del Torio,	P
fusione nucleare. Per	fissione	nucleare si intende la scissione	U
di una bomba a	fissione,	poi confluito	M
decadimento alfa ela	fissione	spontanea hanno una	*T*

Table 4: A concordance of *scissione* and *fissione* showing the changeover to *fissione* over the years. [C(orpus component): F(ermi), A(maldi), P(ersico), U. A(maldi etc.), M(onographs etc.), *T(uttoscienze)*.

Concordance patterns in Table 4 suggest that down to the mid-20th century Italian physicists resorted to Italian language resources to designate new discoveries by assigning special meaning to general words. Over the years, however, the influence of English as the *lingua franca* of science meant that *scissione* was superseded by *fissione* – a loan translation from English. The term is now well-established in general Italian too as examples of its elaboration (*bomba a fissione* and *fissione spontanea*) in the last two rows of Table 4 show.

Afterword

We have attempted to demonstrate the use of language in shaping one of the key sciences of the 20th century, nuclear and subsequently elementary particle physics, by diachronically analysing the texts produced in three important phases of the development of the subject. The lexical choice reflects the state of the development of the subject. Perhaps more important is the notion that accepted concepts reflect in the fabric of the language of the subject – the dictum that frequency of a lexical item correlates with its acceptance is demonstrated quite keenly by our diachronic analysis. The influx of the current major language of science and technology shows itself in the rejection of the original word for *fission* – it is perhaps ironic in that one of the first fission reaction was planned and executed in Rome!

We are currently expanding our corpus of the discovery and adaptation phases to confirm the results thus obtained. The analysis will be extended to studying the compound terms and collocation patterns: our other studies have shown that it is the first 100 open class words that act as the lexical 'infrastructure' of a subject domain.

Acknowledgments

Khurshid Ahmad wishes to acknowledge GIDA (EU5th Framework IST-2000-31123) and FinGrid projects (UK ESRC eScience pilot project).

References

Ahmad, K. & Al-Thubaity, A.M. (2003). Can text analysis tells us something about technology progress. In: I. Makoto & F. Atsushi (Eds.), Proceedings of the Workshop on Patent Corpus Processing (pp. 46-55). East Stroudsburg, PA: The Association of Computational Linguistics..

Ahmad, K. & Musacchio, M.T. (2003) Enrico Fermi and the making of the language of nuclear physics (pp. 120-140). Fachsprache, 25 (3-4).

Altieri Biagi, M.L. (1998) Forme della comunicazione scientifica (pp. 21-73). In: Fra lingua scientifica e lingua letteraria. Pisa/Roma: Istituti Editoriali e Poligrafici Internazionali.

Casadei, F. (1994) Il lessico nelle strategie di presentazione dell'informazione scientifica. (pp. 47-69). In: T. De Mauro (ed.) Studi sul trattamento linguistico dell'informazione scientifica. Roma: Bulzoni.

Dardano, M. (1993) I linguaggi scientifici. In: L. Serianni & P. Trifone (Eds.), Storia della lingua italiana (II-pp. 497-551).

De Mauro, T. (2000) Dizionario della lingua italiana. Torino: Paravia.

DISC (1997) Dizionario italiano Sabatini-Coletti. Firenze: Giunti.

Fermi, E. (1950) Conferenze di fisica atomica. Roma: Accademia Nazionale dei Lincei.

Garzanti (1971) Dizionario Garzanti della lingua italiana. Milano: Garzanti.

Garzanti (1987) Dizionario Garzanti della lingua italiana. Milano: Garzanti.

Garzanti (1991) La nuova enciclopedia delle scienze Garzanti, 3a ed. Milano: Garzanti.

Maltese G. (2003) Enrico Fermi in America. Bologna: Zanichelli.

Zingarelli, N. (2003) Lo Zingarelli 2004. Dizionario della lingua italiana, 13a ed. Bologna: Zanichelli.