Automatic Ontology Extraction from Unstructured Texts

Khurshid Ahmad and Lee Gillam

Department of Computing, University of Surrey, Guildford, Surrey, GU2 7XH, UK
{k.ahmad, l.gillam}@surrey.ac.uk

Abstract. Construction of the ontology of a specific domain currently relies on the intuition of a knowledge engineer, and the typical output is a thesaurus of terms, each of which is expected to denote a concept. Ontological 'engineers' tend to hand-craft these thesauri on an ad-hoc basis and on a relatively small-scale. Workers in the specific domain create their own *special* language, and one device for this creation is the repetition of select keywords for consolidating or rejecting one or more concepts. A more scalable, systematic and automatic approach to ontology construction is possible through the automatic identification of these keywords. An approach for the study and extraction of keywords is outlined where a corpus of randomly collected unstructured, i.e. not containing any kind of mark-up, texts in a specific domain is analysed with reference to the lexical preferences of the workers in the domain. An approximation about the role of frequently used single words within multiword expressions leads us to the creation of a semantic network. The network can be asserted into a terminology database or knowledge representation formalism, and the relationship between the nodes of the network helps in the visualisation of, and automatic inference over, the frequently used words denoting important concepts in the domain. We illustrate our approach with a case study using corpora from three time periods on the emergence and consolidation of *nuclear physics*. The text-based approach appears to be less subjective and more suitable for introspection, and is perhaps useful in *ontology evolution*.

1 Introduction

Literature on intelligent systems invariably refers to a thesaurus of domain objects in the construction of knowledge bases. A 'thesaurus' suggests the existence of a range of words and phrases associated with a concept. The names of the objects form the *terminology of* the domain. The organisation of terminology is discussed under the rubric of *ontology*. Ontology is a branch of philosophy, and some philosophers believe that to understand what ***is*** in every area of reality one should look into the theories of sciences [1]. Quine, one of the proponents of modern ontology, has asked two key questions related to 'philosophy within science' [2]: (i) What are the conditions that lead to talking scientifically? (ii) How is scientific discourse possible? The answer in Quine, by no means exhaustive, is in the *ontological commitment* on the part of a scientist or group of scientists: the scientists observe physical phenomena and articulate them for others in linguistic entities (the controversial observation sentences). This sharing of common roots of reference – physical and linguistic – are,

for us, signs of being committed to the same set, or system, of concepts, and this is the basis of Quinian ontological commitment.

Researchers working on ontological 'engineering' tend to hand-craft thesauri on an ad-hoc basis and on a relatively small-scale. Laresgoiti et al discuss the ontology of an intelligent network control system in which the 'concepts' appear to be derived from an existing data dictionary [3]; Gómez-Pèrez, Fernández-López, and Corcho compose a *travel ontology* without reference to the source of knowledge that comprises 'concepts' such as "American Airlines Flight", "Iberia Flight", "Hotel" and "Japan Location", and a list of relations like "departurePlace" and "placed in" [4].

In this paper we argue that a more scalable, systematic and automatic approach to ontology construction is possible using methods and techniques of information extraction, corpus linguistics, and terminology science to examine archives of a specialist domain of knowledge. The methods and techniques enable the identification of the objects, processes, and concepts that describe an application area, a domain of specialist knowledge, or indeed a whole discipline. We describe a method to automatically extract key terms, and relationships between the key terms, from relatively-large corpora of unstructured, i.e. not markup up, text in particular specialisms, and how international standards (ISO) that have emerged from terminology science can facilitate construction of terminology databases and of the domain ontology. Our system generates hierarchically arranged terms from the text corpora that indicate the ontological commitment of researchers and practitioners of the domain. When represented in one kind of markup, the hierarchically arranged terms can be used as a basis for an ISO-standards conformant terminology, and when represented in an ontology interchange language they can be inspected and refined in an ontological engineering tool like *Protégé* [5]. The hierarchy can be augmented by linguistic pattern analysis to confirm, contract or expand elements of the hierarchy [6]. The method uses the frequency contrast between the general language of everyday use and the special language of the domain to identify key domain words [7], and expands the analysis to the discovery of consistently used patterns in the neighbourhood of these domain words. Our work relates to the emergence of new domains of knowledge and how scientists and philosophers construct an edifice of a new(er) branch of knowledge [8].

Our method identifies and extracts (candidate) terms and (candidate) ontologies from a set of written texts. The *candidate* nature of these results should suggest that we make no claims to treat subjective differences within ontology: on discovering *red wine*, we would, simply, present this as a class of *wine*, assuming this to be what has been discovered. We would leave it to the subject experts, and connossieurs, to debate whether *red* is the value of the attribute of *colour* of *wine*, and how *dry white wine* would now fit into this system. Such distinctions are essentially creative and mental tasks carried out with flair by human beings that is hitherto unmatched: even the most ambitious of ontology project does not attempt such a subjective qualification. Our work is no exception.

The results of our extraction – the terms and ontologies - can be validated by subject experts: the present case study is in Nuclear Physics and this subject has been studied by one of the authors (KA). Within the parameters of author-subjectivity, the results identified are in accord with current findings within the subject. Thesaurus building systems can benefit from automated identification of terms and their inter-relationships within a specific domain of knowledge. The frequency of use of the

terms, and the neighbourhood of the terms, is an indication of how knowledge in the specialism is organised by researchers and practitioners of that domain.

2 A method for extracting ontology

By examining archives of a specialist domain of knowledge, we contend that one can find objects, processes, and concepts that describe an application area, a domain of specialist knowledge, or indeed a whole discipline. Approaches to the identification of domain-specific keywords – the terminology of the domain – generally rely on extensive prior linguistic knowledge and linguistic extraction techniques [9], [10], [11], [12], [13]: use of part-of-speech (POS) taggers predominates. Our treatment differs from these approaches in taking an initially statistical approach, which is suitable for subsequent augmentation using linguistic techniques. It uses the difference between the general language of everyday use and the special language of, for example, physics or philosophy or sewer engineering as its basis. This difference can be determined by comparing the relative frequency of words in texts with their relative frequency in the general language [14]. A special language is a subset, and sometimes an extension, of the natural language of a specialist. Scientists, amongst others, have to convince yet others of the value of their work: repetition is one of the rhetorical devices used in almost all enterprises for informing, exhorting or convincing others of the value of your own view. Evidence of the use of repetition can be found in repositories of specialist documents written in a natural language and adorned by images and tables of numbers. Authors of these specialist documents use a small vocabulary very productively: names of objects are used in singular and plural; a smaller number of words are used to make many of the compound terms and phrases. For example, in a research paper about modern nuclear physics one will find that the term *nucleus* is used 600 times more frequently in the subject than, say, in the 100 million-word British National Corpus – a representative sample of English language [15]. Gillam has refined and extended this contrast, and created a system that generates hierarchically arranged terms indicating ontological commitment within a domain [5].

The method is based on Quirk's notion that frequency of use of words correlates with acceptability of those words as part of the vocabulary [16]:33. The BNC is used as a reference collection of general language and we use the *weirdness index* [14] that has been adapted by smoothing [17], to seed a collocation extraction technique [18]. This analysis produces a hierarchy of terms/concepts via semantic inclusion. The hierarchy can be augmented by a linguistic pattern analysis that may confirm, contract or expand elements of the hierarchy (see [6] for details). By reference to international standards (ISO) for terminology, we can facilitate the construction of terminological resources that have a potentially wider scope of use as thesauri or ontologies: the hierarchy can be exported to an ontological engineering system like *Protégé*. Such a terminology/thesaurus/ontology is then suitable for validation by experts. The algorithm for this is shown in Fig.1.

1. **COLLATE** Text Corpora:
   a. Obtain a general language corpus $S_{General}$ comprising $N_{General}$ tokens
   b. Create a text corpus of documents in a specialist domain $S_{special}$ with $N_{Special}$ tokens
2. **COMPUTE** distribution of all 'words'
   FOR i=1 to $N_{Special}$ /* Compute frequencies of use of single words */
   w = $token_i$;
   IF w ∉ words
   THEN words:= words ∪w & frequency(w)=1
   ELSE frequency(w) = frequency(w)+1
   NEXT i
   FOR i=1 to #words /* Extract frequency f(w) of single word $\mathbf{w_j}$ from $S_{General}$ and $S_{Special}$ */
   weirdness(w_j) := ($f_{Special}$ (w_j)* $N_{General}$)/($f_{General}$ (w_i)+1) * $N_{Special}$)
   NEXT i
   avg_f:=(∑ $f_{Special}$(w_i))/ $N_{Special}$; $\sigma_{frequency}$:=(∑ ($f_{Special}$ (w_i)- avg_f)^2/ ($N_{Special}$*($N_{Special}$-1))
   avg_{weird}:= (∑ weirdness(w_i))/$N_{Special}$; σ_{weird}:=(∑ (weirdness(w_i)-avg_{weird})^2/ ($N_{Special}$*($N_{Special}$-1))
3. **EXTRACT** 'keywords'
   FOR i=1 to #words /* Compute *z-scores*/
   $z_{frequency}(w_i)$:= ($f_{Special}$ (w_i)- avg_f)/ $\sigma_{frequency}$
   $z_{weird}(w_i)$:= (weirdness(w_i)-avg_{weird})/ σ_{weird}
   IF $z_{frequency}(w_i) > \tau_{frequency}$ & $z_{weird}(w_i) > \tau_{weird}$
   THEN keywords:= keywords∪w_i
   NEXT i
4. **EXTRACT** significant collocates of keyword
   FOR i=1 to #keywords /* Build hierarchy */
   FIND $keywords_m$ in $S_{special}$
   FOR j=-5 to +5, j ≠ 0; $f_{coll}(keyword_i, w_{i+j}) := f_{coll}(keyword_i, w_{i+j})+1$;**NEXT** j
   IF $y(f_{coll}(keyword_m, w_{m+k})) > \tau_{collocation}$
   THEN collocations:= collocations∪$(keyword_m, w_{m+k})$
   NEXT i

Fig. 1. An algorithm for extracting 'keywords' and collocates using given threshold values (τ);y is a collocation statistic due to Smadja [17]. Iterative application of step 4 using sets of collocations results are used to produce the hierarchy.

Elsewhere, we have used Zipf's Law [19] to demonstrate similarities in the patterns of frequency of words used in different specialisms: the approach may be generalisable to other specialisms as although the words differ the patterns of use are similar. In this paper, we apply the method to three sub-corpora of nuclear physics to identify changes in the ontology over time, or perhaps *ontology evolution* [20].

3 Text-based ontology: A Nuclear Physics Case Study

3.1 A note on the domain: Nuclear Physics

The evolution of nuclear physics in the 1900's provides an example of how concepts are re-defined (semantic shift) and terms re-lexicalised. Papers start to emerge early in the 20th century describing that the 'indivisible' *atom* was 'divisible' after all and contained a positive *nucleus* surrounded by negatively charged *electrons*. Ernst Rutherford conducted the first of the pioneering experiments in the emerging field of

nucleus physics (*sic*.) and published a number of papers in this emergent field. Rutherford was concerned about the *deflection* of alpha-particles when scattered on selected targets and he noted the *deflexions* using *scintillation counters.* In his later years, he worked to artificially transmutate one element into (many) others by bombarding the element with a beam of *particles* and thus found *artificial radio-activity*. Niels Bohr is regarded as one of the pioneers of modern quantum theory and he produced a model of a stable atom in which the negatively charged particles (*electrons*) precess around the positive nucleus in a *stable orbit* – that is, despite traversing in an electromagnetic field, due to the nucleus, the electrons do not radiate energy. Subsequently, Bohr was involved in nuclear *fission* and produced a model of how a Uranium nucleus, when bombarded by *neutrons*, will split into two fragments, releasing massive amounts of energy. Rutherford and Bohr's work led us to the modern conception of a nucleus comprising the positively charged *protons* and the neutral *neutron* together very compactly by exchanging elementary particles called *mesons.* A system of concepts related to the 'new' structure of matter, in many ways analogous to the planetary system, was established through the frequent use of words (terms) in physics then, especially *atom*, but with a changed meaning, and the adoption of terms from other disciplines, including *nucleus* from botany. The frequent use of these two keywords on their own and in compound terms reflects the ontological commitment of the then modern physicist.

The term *nuclear physics* was first used after the 2nd World War, and due both to its peaceful uses and destructive potential it has received substantive funding and a number of researchers are involved in this field. As time has passed, the subject has focussed on deeper and deeper studies of *nuclear matter*, and one of the current exciting developments is in the field of *exotic nuclei*: nuclear physicists can create highly unstable nuclei and extremely short-lived nuclei in laboratory conditions, and study the behaviour of such nuclei to measure nuclear forces and determine the *structure* of nuclei. Amongst the more recent discoveries are the *halo nuclei* – where *neutrons and protons* are loosely bound to a nucleus much like a halo surrounds our Moon. New structures have been discovered that have been explained by referring to a knot-like structure – the so-called *Borromean rings*. Here, physicists are introducing a new method of studying the structure, redefining the concept of *nucleus* as a stable entity, and then describing newer forms of highly unstable matter – a highly transient element comprising a *halo* around an otherwise *stable core*.

Our task is to investigate whether such key concepts, that would be articulated through frequently used keywords, are automatically extracted from a diachronic study of the texts produced in the three periods in the development of nuclear physics. To this end we have analysed three sets of texts: one written by Rutherford and his co-authors, another by Bohr, and a third that is a random sample of texts published between 1994 and 2004. Rutherford's texts are exclusively from journals; for Bohr we have also included letters he had written to his brother (another physicist) and his wife (non-physicist). The modern nuclear physics texts comprise journal papers, popular science articles and conference announcements. For our comparisons, we use the BNC as a common reference point (reference corpus). See Table 1 for details of these 4 corpora.

Table 1. Composition of the 4 text corpora

Subject	No. of texts	Time Period	No. of Tokens	Text Types
1. Nuclear Physics (*Rutherford*)	17	1908-1932	61,035	Journal Papers (JP)
2. Nuclear Physics (*Bohr*)	16	1920-1950	101,201	JP; Letters (LT)
3. Nuclear Physics (*modern*)	157	1994-2004	564,415	JP, Conference Announcements, Popular Science, Academic Course Details
4. British National Corpus	4124	1960-1993	100,106,029	Various including extracts from newspapers, specialist periodicals, journals and popular fiction in subjects including natural science, social science, commerce and leisure.

3.2 Automatic Extraction of single-word terms and Diachronic Variance

The statistic we use extensively is the *weirdness index* (eqn. 1), a measure of the use of a word in special language compared to its use in a representative corpus of general language texts:

$$weirdness = \frac{N_{GL} f_{SL}}{(1 + f_{GL}) N_{SL}} \quad (1)$$

where f_{SL} is the frequency of word in the specialist corpus, f_{GL} is its frequency in BNC, and N_{SL} and N_{GL} are the token counts of the specialist corpus and the BNC respectively. The disproportionately used words are more likely to be terms of a specialist domain and terms are used to denote *concepts* [7].

Consider the distribution of 10 most frequently used words in each of the three corpora, excluding the so-called closed class or *stop* words (e.g. *the, a, an, but, if*….) as shown in Table 2. The most frequent words in Rutherford include *particle(s)*, *atoms* and *nucleus*. These words are 'disproportionately' used by Rutherford when compared with typical text in English – he uses *particle* 629 times more frequently than is used in the British National Corpus, *atom* 896 times more frequently and *nucleus* 841 times more frequently. There are clues here of the famous scattering experiments – where Rutherford measured the *range* (22 times more frequent) of *alpha* (485 times more frequent) particles emitted by a radioactive source (in *centimetres* or *cm*). The emphasis in Bohr is on the *electrons* (1652 times more frequent), and (the electron) *orbits* (1204 times more frequent); *nucleus* is used less disproportionately in Bohr (652 times) than in Rutherford (841 times). Bohr's more frequent use of *electrons* in an *orbit* should not detract from the fact that the orbit was around the *nucleus*. The word *nucleon* (a hyponym for *proton* and *neutron*) is amongst the most disproportionately used – over 36410 times more frequent in our corpus than in the BNC; *energy* (and its unit *mev –million electron volts*) is amongst the most frequently used. The frequency of *cross, section* and *scattering,* reflects the use of the term *cross-section* in nuclear physics where it is used to refer to a measure of the probabil-

ity of a *nuclear* reaction; *scattering cross-section* is used in determining the determining the *structure* of *nuclei*.

Table 2. Distribution of 10 most frequent single words (terms) in our three corpora – with number in parentheses indicating the rank of the word in a complete wordlist of the corpus

Rutherford			Bohr			Modern		
Token	***Rel. freq***	***Weirdness***	***Token***	***Rel. freq***	***Weirdness***	***Token***	***Rel. freq***	***Weirdness***
particles (10)	1.05%	629	electrons (11)	1.03%	1652	energy (21)	0.48%	39
atoms (18)	0.73%	691	atom (20)	0.60%	1084	neutron (30)	0.34%	1390
number (21)	0.61%	12	electron (25)	0.46%	488	nuclei (36)	0.30%	972
particle (22)	0.59%	847	nucleus (27)	0.44%	652	nuclear (38)	0.29%	36
nucleus (23)	0.56%	841	energy (28)	0.42%	34	cross (42)	0.28%	38
alpha (25)	0.54%	485	theory (31)	0.38%	29	mev (46)	0.27%	7193
atom (27)	0.49%	896	number (32)	0.36%	7	state (47)	0.27%	7
cm (28)	0.48%	231	orbits (33)	0.34%	1204	body (50)	0.26%	10
range (30)	0.45%	22	elements (35)	0.33%	52	nucleon (51)	0.25%	36410
hydrogen (32)	0.42%	348	atomic (39)	0.29%	264	scattering (52)	0.25%	497

The lexical choice of modern nuclear physicist has changed over time and is principally shown by the more proportionate use of the words *atom, atoms* and *atomic* – used around 30 times more in the Modern corpus compared to 800 times or more as was the case for Rutherford and Bohr.

Frequency varies considerably across a corpus of words. Some words are used very frequently and others very infrequently: *neutron* is used 1944 times in the 564115 word Modern nuclear physics corpus; *neutrons* 549 times; *dineutron* 44 times; *multineutron* 5 times; and *tetraneutron, trineutron,* and *neutronization* only once. The average frequency in the Modern corpus is 29.18 with a standard deviation of 445. Much the same can be said about the variation in weirdness across the corpus – the average weirdness is 226.5 with a standard deviation of 1307. The standard deviation of frequency in the British National Corpus is 11,000. Instead of using frequency and weirdness as a measure of disproportionate use, we calculate the *z-scores* (eqn. 2) for both frequency (*f*) and weirdness (*w*):

$$z_i(x) = \frac{(x_i - \bar{x})}{\sigma_x} \tag{2}$$

We can now specify a minimum value of z-scores for frequency and weirdness and use this to automatically select only those words that are above this value, removing the subjective treatment of importance of these words. For a threshold of 0.25, that is all words with a frequency that is above the average frequency (and weirdness) by a margin of a quarter of a standard deviation, we find: Rutherford's corpus has only 8 words that satisfy that criteria in 3446 unique words; Bohr's corpus has 17 amongst 4145 words; and the Modern corpus has 27 words amongst 19341 unique words.

Table 3 shows the 'new' selection together with words that were selected on the use of statistical criteria.

Table 3: Distribution of (8 or 10) most frequent words that satisfy the 0.25*standard deviation criteria across the three corpora. Words in bold are those that were identified manually in Table 2 – the combination removes from consideration those words with low weirdness. Underlining denotes words shared across these topmost – *nuclei* across all 3 with varying importance, *nucleus* and *atom* across Rutherford and Bohr.

Rutherford			**Bohr**			**Modern**		
	z (*f*)	**z (*w*)**		**z (*f*)**	**z (*w*)**		**z (*f*)**	**z (*w*)**
particle	2.45	0.29	**electrons**	4.37	3.18	**neutron**	4.30	0.89
nucleus	2.35	0.29	**atom**	2.50	2.02	**nuclei**	3.71	0.57
atom	2.04	0.32	**electron**	1.90	0.81	**mev**	3.39	5.33
scintillations	1.12	36.9	**nucleus**	1.80	1.15	**nucleon**	3.16	27.68
nuclei	0.92	0.26	**orbits**	1.37	2.27	halo	2.52	0.78
helium	0.89	0.40	**atomic**	1.17	0.36	projectile	1.47	2.27
radium	0.88	3.07	hydrogen	1.05	0.27	proton	1.46	0.27
deflexion	0.30	25.4	nuclei	0.87	1.31	6he	1.39	21.80
			quantum	0.87	0.36	coulomb	1.17	5.17
			stationary	0.80	0.86	breakup	1.12	1.42

Table 3 shows a trace of the *ontological commitment* of workers in nuclear physics over a 100 year period. This is portrayed also in Fig. 2. The commitment to study *energy* – one of the three concepts physicists study, the other two are *force* and *mass* – and the *nucleus* remains the same over the century. What changes over the period is the enthusiasm for the study of unstable systems, hence the word/term *halo*, by way of laboratory created *nuclei* – and we have the word/term *projectile* and a related word *breakup* (referring to the break up of nuclei). This conclusion is all the more gratifying as our system has no **domain** knowledge *per se*. Note that the single words (and the related concepts) denote generic concepts and it is in the specialisation of these potential terms that one can see ontologically motivated hierarchies. This we discuss next.

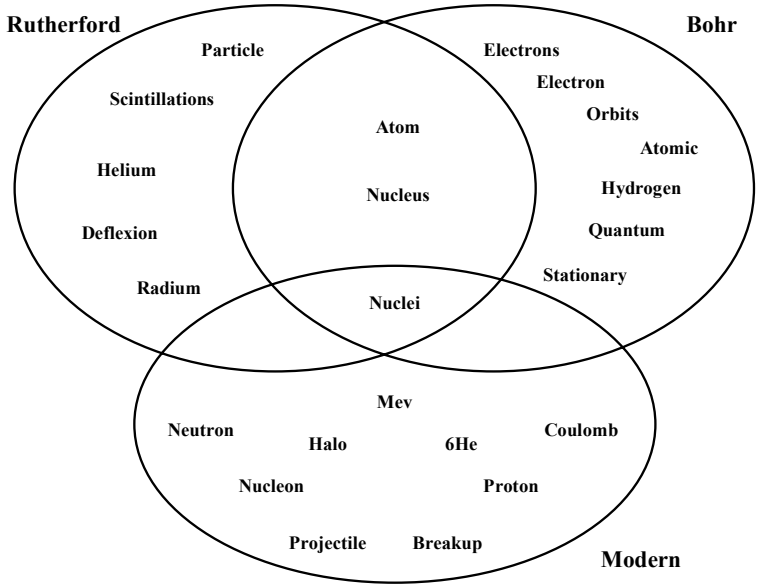

Fig. 2. Lexical sharing amongst the topmost automatically selected words from the three corpora

3.3 Automatic Extraction of Compound Words and Diachronic Variation

What is more important, perhaps, than these frequent words alone is the manner in which the frequent words produce expressions comprising multiple words. The multiword terms help to specify a, perhaps more complex, generic concept – *nuclear energy* is a form of *energy* and is different from *electrical* or *heat energy*; *nuclear reaction* is a kind of *reaction* and *direct nuclear reaction* is a *nuclear reaction* that is different from *compound nuclear reaction*.

Returning to the application of the method outlined in Section 2, for each collection we use a z-score value, manually assigned or automatically derived, to systematically determine the number of keywords for further analysis. Collocation analysis is commonly used in corpus linguistics to identify words that occur frequently within a given neighbourhood of each other, and that are used to convey specific meanings. Corpus linguists make frequent use of *mutual information* and *t-score* statistics to determine the significance of bigrams – two words within the neighbourhood. We have found these metrics to provide limited information with regard to the importance of the individual positions within the neighbourhood. Hence, we have applied the analysis due to Smadja who has argued that significant collocates of a word are within a neighbourhood of five words either side of the word (the *nucleate*) denoted as L1-L5 (left) and R1-R5 (right); Smadja has outlined metrices for quantifying the strength of the collocation including one that isolates *peakedness* (U) of the collocation in the various positions of the neighbourhood together with a z-score: significant collocates have a U-score >10 and z-score > 1. We have implemented Smadja's

method and are able to automatically extract collocation patterns. Consider five of the 7 significant collocates of *nucleus* (Table 4a) and the dominant positions of collocates (Fig. 3.).

Table 4a. Selected collocates of *nucleus* in Rutherford using {U, k}= {10,1}

Collocate	**L5**	**L4**	**L3**	**L2**	**L1**	**R1**	**R2**	**R3**	**R4**	**R5**	**U**	**k**
charge	1	0	19	0	0	17	1	1	2	3	47.2	11.8
helium	1	0	0	0	**19**	0	0	0	4	0	32.0	6.1
hydrogen	2	0	1	0	**14**	0	1	3	0	1	16.4	5.6
atom	1	5	1	0	0	0	0	3	**13**	1	14.8	6.1
theory	1	0	1	0	0	**13**	0	0	1	0	14.6	3.9

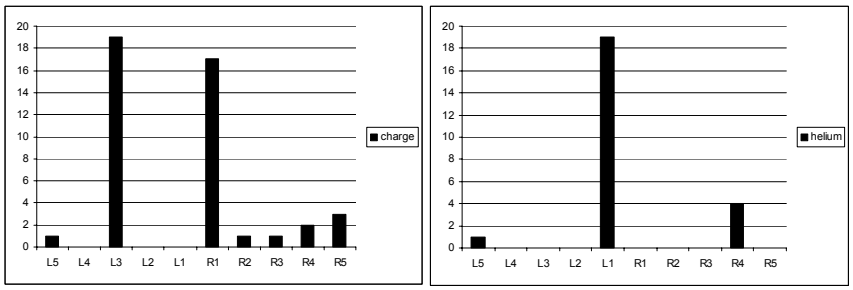

Fig. 3. Collocations in Rutherford with *nucleus* of *charge* and *helium*. Note that dominant positions for *charge* collocating with *nucleus* are L3 (*charge* [] [] *nucleus*) and R1 (*nucleus charge*), while the dominant position for *helium* collocating with nucleus is L1 (*helium nucleus*). Smadja's "z-score k" [18] would select these three patterns for further analysis.

The focus in Rutherford is on the term *nucleus charge* (44 collocates) and there is an enumeration of the nucleus of different elements (*hydrogen* and *helium nucleus)* and a reference to *nucleus theory.* The focus in Bohr is rather different (Table 4b, Fig. 4.) as shown by five (of the 7) collocates where we have *positive nucleus* and *electron(s) +X (+Y) nucleus* are amongst the more common collocates:

Table 4b. Selected collocates of *nucleus* in B**ohr** using {U, k}= {10,1}

	L5	**L4**	**L3**	**L2**	**L1**	**R1**	**R2**	**R3**	**R4**	**R5**	**U**	**k**
round	0	0	5	**38**	0	0	0	0	0	2	127.1	7.4
electron	9	19	**23**	0	0	0	0	4	0	0	68.5	9.2
rotating	0	4	**23**	0	0	0	0	0	2	0	46.5	4.6
electrons	7	**17**	15	1	0	0	5	5	5	6	30.3	10.3
positive	0	5	0	0	**17**	0	2	2	2	0	24.8	4.4

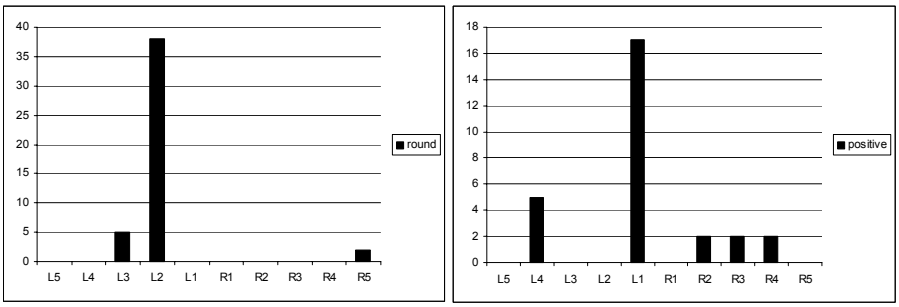

Fig. 4. Collocations in Bohr with *nucleus* of *round* and *positive*. Note that dominant position for *round* collocating with *nucleus* is L2 (*round* [] *nucleus*), while the dominant position for *positive* collocating with nucleus is L1 (*positive nucleus*). Smadja's "z-score k" would again select these two patterns for further analysis.

The collocation patterns in modern nuclear physics are somewhat different – we are in a period where the concept of a *nuclear atom* is established, and *atom* itself goes unmentioned – the collocate *atomic nucleus* only occurs 16 times in the 564,115 word modern nuclear physics corpus and *atomic nuclei* 52 times. What we find instead (Table 4c, Fig. 5.) is the phraseology of reacting nuclei (*target* and *residual*) and unstable nuclei *He*lium-*6* and *Li*thium-*11* nuclei, natural Helium has 4 nucleons and Lithium has 8:

Table 4c. Selected collocates of *nucleus* in Modern Nuclear Physics Corpus using {U, k}= {10,1}

Collocate	**L5**	**L4**	**L3**	**L2**	**L1**	**R1**	**R2**	**R3**	**R4**	**R5**	**U**	**k**
target	3	2	0	1	**114**	1	0	5	11	5	1117	22.4
halo	4	4	2	1	**90**	0	2	5	4	5	684	18.4
compound	0	1	0	0	**41**	0	1	0	1	2	148	7.0
residual	0	0	1	0	**25**	0	0	1	0	0	55.4	3.9
borromean	0	1	0	2	**22**	0	0	0	0	0	42.7	3.6

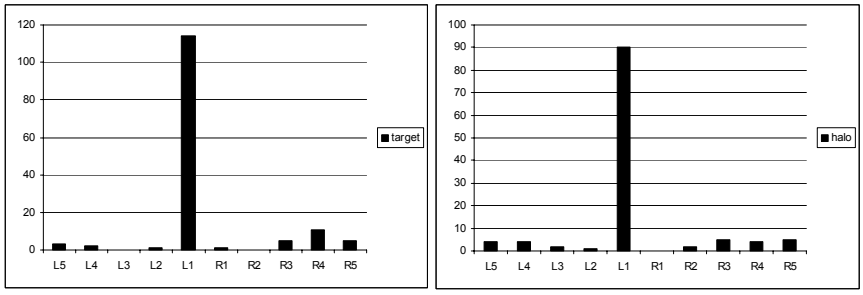

Fig. 5. Collocations in Modern corpus with *nucleus* of *target* and *halo*. Note that dominant positions both collocations is L1 (*target nucleus*, *halo nucleus*), both of which the z-score would select for further analysis.

This kind of analysis is of interest for discovering increasingly more complex *concepts* within specialist text collections. For example, the *halo* was found, as fifth most important keyword, in the modern collection; collocations (Step 4 in the Algorithm) around *halo* include: *halo nuclei*; *halo nucleus*; *neutron halo*; and *halo neutrons*. We further discover the *weakly bound neutron halo* (*f*=2), but the most expanded tree forms under the *halo nuclei*.

The various patterns identified can be used to produce collocational networks [21]. Since we believe that many of these networks may provide evidence of semantic inclusion, we assert *isa* relationships between the words and their collocates and produce hierarchies from these collocational networks (see, for example, [6]). The hierarchies can be used in combination with international standards for the production of terminology interchange formats (using ISO 12620 and ISO 16642) that can be used to populate terminology databases [22]. These formats express concepts, conceptual relations, and provide for items of administrative information, including versioning, and for items of documentary information including sources and contexts of use. There is some degree of overlap with these formats and the so-called ontology exchange languages, and hence the results can also be exported to an ontological engineering tool (Protégé), using a semantic schema such as the Web Ontology Language (OWL), to facilitate knowledge base development and visual inspection and refinement: a domain expert can quickly make on-line corrections to the hierarchy in collaboration with a knowledge engineer using the ontology tool. The role of the knowledge engineer in our method relates exclusively to the construction of the knowledge base and avoids any intuitive input on their part. The (partly-pruned) ontological commitment of modern nuclear physicists in relation to *halo nuclei*, alluded to above, can be seen in Fig. 6, as drawn using the Protégé component ("Tab") OntoViz.

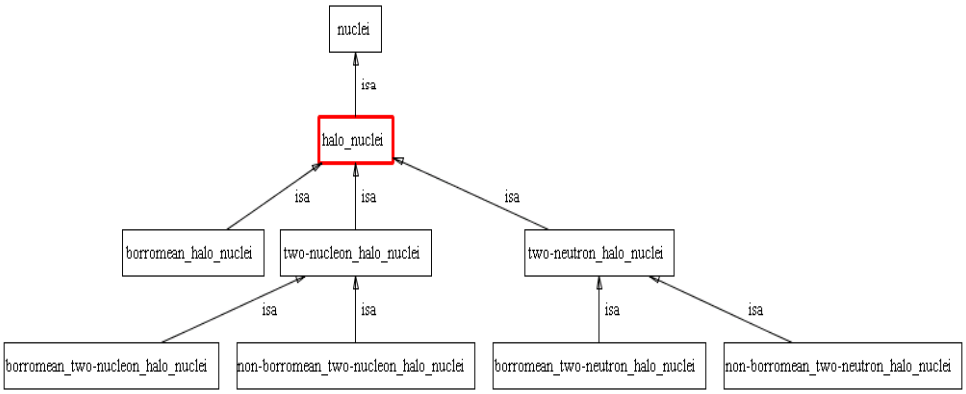

Fig. 6. A candidate hierarchy showing different types of *halo nuclei.*

3.4 Ontology Evolution

Ontology Evolution has been discussed with reference to a wine ontology [20], [23]. The creation of a log of the evolution is emphasised using a differencing operation between ontologies. Without reference to the sources of the original knowledge that went into the production of the ontology, the reasons for the different ontologies may not be easy to discover. The importance of some terms at one time versus other terms at another would result in quite a large log file of changes. Whether such approaches are able to capture, for example, the *atom* changing semantically from being *indivisible* to *divisible* is not clear with from such small examples.

We use the term *ontology evolution* to refer to a change in a specialism's existing repertoire of concepts over a period of time. As the subject of nuclear physics has evolved over time, so the record of the ontological commitment of the workers has evolved. Over a short period of time (c. 25 years from 1900), the concept of a unitary atom was rejected, its constituents were identified experimentally and elaborated theoretically, and by the late 1930s a new field of physics – nuclear physics – had emerged. Researchers now seldom use the term *atom* and are careful in the use of the now generic headword *nucleus*. : nuclear physicists invariably use a qualificatory adjective or another noun when using *nucleus* and its derivatives *nuclear* and inflected form *nuclei*.

We can measure changes in ontological commitment over time, particularly with reference to the changes in importance attributed to these commitments by the authors, by calculating the weirdness index again, this time within the sub-corpora of the specialism. Words that were automatically identified as important for Bohr and that were almost irrelevant for Rutherford, are indicative of changes in the subject and can be identified, again, by high frequency and weirdness values (Table 5a).

Table 5a. Frequency and weirdness values for words of importance in the Bohr corpus, but of low importance in the Rutherford corpus.

Word	f (Bohr)	f (Ruth)	Weirdness
states	231	0	139
stationary	210	0	127
ring	173	0	104
rings	104	0	63
configuration	174	1	52
configurations	79	0	48
quanta	143	2	43
fission	63	0	38
bound	155	2	31
orbits	342	6	29

The 'ideas' that are either common-place within the subject, or have become 'suppressed' for other reasons, will occur with somewhat lower frequency and weirdness values (Table 5b): for Bohr, the atom is slightly more important (> 1) than for Rutherford. The remainder of these words are of lesser importance (< 1), with *radium*, *deflexion* and *scintillations* finding little or no interest at all. Those items that were of importance in Rutherford's work are not the subject of study for Bohr, although the field of study effectively remains the same. However, if they are so common-place

that they no longer necessitate description, they may not be in the newer version of the ontology because that which is understood does not to be discussed.

Table 5b. Frequency and weirdness values for words of importance in the Rutherford corpus, but that have low importance in the Bohr corpus.

Word	f (Rutherford)	f (Bohr)	Weirdness
particles	640	147	0.14
atoms	445	223	0.30
particle	358	91	0.15
nucleus	344	442	0.77
atom	302	606	1.21
scintillations	174	0	0
nuclei	146	227	0.93
helium	142	102	0.43
radium	140	3	0.01
deflexion	60	1	0.01

The same comparison can be made for the Modern Physics corpus, where we find that notions of *stationary*, *ring*, *rings*, *quanta*, *fission* and *orbits* have either become fundamental to the subject, or are suppressed for other reasons (Table 5c).

Table 5c. Frequency and weirdness values for words of importance in the Bohr corpus, but that have low importance in the Modern corpus.

Word	f (Bohr)	f (Modern)	Weirdness
states	231	202	0.87
stationary	210	0	0
ring	173	0	0
rings	104	0	0
configuration	174	11	0.06
configurations	79	30	0.37
quanta	143	0	0
fission	63	0	0
bound	155	117	0.75
orbits	342	0	0

Since the most significant terms, according to our method, are changing in their importance over time, the challenge and need for managing ontology evolution, and for managing the input documents that form a part of this process, becomes significant. Terminology interchange formats, defined according to the International Standards, make provsision for the management of such reference material.

4. **Discussion**

We have automatically extracted hierarchical trees of terms from collections of natural language texts that demonstrate evidence of, and change in, the ontological commitment in physics over a period of time. We have demonstrated the efficacy of the automatic extraction method in a number of domains including nano-technology, forensic science, philosophy of science, financial investment, and epidemiology (for

example, see: [24], [25], [26]). The principal inputs to our system are the collection of texts in an arbitrary domain and the list of general language words. Both the terminology and ontology have a reference point – the text collection: this contrasts with the rather ad-hoc work usually reported in ontological engineering.

Advocates of part-of-speech (POS) tagging might suggest that we ignore POS information at our peril. We have analysed the Rutherford corpus using the Brill tagger in its default (untrained) state. Rutherford's 8 keywords from Table 1 occur as either a kind of noun (NN, NNS or NNP), or as an unknown (UNK). At this level, there would appear to be negligible gain from POS tags *per se*.

Table 6. Part-of-speech information for the 8 keywords selected from the Rutherford corpus (Table 1).

Word	**NN**	**NNS**	**NNP**	**UNK**
atom	298			4
deflexion				60
helium	141		1	
nuclei		139	7	
nucleus	339		3	
particle	323			35
radium				140
scintillations		173	1	

If we consider the *nucleus* in Rutherford, the three patterns we find with POS information are: *hydrogen nucleus* = NN NN*; helium nucleus* = NN NN*, nucleus theory* = NN NN. While this again imparts little additional information and, indeed, one may argue about meronymy since the *nucleus* is a part of *hydrogen*, this may provide some limited evidence useful for mutual validation of our results. Where the expansion is adjectival (for example, *swift atoms* = JJ NNS) determination of whether the relationship is hierarchical, or whether this should be considered as an **attribute** or **value** remains subjective: again, is *red wine* a kind of *wine*, or is *red* a value of the *colour* of *wine*? Such judgements need to be subjectively made, and our objective method does not make provsision for such decisions.

We have explored Hearst's work [27] for augmenting our ontologies, combining phrases including *such as* with our extracted terms and with POS information to enable the bootstrapping of ontologies from unstructured texts [6]. The prior identification of terminological data may circumvent the need for training the POS taggers, which can now be used against the more grammatical elements of the texts. Consideration of the expansions of phrase patterns, for example: *properties of* [] *such as* …., or *characteristics of* [] *such as*…., where [] denotes a term, may provide for further population of the ontology. There are question marks over the scalability of approaches that use POS tagging since the taggers generally require training in new specialisations. The quality of the results is, then, a function of the training plus the coverage of the rules used for identification. Using our expanded method it may be possible to reduce the dependency on the POS tagger.

In other analysis, we have discovered phrases such as *conventional horizontal-type metalorganic chemical vapor deposition reactor*, *ridge-type ingaas quantum-wire field-effect transistors* and *trench-type narrow ingaas quantum-wire field effect tran-*

sistor. We are detecting these phrases without the need for the prior linguistic knowledge that goes into the expectation of existence of specific combinations of POS tags. It may be possible, however, to use statistical validity as a means to generate POS patterns that could be used to identify further elements of the ontology, and may be worth considering in further work. In addition, since we are in a particular specialism, we do not make consideration of different senses (concepts) being indicated by the same term. Indeed, in coining terms and retrofitting new exclusive sense to extant terms, scientists restrict the terms to a single sense. For example, although *nucleus* was, most likely, adopted from biology, it is highly unlikely that it would be used in a biological sense within these corpora. Such considerations may be of value where interdiciplinarity is evident – e.g. biochemistry, although we suspect that the discipline would soon try to remove such ambiguity to ensure good science.

We distinguish between ontology – as *the essence of being* – and ontological commitment – an extant commitment of a group of workers in a specialism as to what that *essence* is and how the essence manifests itself. The commitment shows changes over a period of time, and the change is recorded, howsoever incompletely, in the texts produced by the specialists. What we produce is candidate terminology and ontological commitment, and the statistical metrics we have used, *weirdness* and *collocation strength* metrics – candidate has to be verified and validated by domain experts. Our inputs and outputs are different from other reported systems dedicated to the identification and visualisation of ontology in a specific domain: as compared to other workers in ontology and terminology engineering, we rely far less on our intuition and significantly more on the evidence produced by the domain community. We have presented an algorithm that encompasses the whole life cycle: from the automatic extraction of terms in free texts and onto systems' asserted knowledge representation for automatically populating knowledge bases. The above work will be boosted further by our current efforts in metadata standardisation [28].

Acknowledgements. This work was supported in part by research projects sponsored by the EU (LIRICS: eContent-22236) and by UK research councils: EPSRC (REVEAL: GR/S98450/01). We would like to thank the anonymous reviewers for their comments which we have taken into account in producing this version of the paper.

References

1. Quine, Willard, van Omran.: Theories and Things. Cambridge (Mass) & London: The Belknap Press of Harvard University Press (1981).
2. Orenstein, A.: Willard Van Orman Quine. Twayne, Boston: G. K. Hall. (1977).
3. Laresgoiti, I., Anjewierden, A., Bernaras,A., Corera, J., Schreiber, A. Th., and Wielinga, B. J.: Ontologies as Vehicles for reuse: a mini-experiment. KAW http://ksi.cpsc.ucalgary.ca/KAW/KAW96/laresgoiti/k.html. (1996).
4. Gómez-Pèrez, Asunción, Fernández-López, Mariano., & Corcho, Oscar.: Ontological Engineering. London: Springer-Verlag. 2004.

5. Gillam, L.: Systems of concepts and their extraction from text. Unpublished PhD thesis, University of Surrey. (2004) http://portal.surrey.ac.uk/pls/portal/docs/PAGE/COMPUTING/PEOPLE/RESEARCHERS/GILLAM/PUBLICATIONS/PHD.PUBLISH.PDF

6. Gillam, L., Tariq, M. and Ahmad, K.: Terminology and the Construction of Ontology. Terminology. John Benjamins, Amsterdam. Terminology 11:1 (2005), 55–81.

7. Ahmad K.: Pragmatics of Specialist Terms and Terminology Management. In (Ed.) P. Steffens. Machine Translation and the Lexicon. Heidelberg (Germany): Springer. (1995), pp.51-76

8. Ahmad K. and Mussachio, M.T.: Enrico Fermi and the making of the language of nuclear physics. Fachsprache 25 (3-4). (2003), pp120-140.

9. Maedche, A. and Volz, R.: The Ontology Extraction and Maintenance Framework Text-To-Onto. Workshop on Integrating Data Mining and Knowledge Management. California, USA (2001)

10. Maedche, A. and Staab, S: Ontology Learning. In S. Staab & R. Studer (eds.): Handbook on Ontologies in Information Systems. Heidelberg: Springer (2003).

11. Faure, D. and Nédellec, C.: Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM. LNCS 1621. Springer-Verlag, Heidelberg. (1999) 329-334.

12. Faure, D. and Nédellec, C.: ASIUM: Learning subcategorization frames and restrictions of selection. In Y. Kodratoff, (Ed.), 10th Conference on Machine Learning (ECML 98), Workshop on Text Mining, Chemnitz, Germany. (1998).

13. Mikheev, A. and Finch, S.: A Workbench for Acquisition of Ontological Knowledge from Natural Text. In Proc. of the 7th conference of the European Chapter for Computational Linguistics (EACL'95), Dublin, Ireland. (1995) 194-201.

14. Ahmad, K. and Davies, A.E.: Weirdness in Special-language Text: Welsh Radioactive Chemicals Texts as an Exemplar. Internationales Institut får Terminologieforschung Journal 5(2). (1994) 22-52.

15. Aston, G. and Burnard, L.: The BNC Handbook: Exploring the British National Corpus. Edinburgh University Press (1998).

16. Quirk, R.: Grammatical and Lexical Variance in English. Longman, London & New York (1995)

17. Gale, W. and Church, K. W.: What's wrong with adding one? In Oostdijk, N. and de Haan, P. (eds.): Corpus-Based Research into Languge: In honour of Jan Aarts. Rodopi, Amsterdam (1994), 189-200

18. Smadja, F.: Retrieving collocations from text: Xtract. Computational Linguistics, 19(1). Oxford University Press. (1993), 143-178

19. Zipf, G.K.: Human Behavior and the Principle of Least Effort. Hafner, New York. (1949).

20. Noy, N.F.; Musen, M.A.: Ontology versioning in an ontology management framework. Intelligent Systems 19 (4). IEEE Press (2004), 6-13

21. Magnusson, C. and Vanharanta, H.: Visualizing Sequences of Texts Using Collocational Networks. In Perner, P. and Rosenfeld, A. (Eds): MLDM 2003, LNAI 2734 Springer-Verlag, Heidelberg. (2003) 276-283

22. Gillam, L., Ahmad, L., Dalby, D. and Cox, C.: Knowledge Exchange and Terminology Interchange: The role of standards. In Proceedings of Translating and the Computer 24. ISBN 0 85142 476 7 (2002).

23. Noy, N. F., Klein, M.: Ontology evolution: Not the Same as Schema Evolution. Knowledge and Information Systems, 5 (2003).

24. Ahmad, K., Tariq, M., Vrusias, B. and Handy, C.:. Corpus-Based Thesaurus Construction for Image Retrieval in Specialist Domains. ECIR 2003, LNCS 2633. Springer Verlag, Heidelberg (2003), 502-510.

25. Gillam, L. and Ahmad, K.: Sharing the knowledge of experts. Fachsprache 24(1-2). (2002), 2-19.

26. Gillam, L. (Ed): Terminology and Knowledge Engineering: making money in the financial services industry. Proceedings of workshop at 2002 conference on Terminology and Knowledge Engineering. (2002).

27. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. Proceedings of the Fourteenth International Conference on Computational Linguistics. Nantes, France. (1992), 539–545

28. Gillam, L.: Metadata descriptors: ISO standards for terminology and other language resources. Proc. of 1st International e-Social Science Conference. (2005).