A natural language approach to information management: tracking scientific advances through the structure of words

Andrew Hippisley and Chara Karavasili

University of Surrey
Guildford, Surrey, UK, GU2 7XH
a.hippisley@surrey.ac.uk

Abstract

In scientific texts specialist words, or terms, express conceptual knowledge. We show that by looking at the use of a term and its family of derivatives over time we can have a tangible picture of how an underlying concept has evolved in scientific advances. This is because the structure of a word encases a core idea and how that idea has been extended in a particular direction. This paper is an outline of a research programme with some preliminary results from an analysis the term nucleus in a nine million word corpus of nuclear physics articles written over thirty-six years, from 1969 to the present day, representing a body of specialist knowledge of recognized growth over time.

Introduction

It has been demonstrated that text-based information management can benefit from insights from theoretical studies of language. In the 'natural language' approach to information retrieval / extraction there are cases where simple string-based methods have been enhanced by considering the grammatical structures in which key words occur in order to "uncover certain critical *semantic* aspects of document content" (Strzalkowski et al. 1999). We will present a natural language approach to uncovering scientific developments over time. We look specifically at the diachronic changes in scientific terminology. If conceptual knowledge is expressed by words, then emerging concepts require new words (see Felber 1984). In some cases an existing word denoting an already established concept is used as a base of a new word for a related, derivative concept. This is possible because the grammar of a language provides the mechanisms for deriving a new word from an existing word. By referencing this system we can show how the layering of a word discloses the central idea it is based on, and how that idea has been extended in a particular direction. For example the relation between *magnet* and *magnetise* indicates a move from object to process, and that between *magnetise* and *demagnetise* expresses a process and its reversal. This is an example affixal word formation. However word formation can also be a product of compounding, for example *nucleus* derives *heavy nucleus* and *light nucleus*, two specific kinds of nucleus. The scientific knowledge we will consider is nuclear physics, an area where definable conceptual shifts have taken place, and the principle medium through which these shifts are discernable is natural language texts.

Natural language approach: word structure

We assign terms to derivational families, sets of word that are structurally related according to established word formation and compounding rules. Terms may be related in two ways: through affixation, for example *magnet* → *magnetise*, or compounding, for example *nucleus* → *proton nucleus*. In cases of affixation we can make use of lexeme-based word formation rules (WFRs) which serve as abstractions over the relationship between morphologically related terms (see Aronoff 1976, 1994). The WFR for *magnet* and *magnetise* is given below It records the changes in a base term at the three levels of description: at the formal level the *–ise* suffix has to been added to the stem; at the syntactic level the class has changed from Noun to Verb; and finally at the semantic level the meaning of the base term has been modified in some way.

'-ise' Word Formation Rule

Base term	→	Derived term
Form:	/stem-/	/stem-ise/
Syntax: Noun		Verb
Semantics:	*'X'*	'make become *X*'

In cases of compounding we will make use of the head-modifier principle where the linear arrangement of the elements may reflect how the compound term is related to its base. The rightmost element, identified as the head, acts to name the general (semantic) category to which the whole word belongs; the leftmost element, the modifier, distinguishes this member from other members of the same category (Zwicky 1993). Thus for *proton nucleus* the category of NUCLEUS has sub-categories of which PROTON NUCLEUS is one.

Method

A preliminary step is to determine which physics terms to use for our investigation. Our choice will be guided by the terminology used in the corpora of writings of two founding fathers of nuclear physics. Using data on frequency distributions as a guide, and the morphological simplicity of their structure, base key terms considered to underlie the terminology that has developed will be elicited. Base terms will be morphologically simpler and head a derivational family of terms. Members of the base's family will then be supplied by morphologically

related terms in Rennie's (2002) *Dictionary of Atomic and Nuclear Physics*, an exhaustive list of some 2000 terms found in writings on nuclear physics from its origins to the present day.

We also have to specify the morphological relationship between terms belonging to a derivational family, WFRs for affixallly related terms and compound rules for compound terms. The extensive work carried out on English derivational morphology could provide the specification of these rules: for example Marchand (1969), Aronoff (1976), Bauer (1983), Beard (1995), and Flood's (1960) collection of affixes used in scientific terms.

Hypotheses

To explore the relationship between word structure directionality and diachronic directionality we wish to test a number of hypotheses that link frequency distributions of morphologically related terms with developments in the concepts that underlie them.

Hypothesis H1A: There is a relationship between morphological layering and text date
If an increase in morphological layers of a term maps onto the extensions of the underlying concept we expect there to be a relationship between the increase of morphological layers of a term and an increase in time. Given a base-derivative pair, the dates of texts in which the base and derivative appear may overlap but the first appearance of the derivative will not precede the first appearance of the base

The related hypothesis is H1B:

Hypothesis H1B: There is a relationship between frequency, morphological layering and text date
Given the summed frequency of occurrences of a derivative and its base in the earliest texts, an increase in proportion of derivative occurrences will be related to an increase in dates of the texts in which they are found.

Hypothesis H2: There is a relationship between term frequency, date of text and the establishment of a concept
When comparing the number of occurrences of a key term K in the earliest text it appears in, with the number of occurrences of the same key term in texts of later dates, KF2, KF3...KFn, then an increase in the KFn : KF1 proportion as n increases indicates the further establishment of the underlying concept. The assumption is that the association between the keyword's primary frequency KF1 and its subsequent 'temporal' frequencies KFn calculated for K's occurrences in texts of incrementally later time periods is related to the evolution of the concept conveyed by the keyword K.

Hypothesis H3: There is a relationship between the increase in size of a base term's derivational family and the establishment of the concept that underlies the base
A registered increase within the diachronic corpus of the number of types that are morphologically related to a base term is evidence of a shift of the concept underlying the base term towards the core of the specialist domain. This is partly based on Dixon's (1982) lexical assimilation claim that the more assimilated a word is in the language's lexicon the larger its number of derivatives; Baayen et al. (1997) present frequency and psycholinguistic evidence in support of this. Lexical assimilation measured in this way is evidence of the establishment of the concept underlying the base word.

The preliminary results we report here relate to the last of these hypotheses, Hypothesis H2.

Pilot study

Of the hypotheses presented above we report on preliminary results of Hypothesis H2, the frequency of occurrences of a term will increase diachronically suggesting the concept underlying the term is becoming embedded in the discipline. As a pilot study we selected *nucleus* (*nuclei* plural) as the key term.

Corpus

To do this we have compiled a corpus of nuclear physics articles published between 1969 and the present day. Specifically the corpus contains the abstracts of journal papers on nuclear physics published by *Physical Review*, a major physics journal. Note that the use of abstracts rather than full texts is not unusual in information retrieval / extraction as abstracts can be viewed as 'handy full document distillations' (Sparck Jones and Willett 1997: 90). The corpus contains n abstracts with a total token frequency of just over nine million.

To organise the corpus diachronically we have referenced the month and year information that is provided with the abstract. We have structured the corpus into diachronic stages of around ten years. Table 2 gives the diachronic stages with number of abstracts and token frequency at each stage.

	Stage 1	*Stage 2*	*Stage 3*	*Stage 4*
dates	1969-1978	1979-1988	1989-1998	1999-2004
abstracts	9840	12734	15273	9538
tokens	1,638,936	2,458,522	3,081,303	1,970,031

Table 1: Diachronic stages

From the table we can see that stages have very different token frequencies. Thus Stage 3 has over three million tokens whereas stage 1 has less than one million. This means that we must be careful to compare not only the absolute frequencies of key terms occurrence at each stage but the proportions of occurrences of the key term at each stage. Finally the last Stage, Stage 4, is 'incomplete', representing just over five years whereas other stages represent 10 years.

Corpus analysis

We restricted the study to members of the key term's derivational family that were strictly compounds. And within the set of compounds we looked only at those headed by *nucleus*, i.e. ignoring examples where *nucleus* is the modifier. Even with these restrictions the size of the derivational family is considerable in the corpus we examined: there are over fifty collocations which could be viewed as compound terms, and many more which are

frequently occurring modified NPs some of which could have compound status.[1] In (1) we give a small sample of the derivational family of *nucleus*.

(1) Nucleus
compound nucleus
deformed nucleus
finite nucleus
heavy nucleus
light nucleus
odd nucleus
proton nucleus
residual nucleus
…

Evidence suggesting that the concept expressed by *nucleus* has evolved in the discipline will be an increase in frequency of occurrences of its derived forms, i.e. members of the derivational family. Part of this evidence comes from comparing the frequency of nucleus as a single word term with the frequency of nucleus as part of a multi-word term. In other words we can compare the frequency of nucleus as string and the frequency of nucleus as a substring. The comparison is given in Table 2 and shows that close to all occurrences of the string nucleus are as substrings of complex expressions. The suggestion is the general development of the core concept in various directions.

Overall freq.	Freq. as string i.e. *nucleus*	Freq. as substring, e.g. *proton nucleus*	Proportion as substring
83,935	2803	81132	97%

Table 3: *nucleus* as string and substring

The development of the concept is expressed as complex forms rooted in the form that express the concept. In this study the complex forms are compounds headed by the concept's expression. If the complex form expresses the particular direction in which the core concept has developed we can see how the specific developed concepts themselves develop. A chronological increase, or decrease, of occurrences of these headed compounds will suggest that a particular elaboration, or development, of the concept has itself become more, or less, established in the discipline. In Table 1 above we showed how the diachronic corpus has been split into chronological stages. By splitting up the corpus in this way we can compare the frequency of occurrences of a compound terms at each stage. Table 3 traces the chronological frequencies of a sample of compounds headed by *nucleus*. For the dates of the stages, and token frequency of each stage, please refer to Table 1. The first column describes the compound in terms of its modifier. Thus the for the first example we have frequency information for the multi-word term *compound nucleus*. The table gives the absolute frequency of occurrences, but due to differences in token size for each stage it also gives the proportion of the stage's tokens that is represented by the key term. So for compound nucleus the 463 occurrences at Stage 1 represents 0.028% of all tokens of this stage but the 485 occurrences at Stage 3, a much larger sub-corpus, represents only 0.016%. Finally it should be noted that the frequency includes both singular and plural forms, i.e. *compound nucleus* and *compound nuclei*.[2]

	Stage 1	*Stage 2*	*Stage 3*	*Stage 4*
modifier				
compound				
abs. freq.	463	423	485	218
%	0.028	0.017	0.016	0.011
deformed				
abs. freq.	95	215	821	797
%	0.006	0.009	0.027	0.040
finite				
abs. freq.	40	93	300	144
%	0.002	0.004	0.010	0.007
heavy				
abs. freq.	84	105	249	112
%	0.005	0.004	0.008	0.006
light				
abs. freq.	100	200	255	108
%	0.006	0.008	0.008	0.005
odd				
abs. freq.	24	35	172	160
%	0.001	0.001	0.006	0.008
proton				
abs. freq.	232	1886	806	555
%	0.014	0.077	0.026	0.028
residual				
abs. freq.	94	153	119	57
%	0.006	0.006	0.004	0.003

From the table we see that the picture is mixed. Some terms increase their frequency over time, some show a decrease, and some appear to peak and then tail off.

Increase in frequency

When we compare Stage 1 and Stage 3 we see some of the terms show a there is an increase in frequency that follows the chronology. This is true for *deformed nucleus* which dramatically moves from 0.006% to 0.027%, *finite nucleus* (from 0.002% to 0.010%), *heavy nucleus* (0.005% to 0.008%), *odd nucleus* (0.001% to 0.008%), and *proton nucleus* (0.014% to 0.026%). For two of these, *deformed nucleus* and *proton nucleus*, the trend continues to Stage 4 with *deformed nucleus* again seeing the most dramatic rise: it has a 0.006% proportion at Stage 1 and by Stage 4 this has increased to 0.040%. This can be shown in the dispersion plot for the key term (generated from *WordSmith)* in (2).

(2)

1 Distinguishing compounds from phrases is not always straightforward. See Lieber and Sproat (1992) for a detailed X-bar approach to distinguish 'true' compounds which are lexical objects, from phrasal categories, which are syntactic objects.

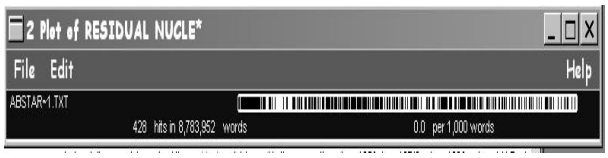
2 In fact the plural form of these compounds has a much higher token frequency than the singular.

.

Decrease in frequency

A new elaboration of a core concept may also lose prominence in a discipline over time, and there is some evidence of this in our data set. For example *compound nucleus* has the relatively proportion of 0.028% at Stage 1 but this decreases at each subsequent stage. A similar decline is see in *residual nucleus*, for which we show the dispersion plot in (3).

(3)



Peaking in frequency

Finally there are case where terms' use appears to peak and then tail off. One such case *proton nucleus* which dramatically increases from Stage 1 to Stage 2 where it peaks with a relatively high proportion of 077% then decreases significantly at the subsequent stages.

This small sample shows the range of possibilities. Though not represented here there is a fourth frequency behaviour that ought to be mentioned. There is a high number of compound terms headed by *nucleus* that appear at one stage only and never appear again. This suggestions a difficulty in establishing a particular concept within the discipline. An example is *liquid nucleus* which appears at the first Stage but not any subsequent stages.

Summary

We have given a brief overview of a programme of research that takes a natural language approach to information management by seeing the parallel between a word and the concept it conveys, and word's derivative, affixal or compound structure as a tangible index of elaboration or development of the embedded concept. The assumption is that knowledge is chiefly disseminated through natural language texts. Over time, concepts conveyed by texts evolve, as registered by the *knowledge life-cycle*. The linguistic properties in text, such as the rules governing the structure of words, play a valuable methodological role for leveraging embedded knowledge, or content. This can demonstrated by a text analysis of a diachronic corpus that considers the structural relationship between terms on the one hand, and their frequency distributions over time on the other.

References

Aronoff, Mark (1976). *Word Formation in Generative Grammar*. Cambridge, Mass: MIT Press.

Aronoff, Mark (1994). *Morphology by itself: stems and inflectional classes*. Cambridge: Cambridge University Press.

Bauer, Laurie (1983). *English Word-Formation.* Cambridge: Cambridge University Press.

Beard, Robert (1995). *Lexeme-Morpheme Base Morphology: A General Theory of Inflection and Word Formation.* Albany: State University of New York Press.

Felber, Helmut (1984). *Terminology Manual.* Paris: UNESLO International Information Centre for Terminology.

Flood, W. E. (1960). *Scientific words: their structure and meaning*. London: Oldbourne.

Marchand, H. (1969). The Categories and Types of Present-Day English Word-Formation. München: Beck.

Rennie, Richard (2001). *Facts on File Dictionary of Atomic and Nuclear Physics*. New York: Facts on File.

Strzalkowski, T. (1999). *Natural Language Information Retrieval*. Dordrecht: Kluwer.

Zwicky, Arnold (1993). Heads, bases and functors. In: Fraser, Corbett and McGlashan (eds) *Heads in Grammatical Theory*. Cambridge: Cambridge University Press. 292-315.

Martin, L.E. (1990). Knowledge Extraction. In Proceedings of the Twelfth Annual Conference of the Cognitive Science Society (pp. 252--262). Hillsdale, NJ: Lawrence Erlbaum Associates.
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx