THE SCIENCE OF SCIENCE IN SPACE

Christoph Haase
University of J.E. Purkyně, Ústí nad Labem

This overview article focuses on epistemological questions raised in the context of academic reflection upon the outcome of research and its representation in actual texts. It tries to develop a dichotomy between formal and functional approaches and illustrates this by reviewing the results of several years of research on the SPACE corpus. This new typer of parallel corpus will be described and a few results published elsewhere reviewed under the vantage point of developing a "Science of science". Further, it will be demonstrated that the scientific method and the application of rational thought itself depend on linguistic structures. Its study can be supported by large academic corpora and their processing.

Introduction

The question of representing academic content has always been a question of the representation of complexity. When research started on a corpus known under the acronym SPACE we had the intention to look at academic domains which have a built-in complexity. The intention included an initial doubt that had been raised elsewhere concerning a hypothesis which is known as the complexity hypothesis. The complexity hypothesis (cf. Haase 2010b among others) explains that the language used to discuss an object of study should proportionally reflect the complexity of its object of study. This would mean that the most complex research topic that humankind knows of should merit the use of the most complex language. Those most complex ideas today are found almost exclusively neither in the field of linguistics, nor in the field of language studies or social sciences, but they involve an extremely high level of abstraction and fields of math that had to be invented in order to describe the phenomena at hand. The thought structures necessary involve an incredible depth of bringing transparency to abstraction in fields like quantum theory and cosmology.

The linguistic underpinnings for the decision to compile a corpus with the acronym SPACE (Specialized and Popular Academic Corpus of English) were laid out in Haase, 2007 (published in REAL Studies 2). The corpus contains a number of texts that initially take us back to a few ideas about general discussion of the theory of science and what could be considered a "science of science".

If the linguistic features in the language of the most advanced practitioners in the natural sciences are compared and the question is raised whether they also have beautiful, complex, and creative language to match their research topics, the answer to that is, in all probability no.

This contribution intends to point out a few reasons why this might be the case. This takes us to the concept of "aboutness" in science.

What we can expect from academic texts in general is that the research culture demands them to be objective because any error that is necessarily part of the research process is ideally small: errors introduced by the experimenter, by miswritten judgments or wrong interpretation. The objective information to be imparted is empirically found to be reified in nominalizations, a very conflated and dense style, an overuse of passive and a high degree of semantic packing. This is a common denominator as demonstrated below:

> 0082PN GFP expression observed in the gustatory neurons of the labial palps and leg tarsal segments (Fig. 1 C and D) was suppressed by targeted GAL80 expression (data not shown), as expected from the previous observation that the 3.3-kb Cha regulatory DNA directs gene expression in most if not all chemosensory neurons in the peripheral nervous system (23, 41). Concomitant with the further restriction of the GAL4 activity in C309 by the Cha3.3kb-GAL80 construct, the temperature-induced courtship chain formation and head-to-head interactions were suppressed completely

The example text from SPACE (#0082PN) has no agents in the subject positions of its propositional structures, it does not express agency (who does what) - it expresses factuality (what happened). This is the common stylistic denominator of academic writing. For a linguistic analysis, this is the secondary part, because linguists are not the expert target group (which in this case is genetics). The other common denominator is that in the natural sciences, there is little space for subjectivity.

The common ground can be summarized as follows:

- relatively few markers of subjectivity in natural sciences
- thus: objective account of the author's involvement/participation
- author commitment: often stereotypically lexicalized (in modal adverbs etc.)

The third point returns to objectivity through the backdoor: The author's commitment or involvement should be objective in a conventionalized sense. It can be scaled by hedging and very often this happens stereotypically. This means not that authors really scale their judgment down or up to a level they are really convinced of but to a level that is expected by the requirements of the genre and the text type, in one word: convention. For instance, by using the expected modal adverbs the linguistic scaling represents a compromise between the intended and the expected. Since there is not an infinite amount of modal adverbs the stereotypical lexicalization is repeated in most texts.

A number of other markers can be found in the following example:

> AX0039 indicate a presence well within … current observation bounds could cause early star formation at a level sufficient to explain the high reionization redshift

Here, the reader draws conclusions from involvement and commitment cues. They are studied in closer detail as modal adverbials (for example in Haase, 2012) or as hedge expressions, cf. Beyer (this volume).

Science and "aboutness"

An epistemological battleground

In what way does this relate to linguistic analysis? Two major (and a few minor) reasons will be suggested in this contribution which, to my knowledge, has never been brought in context with academic writing. The initial point of this is concerned with what academic language does for us as practicioners on the one hand. On the other hand, we need to see what we can do with it and exactly how we do it. Is the language we use therefore really *about* the science that we practice or is it more a reflection of ourselves? This generates the two major approaches.

The formal approach

The first approach, which is probably the less creative, says that language is simply one part of many other cognitive skills; one other cognitive skill is for example rational thought.This however, may also be equivalently expressed in other modules of cognition. That means that mathematics for instance is a short hand for a conceptualization of something that really, phenomenologically *happens* in nature.

In sum:

- language is only one module among other cognitive skills
- rational thought and scientific modeling may rely on other modules
- math is a shorthand for a conceptualized thought process directly related to nature
- numbers are "out there"
- extraterrestrial civilizations will have the same math
- scientific revolutions resemble "glimpsed" shortcuts

Thus: the role of language is at best secondary.

The falsifiability of this approach is probably low even though the main protagonists –Feynman or Penrose in the natural sciences, Johnson-Laird and others in the social sciences (psychology) are on the more abstract end of the continuum. The falsifiability is low because this approach is not free of its own mysticism as in its conception of free will as a quantum phenomenon and intelligence not as an objectivist, plannable resource but as the ability to find shortcuts in the description of the fabric of reality.

Our number system – as anthropomorphic as it might appear – reflects reality directly, not conceptually. Prime numbers or π are *out there* in nature, we have not created them, they are not a construal of our sociopsychological personæ. Should we find extraterrestrial civilizations, their math would rely on the same fundamentals. Anything that we find in a process of scientific discovery is a shortcut within these configurations.

This means for language that it is at best secondary and again this approach is very hard to falsify because no means of comparison can be given. This is in no way a fringe assumption as luminaries like Richard Feynman prove.

To come back to academic writing then, does this mean that the scientific mind is somehow lost in space? Does it try to wrap itself around phenomena that are not really part of language but that are glimpsed hints of reality? If there is some truth in this approach then the study of academic English or academic language may as well come to a halt.

The functional approach

There is an alternative approach to the previous paragraph. According to this approach, language is "about" the world, thus science follows from rational configuration (and re-configuration) of linguistic objects in the minds of the practitioners of science.

This approach can be summed up as follows:

- math is a language
- numbers are discrete representations of human body plans
- extraterrestrial civilizations will have fundamentally different math/science
- scientific revolutions can be planned
- thus: the role of language is primary
- falsifiability of this approach:high
- protagonists: Fodor, Dennett, Vienna Circle, Cognitivists, basically all people who study academic language

If language is actually about the world and science relates to the constant reconfiguration of linguistic objects that take place in the mind, then science is also *in* the language.We can see this if we agree with the assumption that math is a language, that numbers represent something that emerges out of the human body plan. We have a decimal number system that relates to the ten fingers for instance. If humans had eight fingers, we would in all probability have a octal number system. In this way, our numbers are a representation of the human body plan. Should we discover extraterrestrial civilizations, their math would be completely different from ours. This also means that via language we can plan our scientific revolutions. Language takes the primary role, it can be falsified because we have the test, we can see if we are

successful with them, why not, and notable academics stand for this approach: the Vienna circle with Wittgenstein, basically all people in cognitive linguistics and also all protagonists who study academic language.

Obviously, science is successful in both approaches given that Richard Feynman counts among the superior minds of second half of the 20th century. Therefore the matter is not really decided. For us, the job is to make corpora and run tests on texts that actually occur in academia.

Questions (among others) a corpus can help answering

Among the questions that a corpus such asSPACE can help answering, the following subset is related to the epistemological discussion.

a) Which forms are "about" which study objects/processes?

b) Which forms are "about" which truth values assigned to a)?

c) Are the linguistic structures isomorphous with the scientific phenomena?

d) If yes, can the linguistic structure somehow be optimized?

e) Thus, "better" language leads to better science?

f) If yes, is linguistics "the science of science"?

To break down the first question to corpus level considerations we need to look at the lexical items involved that describe the objects (nominal expressions) and processes (verbal expressions). A glance at the academic wordlist and any other frequency list generated out of SPACE shows that the specificity of the objects is mirrored by highly infrequent lexical material from expert knowledge. The study of this is covered in Haase 2009c on lexical-semantic criteria.

The truth values that are assigned to the processes by way of quantification open a very systematic escape clause for the researcher: Independently of the question whether the truth value is a quantificational process in the mind of the beholder or whether there is binary truth (0 and 1, false and true), modality and hedging are systematic and conventionalized ways (see Haase 2011c, 2011b and 2008f.)

a) and b) together enable us then to ask the following:

Are linguistic structures isomorphous with the scientific phenomena? What re-appears here is again the complexity question through the backdoor. If the phenomena are complex, which they undoubtedly are, then the linguistic structures used should also be complex. If the answer to c) then is yes, by consequent and evolutive conventionalization, can the use of the linguistic structures be optimized? The answer to this question opens up a wide field that involves not only epistemological but also ideological aspects. In his seminal paper on English as an academic language Swales terms it to be either a "Tyrannosaurus rex" or the "triumphalist" mode of expression for the global academic gatekeepers (Swales 1997: 376). He may overlook however

that a Babylonian tapestry of academic publications in the native languages of their practitioners is actually a hindrance for science. In that sense, the English language does represent an effective and highly optimized code for academic interchange on its own and independently. The genre approach in which the genre is actually owned by the research community that practices it goes a long way in conventionalizing its mechanisms and thus facilitate understanding.

Finally, the grammaticalization and conceptualization of space (Haase & Schmied 2011a) and causation is deeply ingrained in the language (Haase 2010a). By obtaining a direct or indirect mapping between phenomenological causes and effects to linguistic structures, the linguistic structures actually represent the causality they describe (Haase 2009b, 2009a).

In the end, if the answer to all questions is positive and a better language is in service of better science then linguistics may be considered the science of science.

SPACE – A brief overview

In the latest corpus built (v.02 from 2011) we have added an amount of around 800,000 new words, building up from sciences: From the physical field and from the bio sciences field, thus it shows a relatively strict separation into a hard and a “soft” branch. These new additions are original publications, which are partly free from copyright:

from arXiv, a pre-print server fro rapid, non-peer-reviewed access. (fig. 1)

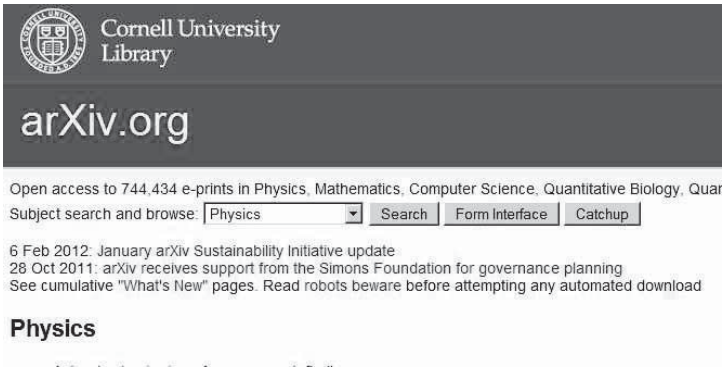


Fig. 1: The *arXiv* website

and from the public Proceedings of the National Academy of Sciences (PNAS), science which is supported by the National Science Foundation of the United States. The research results that originate from publicly funded research are therefore public domain and the (peer-reviewed) articles are free of copyrights (fig. 2.)

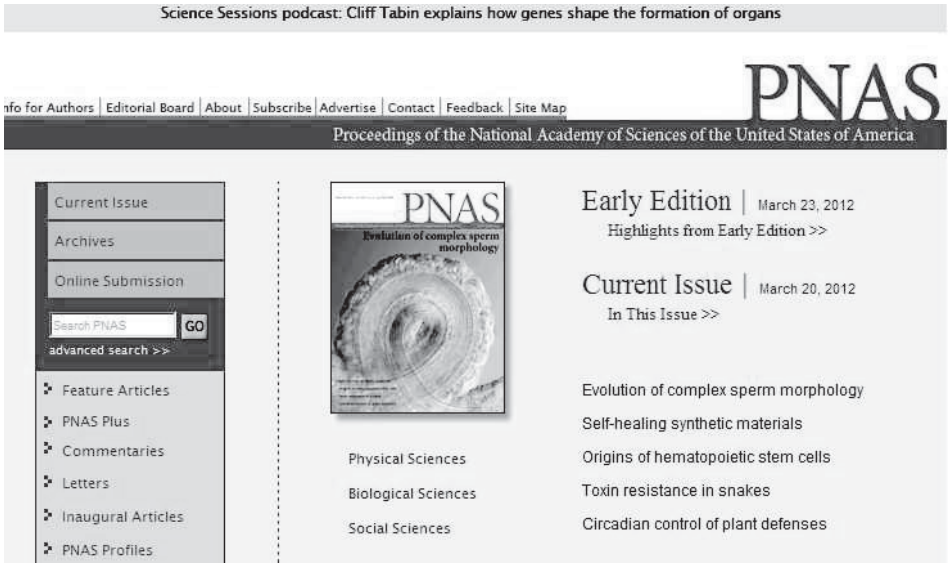


Fig. 2: The *PNAS* website

To parallel the science texts with a means of comparison and also to make it slightly more palatable to students who are neither biologists nor physicists, a parallel structure has been built into the corpus: the so-called popular component. The popular component is exclusively from the *New Scientist*, the leading popular academic journal world-wide today (fig. 3).

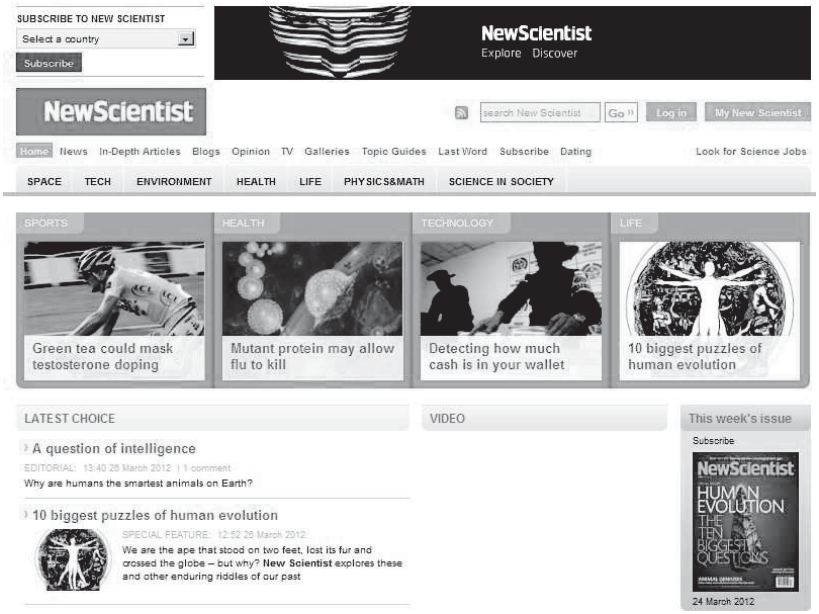
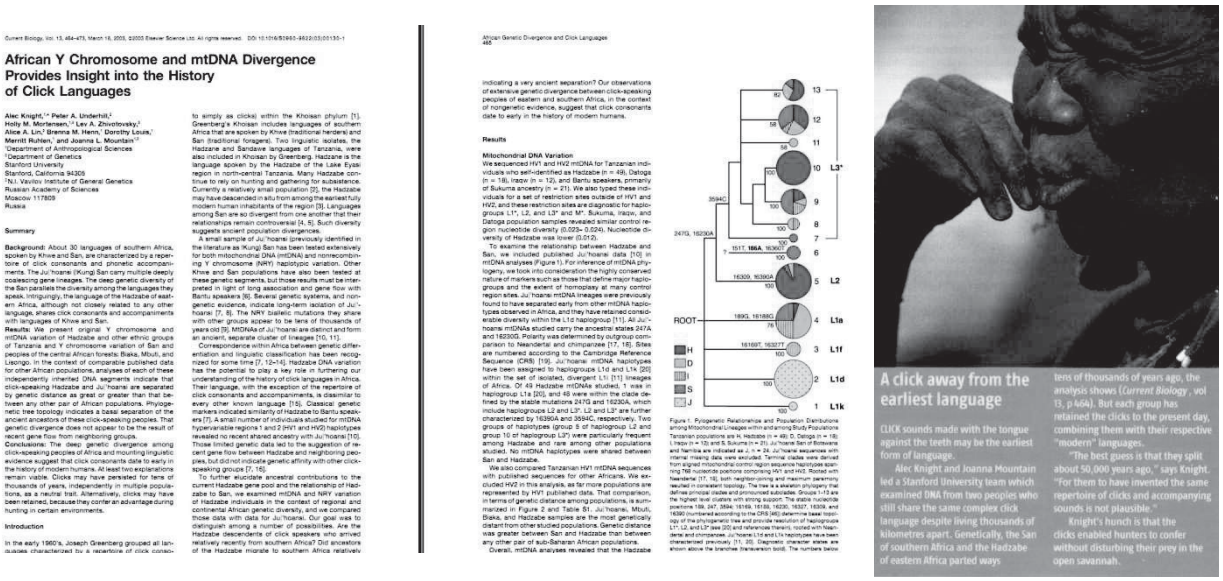


Fig. 3: The *NEW SCIENTIST* website

When the different text types with similar content are compared, even at first glance significant differences can be observed in quantity and presentation. Shown below are the layouts of two articles on click languages, the original article and its popularized counterpart. It is obvious that in the second version this text has been condensed down to a very short summary with a picture-to-text ratio of about 60:40. The picture is a stock photo and unrelated to the original research.

Even though the titles are very different, at least both deal explicitly with click languages but the second title is definitely more palatable to a non-expert in this field.

The different paragraphs in the original article indicate a general tendency of integrating the different text types in academic writing. It starts with a section called *summary* (where in most publications an *abstract* would be expected), continues with one column/half a page of introduction and adds a relatively complex diagram on page two of this 2-column layouted text.



African Y Chromosome and mtDNA Divergence Provides Insight into the History of Click Languages

A click away from the earliest language

Fig. 4: Original and popularized variant (*Current Biology* 13 and *New Scientist*)

A brief breakdown of the fields and domains in the latest incarnation of SPACE can be found in Tab. 1.

Subcorpus	Descriptors	word count
arXiv	physics, astrophysics, quantum mechanics	809,320
New Scientist – physics	physics, astrophysics, computer science, quantum mechanics	203,470
Proceedings of the National Academy of Science (PNAS)	biochemistry, genetics, genetic engineering, microbiology	267,105
New Scientist - biosciences	biochemistry, genetics, genetic engineering, microbiology	30,499
Public Library of Science – Medicine (PLoS),	medicine, virology, clinical psychology, public health	217,254
New Scientist – medicine	medicine, virology, clinical psychology, public health	17,050
Total		**1,544,149**

Tab. 1: SPACE domains and word counts

The description of the sub-corpora is a coarse-grained method to conveniently summarize the specializations of the respective fields. Again, a direct comparison shows how much information is lost when we look at the real domains.This is also related to the broader epistemological question of how much knowledge is generated by means of the ontology itself. This raises the following questions:

Do better ontologies facilitate better science?

Does better mean more complex/fine-grained/sophisticated?

Is there an ontological optimum?

Knowledge generation and knowledge transfer

Ontology benefits

The arrangement of subcorpora indicates a superficial understanding of the classification of research into their respective (albeit shallow) branches. This however represents only an outtake of the ontology suggested and used by the publishers, an ontology in the sense of the fields and subfields that we find in those publications falling under biological sciences in the proceedings of the National Academy.

Ontologies are considered as one of the pillars of the Semantic Web initiative (cf. for example semanticweb.org) in which "complex forms of knowledge organization systems are represented in a machine-readable, formal language" which are needed "to provide the semantic layer for the Web" (Weller 2010: 3). In the SPACE ontology this takes the form of "general *concepts* in this domain" (ibid, italics in the original). This is a function of convenience as it provides easy access to these disciplines. At the same time, this convenience comes at a cost. The generation of knowledge that is achieved by creating a meta discourse of the sciences by linking different parts of research, approaches and also individual researchers is lost in this way.

Knowledge generated by ontologies

The PNAS ontology (Biosciences) is an alphabetical, 1-tier list.

Biological Sciences	Medical Sciences
Agricultural Sciences	Microbiology
Biochemistry	Neuroscience
Cell Biology	Pharmacology
Developmental Biology	Plant Biology
Ecology	Population Biology
Evolution	Psychological and Cognitive Science
Genetics	Sustainability Science
Immunology	Systems Biology

The SPACE ontology (Biosciences) has exactly three items in it:

Microbiology Genetics (aka Molecular biology) Biochemistry

Thus, in direct comparison the PNAS ontology seems much more helpful. The SPACE ontology however has an integrative benefit. It leaves out branches not covered at all by SPACE (like the Medical Sciences) and on the other hand does not suffer from the pitfall of double assignment as many papers would. (E.g. a genetics engineering paper that supports sustainability of agricultural techniques would be difficult to classify). The logic of SPACE is a different one, it takes the granularity of the research objects and transfers it to a granularity of its ontological import: microbiology concerns bacteria and viruses, genetics the building blocks of life (DNA, RNA, thus basically huge and very complex molecules) and biochemistry takes one more step into the world of even smaller components (organic chemicals, partly very simple substances etc.).

More interesting differences emerge when we consider the *arXiv* ontology (physical sciences, "hard" science). Below, only a subset is represented:

Physics

* Astrophysics (astro-ph new, recent, find)

includes: Cosmology and Extragalactic Astrophysics; Earth and Planetary Astrophysics; Galaxy Astrophysics; High Energy Astrophysical Phenomena; Instrumentation and Methods for Astrophysics; Solar and Stellar Astrophysics

* Condensed Matter (cond-mat new, recent, find)

includes: Disordered Systems and Neural Networks; Materials Science; Mesoscale and Nanoscale Physics; Other Condensed Matter; Quantum Gases; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons; Superconductivity

* General Relativity and Quantum Cosmology (gr-qc new, recent, find)
* High Energy Physics - Experiment (hep-ex new, recent, find)
* High Energy Physics - Lattice (hep-lat new, recent, find)
* High Energy Physics - Phenomenology (hep-ph new, recent, find)
* High Energy Physics - Theory (hep-th new, recent, find)
* Mathematical Physics (math-ph new, recent, find)
* Nuclear Experiment (nucl-ex new, recent, find)
* Nuclear Theory (nucl-th new, recent, find)
* Physics (physics new, recent, find)

includes: Accelerator Physics; Atmospheric and Oceanic Physics; Atomic Physics; Atomic and Molecular Clusters; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; Data Analysis, Statistics and Probability; Fluid Dynamics; General Physics; Geophysics; History and Philosophy of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics Education; Physics and Society; Plasma Physics; Popular Physics; Space Physics

* Quantum Physics (quant-ph new, recent, find)

Mathematics

* Mathematics (math new, recent, find)

includes (see detailed description): Algebraic Geometry; Algebraic Topology; Analysis of PDEs; Category Theory; Classical Analysis and ODEs; Combinatorics; Commutative Algebra; Complex Variables; Differential Geometry; Dynamical Systems; Functional Analysis; General Mathematics; General Topology; Geometric Topology; Group Theory; History and Overview; Information Theory; K-Theory and Homology; Logic; Mathematical Physics; Metric Geometry; Number Theory; Numerical Analysis; Operator Algebras; Optimization and Control; Probability; Quantum Algebra; Representation Theory; Rings and Algebras; Spectral Theory; Statistics Theory; Symplectic Geometry

Nonlinear Sciences

* Nonlinear Sciences (nlin new, recent, find)

includes (see detailed description): Adaptation and Self-Organizing Systems; Cellular Automata and Lattice Gases; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons

Computer Science

* Computing Research Repository (CoRR new, recent, find)

includes (see detailed description): Artificial Intelligence; Computation and Language; Computational Complexity; Computational Engineering, Finance, and Science; Computational Geometry; Computer Science and Game Theory; Computer Vision and Pattern Recognition; Computers and Society; Cryptography and Security; Data Structures and Algorithms; Databases; Digital Libraries; Discrete Mathematics; Distributed, Parallel, and Cluster Computing; Emerging Technologies; Formal Languages and Automata Theory; General Literature; Graphics; Hardware Architecture; Human-Computer Interaction; Information Retrieval; Information Theory; Learning; Logic in Computer Science; Mathematical Software; Multiagent Systems; Multimedia; Networking and Internet Architecture; Neural and Evolutionary Computing; Numerical Analysis; Operating Systems; Other Computer Science; Performance; Programming Languages; Robotics; Social and Information Networks; Software Engineering; Sound; Symbolic Computation; Systems and Control

Quantitative Biology

* Quantitative Biology (q-bio new, recent, find)

includes (see detailed description): Biomolecules; Cell Behavior; Genomics; Molecular Networks; Neurons and Cognition; Other Quantitative Biology; Populations and Evolution; Quantitative Methods; Subcellular Processes; Tissues and Organs

The SPACE ontology (physical sciences) looks like this:

Cosmology Particle physics Quantum physics

Again, the rationale starts out with the macrophysical of large dimensions and ends with the most subtle phenomena known to science at the quantum level of description. Here, the overlap is more frequent (micro- and macrocosmos can be linked in intricate ways) but the insightfulness of the short ontology (and this is not so obvious at first glance) is of course that it reflects the fundamental forces in nature: Cosmology, being the science of gravitation, particle physics where no process involves gravity directly

but instead most processes involve the strong nuclear force and quantum physics where the electromagnetic and the also the weak force have a role.

Thus, the structure applied and mapped onto the knowledge represented becomes part of the knowledge itself. A technical variant of this are ontologies within the semantic web initiative: e.g. Dublin Core, a set of standardized semantic metadata:

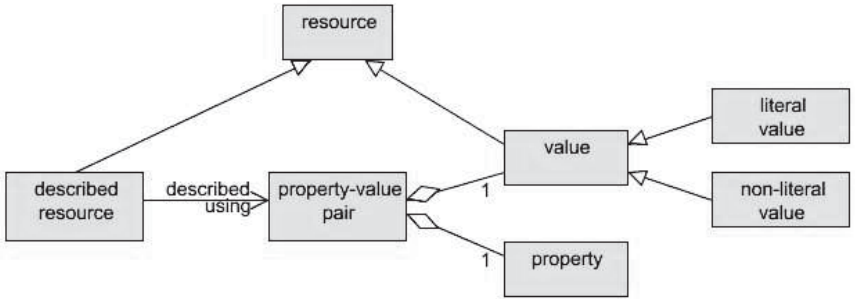


Fig. 5: Dublin Core resource model (from dublincore.org)

The central element is a linguistic one related to predicate calculus: the property-value pair. By assigning values to properties, the property can be scaled in the same way as an utterance can be scaled by modality. It further picks out a resource out of the pool of available resources and transforms it into a described resource. It can then be networked to other resources (top box). These technical realizations of ontologies help to generate knowledge out of the meta-data. Further, they can even be parsed automatically, thus creating networks of components of knowledge.

A case study in ontological research

Measuring word recognition of lexical items with differing degrees of semantic difficulty is relatively easy and leads to repeatable results within the standard model by Marslen-Wilson and others (Marslen-Wilson et al. 1994).

A linguistic ontology basis for text profiling

The high specialization of the lexical items in the original texts and its transformation to a much smaller array of general-academic terms in the popularized texts is interesting from a semantic point of view. It seems obvious that the thrust of the scientific argumentation lies in the use of words. Only highly specialized words enable science. This is intuitively obvious when we consider the difference between base-level categories and prototypes as suggested by Rosch. If we compare the vertical taxonomy by Rosch (see Evans & Green 2006: 256) we find the basic-level categories

with the highest degree of inclusiveness near their prototypes (the horizontal taxonomy), i.e. the categories acquired earliest, used most frequently and recognized and recalled most rapidly. These are words at a level like *dog* and *chair*. They provide surface access and establish a common ground in discourse. They also prevent any kind of scientific thinking. Therefore, if linguistics can at all inform the sciences to facilitate the argumentation the study needs to focus on the ontologically "deep" lexical items with a near-zero degree of inclusiveness. In fact, it is the exclusiveness of these categories that lends them their academic/scientific status.

The following table displays the difference indicated above: The lexical items were extracted from the same base material, a text on "Experimental hints of Gravity in Large Extra Dimensions?" (0007AX). The central column shows the academic items which have little use outside this highly specialized field. The popular text even tries a hand on boosting the message with imprecise but impressive metaphors like *dead stars* and *rogue comets*.

	academic text **0007AX**	**popular academic text** **0007NS**
markers of specialization	*conjectures, compactification, coalescence, planetesimals, angular, mesoscopic, gauge field, accretion, radial drag*	*dead stars, cloud of gas, hot star, proto-planetary disc, rogue comets*
markers of vagueness	*suggest X may have, should detect Rc, deviations are weak, may be turbulent*	*it may be hard, can be slow, they probably rebound, could charge up*

Tab. 2: Semantic complexity and ontological depth

If we therefore assume that the ontological depth can be seen as a marker of the argumentative prowess in an academic text then we can use this to systematize this as a lexico-semantic function and use it in automatic text profiling. We can do this for two reasons. First, it can help compare texts and measure their difficulty and second, to obtain data from recognition tests to match and correlate them with the words that are impressionistically felt to be hard. In an additional step the text could then be re-phrased by the author.

In order to make this feasible, a very solid and extensive data basis was needed. Within the SPACE project we settled on WordNet (www.wordnet.princeton.edu) because it can be implemented freely and with relative ease.

An entry from WordNet is displayed below:

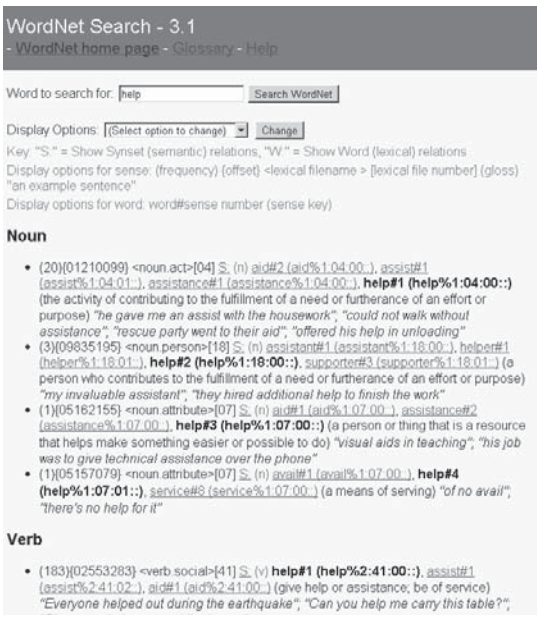


Fig. 6: A WordNet entry

Complexity Analyzer (Complexana)

Using the WordNet implementation as a basis we developed a tool that enables a quantitative comparison with different texts. The tool was written in Perl, using a license-free Linux implementation of WordNet in a compiled form. This implementation is a working environment that can be packaged up and bundled with the tool in a self-contained executable file.

The application requires the input of a raw text (txt format) as a user interaction. The first step uses a part-of-speech tagger (the free TreeTagger was fully implemented) to tag the entire text. TreeTagger provides overall robust accuracy and is even superior when nouns are concerned. At the same time types and tokens in the texts are counted. The tagging is the first process because ComplexAna uses exclusively the nominal items to profile the semantic complexity of the text. The tagged file is saved.

In the second step ComplexAna extracts all nominal items that were identified in the tagged text. These items are written in a separate file. We also added extended functionality for stoplists and better control for excluding items that generate false scores (discussed in section 4).

In the third step all nominal items from the text are queried in the implemented version of WordNet. From the query results the position of the item in the ontology, its so-called semantic depth is calculated. This score is coupled with a series of terms that

are also calculated in dependence on the result of the WordNet query. These can be seen in Fig. 7:

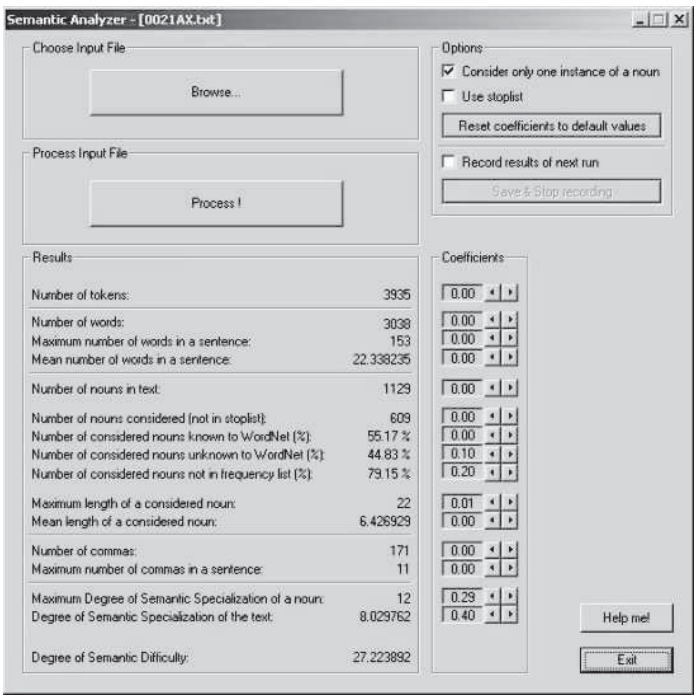


Fig. 7: ComplexAna v.1.2

The parameters are used for correction terms that influence the main parameter, the degree of semantic specialization of the nouns.

Finally, a single score is calculated that summarises the semantic complexity of the text. This is a dimensionless number. It works only in comparison with the numbers obtained from other texts.

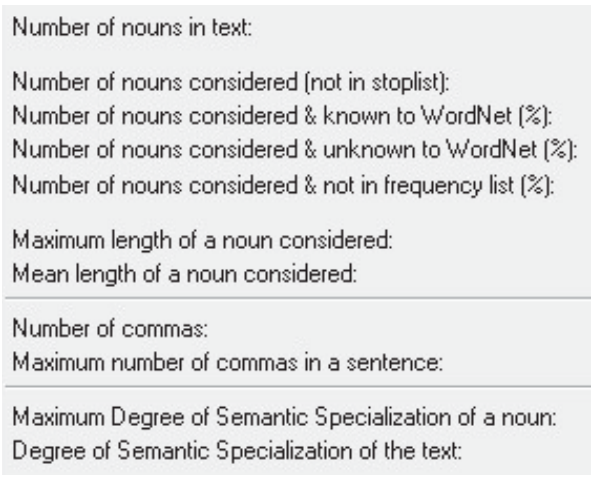


Fig. 8: Nominal parameters for automatic semantic profiling in ComplexAna