The semantics of continuous quantities in natural languages: models and applications in information extraction and integration

Shenghui Wang, David E. Rydeheard, Mary McGee Wood and David S. Brée

School of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL, U.K.
{shenghui.wang, david.e.rydeheard, m.m.wood, david.s.bree}@manchester.ac.uk

Abstract. We investigate the semantics of continuous quantities such as colour and shape as they occur in natural languages. To do so, we introduce a mathematical framework which allows us to define semantics uniformly across different continuous quantities. The original motivation for this was the analysis of texts in descriptive sciences and we begin a series of experiments testing proposed semantic interpretations against these texts and the natural phenomena being described. Finally, we describe applications of the semantics in information extraction and integration.

1 Introduction

Natural languages are finitary, in having a finite (if changing) vocabulary and a finite number of formation rules for sentences, whereas the physical world appears to be continuous. One of the major roles of natural language is to describe the physical world and yet there is clearly a mismatch between the resources of language and the variation of continuous quantities in nature. This becomes particularly acute for the descriptive sciences where precise description of colour, shape, sound, texture, and spatial and temporal quantities and their variation is essential.

The way we build expressions in natural languages for describing continuous quantities uses a very limited vocabulary for phrase formation and is the same over a wide range of such quantities. That is, there is a common terminology, independent of the quantity under consideration, for describing, for example, ranges of values, intermediate values, modified and transformed values, proximity and equivalence of values, combinations of values of different quantities, and logical expressions involving values and ranges.

In this paper, we consider the proposal that not only is there a common terminology across continuous quantities, but also a common semantics of these expressions. That is, there is a semantics not simply of particular phrases, but a generic account giving a meaning to each phrase formation independent of both the choice of quantity we are describing and of the choice of model for interpreting a particular quantity. Different interpretations arise from different choices of model but each via a uniform schema of interpretation.

Our interest in this topic arose from a very practical problem. In the descriptive sciences, such as botany (our primary example), there are large corpora of written knowledge. In botany, in particular, there is a vast collection of parallel (and often independent) descriptions with the same species described by many different authors. Because of this large and disparate body of text, it is tempting to try to automate some of the analysis. With such an automation, we can compare different descriptions to assess how much they agree, we can collate and compare knowledge, and we can provide tools for information extraction and integration. However, to achieve all of this, we need to be able to analyse the constituent descriptions, which are primarily descriptions of continuous quantities, and this analysis requires semantics – hence the topic of this paper.

In proposing a semantics for phrases in a natural language, it is often considered sufficient to justify the semantics by the degree of correspondence between the formal meaning and 'what we know the phrase means'. However, with descriptions of physical phenomena, there is an important additional ingredient, namely the physical quantities themselves, which we may measure and analyse to set up a three-way comparison between natural language scientific descriptions, the semantic modelling, and the actual physical phenomena themselves.

This is a large and wide-ranging project covering aspects of computational semantics, natural language processing, information extraction and integration, mathematical modelling, scientific description, and measurements and analysis of both texts and natural phenomena. This paper serves as a preliminary report on the methods, implementation and results achieved so far. We set out the semantics and briefly consider a few models of physical phenomena and how data gathered from nature supports such modelling. We also discuss the role of semantic information gathered from the statistical analysis of texts in descriptive sciences. Finally, we describe techniques for information integration and extraction based on this semantics. The mathematical background (metric spaces), the experimental data and its analysis and much other supporting material is to be found on the project website [1]. We see the project as very much experimental, being a series of experiments in practical semantics and its applications.

As for other work in the area, there has been considerable work on the automated processing of text, handling colour [6] or shape terms, and work on database aspects of storing and retrieving information based on semantic ideas including those of colour and shape (see, for example, [15]). Metrics have been proposed in various approaches to semantics, including in the work on 'conceptual spaces' (see e.g. [9]), spatial relationships (see e.g. [21] and [7]) and programming language semantics. The relatively new field of morphometrics [17] covers the measurement and analysis of size and shape in nature. Automated analysis of botanical texts is of considerable current interest, including work by the some of the authors of this paper (see [14], [24], [25], and [26]).

2 Descriptive Phrases for Continuous Quantities and their Semantics

We begin by looking at some of the phrase formations in English for describing continuous quantities, and consider semantic aspects of these phrases. We define the semantics in a metric space (for the mathematical background see standard texts, for example [11], and also [1]). Metric spaces appear to provide workable models of sufficient generality, and the notions of proximity and geodesics (shortest paths) that they provide correspond well with the semantics. At the end of this section, we discuss how to choose appropriate metric models for the semantics.

2.1 Simple Words

Consider simple words describing particular values of continuous quantities, for example *blue* for colour, or *obovate* for leaf shape (see Figure 2 for some pictures and names of leaf shapes). What do these denote? We shall use the standard semantic brackets for the denotations: thus what kind of objects are $[\![blue]\!]$ and $[\![obovate]\!]$?

Points in a metric space are the location of exact values. For continuous quantities, as they occur in nature or as described in natural languages, collections of points rather than single points are the appropriate denotation for these words, allowing points close to and indistinguishable from each other to be collected together in the denotation. Thus words describing continuous quantities are interpreted as *regions* of a metric space. The shape and extent of such regions clearly depends on quantity being modelled and the modelling space as well as the exactness of the descriptive word. Some formal terminological schemes for continuous quantities, for example for colour descriptions [3], [4], [12] and for leaf shapes [22], not only allocate points to names, but also delimit appropriate regions of the space. For other quantities, regions are determined by the way conventional meaning is represented in the model.

2.2 Modifiers

Consider phrases such as *pale yellow* or *deep reddish purple*, or, for shape, *broadly elliptic* or *narrowly linear*. Each of these uses a *modifier* to change the denotation of the main phrase. Thus $[\![pale]\!]$ and $[\![broadly]\!]$ are operators or transformations on the space of colours and leaf shapes, respectively.

Modifiers are interpreted as *continuous transformations* of a metric space into itself. The continuity here reflects the fact that physical indistinguishability is a coarser relation than limiting proximity in metric models. Other preservation properties (e.g. isometry) vary with modifier and model.

2.3 Intermediates and Ranges

Consider phrases such as, for colour, *pink to purple* or, for leaf shape, *ovate to elliptic*.

What are the semantic objects: $[\![\textit{pink to purple}]\!]$ or $[\![\textit{ovate to elliptic}]\!]$? Clearly, they are a delimited range of values between the end-points. But which values are in the ranges given? How red does pink have to be before it steps outside the range of colours in *pink to purple*? Is *obovate* in the range *ovate to elliptic*?

Paths in a metric space are continuous maps from a closed interval of reals to the space. We interpret ranges in terms of *shortest paths* (also known as *geodesics*). We consider only spaces where geodesics always exist between pairs of points. However, there may be more than one geodesic between two points (and sometimes an infinity).

Interpreting $[\![A]\!]$ and $[\![B]\!]$ as regions of the space, $[\![A \textit{ to } B]\!]$ is the set of points determined by the geodesics between points in $[\![A]\!]$ and points in $[\![B]\!]$. What does this set look like? This depends on the modelling space and in many cases it will form a region of the space. However, there is a subtlety in this use of geodesics, namely that given two geodesics between a pair of points, they may be continuously deformable into each other, or they may not, meaning that the geodesics take distinct routes. Ambiguity in the phrase is indicated when the interpretation splits into a number of distinct geodesic routes.

A related form of phrase commonly used is *violet-blue* or, for shape, *ovate-lanceolate*. Like single words, these compounds denote a particular region of the space, rather than paths in the space. The points in $[\![A\text{-}B]\!]$ are determined as particular points along the geodesic paths from $[\![A]\!]$ to $[\![B]\!]$. Exactly which points depends upon the model: half-way along is appropriate for some models but not others. Indeed, some formal terminological schemes for continuous quantities (including for colour and for leaf shape) specify the location of these intermediate values.

2.4 Mixed Forms and Logical Operators

Another form of expression for continuous quantities is common: those which mix several types of quantity. Examples of this are (combining colour and frequency) *violet-blue, rarely pink or white*, or (combining shape and frequency) *usually broadly ovate, occasionally deltoid*, or (combining size with shape - a hypothetical example) *small elliptic to large ovate*. Each of these combines two quantities, and an interpretation will consider pairs of semantic objects as though they were themselves semantic objects with appropriate notions of proximity, paths etc. That is, the interpretation takes place in a *product metric space*. Various products are available, reflecting the fact that metrics for the two quantities can be combined in a variety of ways.

Logical operators naturally occur in the description of continuous quantities and the complexity of their variation. For example, (for disjunction) *obovate to narrowly elliptic or narrowly obovate*, or *violet-blue, rarely pink or white*. Similarly, we meet conjunction and, in more complex expressions, negation. Interpreting values as regions of space, these operators combine regions of space. The standard boolean operations on sets appear to provide suitable interpretations, which we have incorporated into automated reasoning tools suitable for querying and for information integration.

2.5 Equality and Degrees of Semantic Equivalence

When looking at different (independent) botanical texts describing similar collections of species one is struck by the fact that rarely do they syntactically agree – rarely is the same phrase used to describe the same attribute of the same species in different texts.

The question is: Can we quantify the amount or range of agreement and disagreement in such descriptions? What is the relevant equality on semantic objects? For phrases describing continuous quantities, exact equality is likely to be too fine a distinction and a notion of proximity (closeness) is more appropriate.

An important application of this lies in combining multiple parallel descriptions into an integrated form. This is the process of data integration and clearly requires these semantic considerations in order to capture the overall information without redundancy and detecting any inconsistency that is present.

To compute the degree of equivalence and proximity of regions of the modelling space, we need suitable representation of regions and calculations of distances between such regions. We have built computational representations of regions as delimited ranges of parameters in multidimensional spaces. For some modelling spaces this is adequate, others require more complex representations, depending upon both the space and its metric.

2.6 Appropriate Models

With these interpretations in place, the key to providing a semantics for a continuous quantity lies in choosing an appropriate model as a metric space. The interpretations above provide criteria for the notion of 'appropriateness' of a model: the extent to which the interpretations match the intended semantics being the determinant of the suitability of the model. For some quantities (e.g. colour), there has already been work on determining appropriate metric models (which we discuss later). In general, the ingredients which contribute to successful models and the criteria of success are still under development.

3 Models

We briefly survey several models of natural phenomena to illustrate how the above semantics finds interpretation and we discuss the results of experiments in the measurement and analysis of these phenomena and how these results allow us to evaluate models.

3.1 Shape

Describing shape in natural languages is not easy. However, for botanists, the shape of plant parts is a key factor in the description and determination of taxa. Botanists have therefore devised a method and terminology for describing the shape of plant parts [16]. As a simple exercise in modelling plant shapes, we consider leaf shape.

There are numerous possibilities for modelling shape, including, interpolation techniques and curve-fitting, formulae for generating shapes, and transformational methods and those of morphometrics.

We consider a simple example of modelling, using polynomial curves. So that we can illustrate how the semantics of shape can be defined, we keep the modelling simple, restricting to leaves that are bilaterally symmetric about the main vein, and entire (without teeth or lobes), and consider only the overall leaf shape. See Figure 1 for a few examples of standard leaf shapes generated by a different choices of coefficients for a fixed polynomial.

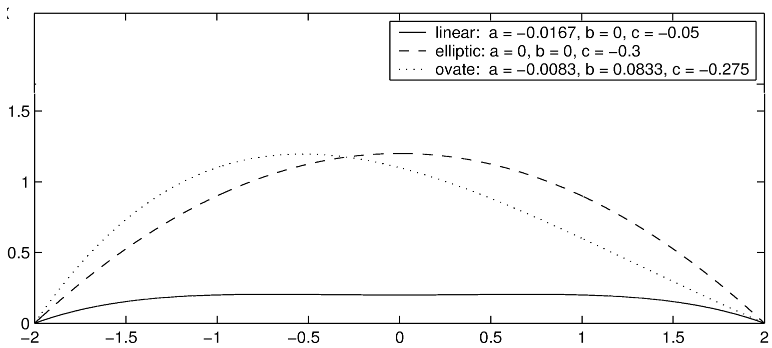

Fig. 1. Examples of leaf outlines generated by polynomials $(x+2)(x-2)(ax^2+bx+c)$.

What metric are we to use on these simple shapes? The choice of metric determines the semantics of expressions for leaf shape, so it is worth exploring the different semantics that metrics produce, and then compare these to botanical texts describing real leaves.

Let us begin with the Euclidean metric on the ordinates. The geodesics for this metric are, of course, straight lines. This means that there is only one path from one shape to another and that it progresses by making proportionally equal steps as the outline progresses from one shape to another. Thus, in this metric, the shape *lanceolate* is not in the range *linear to elliptic*, as *lanceolate* requires us to inflate in some regions more than others in order to achieve the asymmetry and then to complete the transition by a compensatory increase elsewhere. Note the importance of symmetry and asymmetry in our assessment of valid paths. Indeed, the Euclidean metric preserves symmetry (and the degree of asymmetry) in its geodesic paths.

Now let us consider the city-block metric: The notion of proximity is the same as for the Euclidean metric. However, geodesics differ for these two metrics. For the city-block metric, there are multiple geodesic paths which are generated by monotonic movements at each point. Thus values of the ordinates can move at different rates between their end-points, and in this metric *lanceolate* is in the range *linear to elliptic*.

Regions for interpreting shape names are determined by standard definitions [22]. Modifiers for leaf shape descriptions, such as *broadly* and *narrowly* are interpreted in this model simply as scaling by constants: a multiplier of $3/2$ suffices for *broadly*, and $1/2$ for *narrowly*.

Clearly, this model allows simple semantics for a restricted range of leaf shapes. There are many other approaches to modelling leaf shape [17]. One that we have investigated in detail is based on a formula [10] for generating shapes. A sample of such generated shapes approximating leaf forms is depicted in Figure 2. We have defined suitable metrics on the parameters generating these shapes and shown how we can then interpret leaf shape expressions using this formula. This provides a model with computational properties quite different from those of the polynomial model.

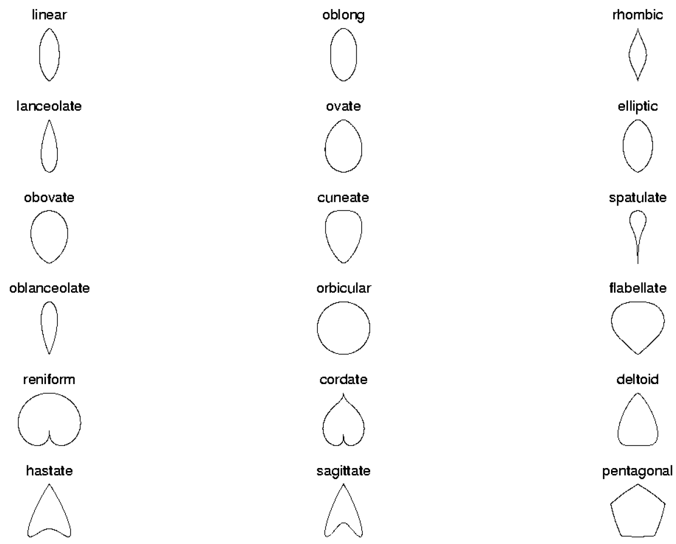

Fig. 2. Some leaf shapes generated by the formula of [10].

3.2 Analysis of Natural Phenomena: Leaf Shape

As mentioned above, a key aspect of this approach to semantics is that we can evaluate it against the phenomena under description, in this case, leaf shape.

We have undertaken a range of experiments (see [1] for data, results and analysis) in sampling and measuring leaves of several species, statistically analysing the results, and comparing this with models of leaf shapes, such as those above, and with botanical descriptions of the species.

As an example of the conclusions we can draw from such experiments, it appears that of the two metrics discussed above for the polynomial modelling of leaf shape, the Euclidean rather than the city-block better captures the interpretation of the botanical description of variation in nature.

What has become abundantly clear during these experiments is the real variability of nature and the rather loose coupling of this with botanical description. However, our experiments suggest an intriguing possibility, namely using the real measurements and the accompanying models to generate precise descriptions of the quantities and their variation as they occur in nature. This goes beyond the integration of descriptions and would bring to the fore the 'descriptive' part of 'descriptive science'. It would be interesting to see the phraseology needed to capture fully nature's variation.

3.3 Colour

As well as leaf shape, we have also investigated the use of colour in botanical texts, showing how some standard colour models suffice for the interpretation of colour expressions. Again, we have evaluated these interpretations against actual colour in plants.

The direct modelling of colour is via light spectra, however for computational purposes we choose one of the standard three-parameter models [8]. Those which are closer to human visual perception such as the Hue-Saturation-Lightness (HSL) model, or the Hue-Chroma-Lightness model [20], are clearly more appropriate for semantic modelling.

We have investigated the HSL model as a basis for the semantics. The requisite metric is Euclidean distance and geodesics are straight lines. When depicted, these paths do appear to be in approximate accord with our expectations of colours in colour ranges.

To interpret colour phrases, standard interpretations for basic colour terms *red*, *blue* etc. and for colour modifiers *pale*, *deep* etc. are used (see, for example, [3], [4], [5], [12]) and, for combined expressions, the metric semantics presented here.

To compare the semantics with actual colours in nature we use digital images. Because of the extensive processing of pixel information in these images, the resultant colour information is only an approximation to the original, yet even this simple analysis has provided useful results (see [1]). The overall conclusions of this analysis are similar to those of leaf shape, allowing us to evaluate models, but also exposing how botanists differ in their descriptions of colour for a species, and that nature's variation is complex and no simple descriptions suffice in general, though this analysis suggests methods of refining descriptions to better match nature. Semantic information integration, which we discuss below, is clearly one aspect of this multiplicity and variability.

4 Statistical Analysis of Texts

The statistical analysis of large corpora of texts can reveal much about how we use words, and how meanings and proximity of meaning are reflected by contexts. Botanical texts are highly specialised in several ways and it is interesting to see how general textual analysis copes with the highly structured nature of these texts, the presence of multiple parallel descriptions and the specialised limited vocabulary. With botanical texts there is a further complication, namely that frequency of occurrence is mediated not only by meaning but also by frequency in nature. A naive reading of contextual data may therefore not yield the correct metric information.

As source texts, we have used the online floras available in eFloras. We have written a small automated parsing and text extraction system to identify parts of plant descriptions and contexts for words. With this in place, we have run analyses of shape terms, using the methods of Latent Semantic Analysis [13], in particular, using the singular value decomposition of matrices of words against contexts followed by 'dimension reduction'. The resultant matrices indicate transitive notions of occurrence across multiple contexts. We have used these techniques in two ways: Taking into account the parallel nature of the descriptions and ignoring it. Both analyses appear to give similar results. For example, using just 9 words for leaf shape, the similarity orders in the table below resulted. Each line here is a list of words in increasing contextual distance from the first word.

linear: lanceolate ovate oblong elliptic obovate
oblong: ovate lanceolate elliptic linear obovate
rhombic: ovate linear lanceolate obovate oblong
lanceolate: ovate oblong linear elliptic obovate
ovate: lanceolate oblong elliptic linear obovate
elliptic: ovate oblong lanceolate obovate linear
obovate: ovate oblong lanceolate elliptic linear

These results are in good accord with the models of leaf shape described above. Moreover, this analysis also provides a means of building metric models (see [13]) which themselves may be compared with existing models or used directly to produce a semantics.

5 Applications: Information Retrieval and Integration

Semantics has a fundamental role in the automated processing of text, in information extraction, in database ontologies and in description logics for database interrogation, where the logical aspects of querying are combined with models of the relevant structures, including models of continuous quantities as we present in this paper. Moreover, a semantics is necessary for proper data integration:

combining parallel descriptions into single descriptions without redundancy, detecting inconsistency, and with due regard for the range of variation amongst descriptions. A metric semantics allows us to identify and compare regions of space so that consistency can be described in terms of proximity and integration is the assembling of descriptions of compound regions. When incorporated into ontologies and description logics, such semantic tools can provide powerful mechanisms for the organisation, processing and exploration of knowledge bases. For this to be effective, the semantics needs to be in a suitable computational form and the incorporation of this into a description logic should extend the automated reasoning ability without a significant deterioration in its performance.

An ontology-based approach to extracting and integrating parallel information from multiple (botanical) textual sources is described in [14], [24], [25] and [26]. This work however does not support a semantics of the descriptive phrases which form a key component of such texts. In more recent work, Wang and Pan [23] have incorporated into this ontology the semantic modelling described in this present paper and shown how it may be combined with a description logic reasoner (FaCT-DG in OWL see [2] and [18]).

The process of incorporating semantic information into ontologies and description logics begins by representing phrases describing continuous quantities as a datatype in the description language. A subsumption relation is defined over these phrases so that more general expressions subsume more specialised expressions. Based on this subsumption relation, the description logic reasoner may then process queries of the following forms (1) exact matching, (2) subsumption, (3) specialisation (converse of subsumption), and (4) overlap (non-empty intersection).

To handle integration of descriptions, explicit calculation of regions in the modelling space is required using the semantics of this paper. Then the metric allows us to assess the similarity of two descriptions and decide how a combined description should be formulated. This is based on an iterative process described in [23] which uses simplified calculations of regions. The result of this process is not only an integrated result, but also a record of how the individual descriptions relate to the integrated result and to each other.

This scheme has been implemented for colour descriptions based on the HSL model. The results of a series of experiments on botanical texts, both inferential matching and description integration, have been encouraging. We intend to use similar techniques to incorporate the semantics of shapes as described in this paper, and more generally to extend the techniques to incorporate the semantics of other continuous quantities.

6 Discussion

The purposes of this paper have been broad indeed – not only to describe a general semantic framework for continuous quantities in natural languages but also to present a range of models for several such quantities and to show how models may be assessed against both the physical phenomena being described and semantic information arising from textual analysis. Because of this range of issues, we have been able to introduce only briefly some of these aspects and clearly much work remains to be done.

Models arise from a wide variety of sources: (1) standard mathematical constructions of metric spaces, (2) physical phenomena (e.g. light spectra), (3) technology of devices (e.g. three-parameter models of colour), (4) models of perception (e.g. other three-parameter models of colour), and (5) textual analysis of the use and context of words. Some models have a degree of appropriateness for their role in semantics built into their construction. It is perhaps unsurprising that these models are adequate for the task and that proximity and geodesic paths really do reflect a semantic reality which we can compare with the physical phenomena under discussion. As we have indicated in the paper, this wealth of models and their suitability for semantics is worthy of considerably more investigation.

Another aspect of models under investigation is their role in computation, especially in text processing and in information structuring and retrieval. This provides further requirements on models in terms of their suitability. Even the small range of models we have described exhibit a variety of computational behaviours. One can enumerate such computational aspects of models: for example the efficiency of computing semantic values, the computation of proximity, regions and geodesic paths, and the ease or difficulty of incorporating these computations into ontologies and into automated reasoning systems.

At the moment, we have investigated only two continuous quantities. Obviously, there are many other quantities present in botanical texts (and elsewhere). Many appear to fit fairly straightforwardly into this framework. However, interesting issues arise in modelling some kinds of quantities, for example temporal aspects of plant development and spatial aspects of the arrangement of plant parts. It is clear that this is just the beginning of a far more extensive investigation of the semantics of continuous quantities, leading us to consider other continuous quantities, applications other than botany, other models and more data from nature and from texts.

References

1. Manchester Computational Botany Website: http://www.cs.manchester.ac.uk/flora
2. S. Bechhofer, F. van Harmelen, J. Henderson, I. Horrocks, D.L. McGuiness, P.F. Patel-Schneider, and eds., L.A.S. (2004) OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref.
3. T. Berk, L. Brownston, and A. Kaufman. A new color-naming scheme for graphics languages. *IEEE CG&A*. Vol 2. No. 3. pp 37-44. IEEE 1982.
4. T. Berk, L. Brownston, and A. Kaufman. A human factors study of color notation systems for computer graphics. *Comm. ACM.* Vol 25. No. 8. pp 547-550. ACM, August 1982.
5. D. Conway. An experimental comparison of three natural language colour naming models. *Proc. East-West International Conference on Human-Computer Interaction*. St. Petersburg, Russia. pp 328-339. 1992.

6. Mike Dowman. A Bayesian Approach To Colour Term Semantics. Lingscene, Volume 1. 2001.
7. Max J. Egenhofer and A. Rashid B.M. Shariff. Metric Details for Natural-Language Spatial Relations. *ACM Trans. Inf. Syst.* 16(4) 295-321. 1998.
8. James D. Foley, Andries van Dam, Steven K. Feiner and John F. Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley, 1996.
9. Peter Gärdenfors. *Conceptual Spaces.* MIT Press. Cambridge, MA. 2000.
10. Johan Gielis. A Generic Transformation that Unifies a Wide Range of Natural and Abstract Shapes. *American Journal of Botany*, 90(3), 333-338. 2003.
11. J.R. Giles. *Introduction to the Analysis of Metric Spaces.* Australian Mathematical Society, Lecture Series 3. Cambridge University Press.1987.
12. J.M. Lammens. *A computational model of color perception and color naming*. PhD Thesis. State University of New York at Buffalo. 1994.
13. T.K. Landauer, P.W. Foltz and D. Laham. Introduction to Latent Semantic Analysis. *Discourse Processes* Vol 25. pp 259-284. 1998. See also the website: http://lsa.colorado.edu.
14. Susannah J. Lydon, Mary McGee Wood, Robert Huxley and David Sutton. Data patterns in multiple botanical descriptions: implications for automated processing of legacy data. *Systematics and Biodiversity* 1(2) 151-157. 2003.
15. S. Marchand-Maillet et al. Viper's CBIRS webpage. Online address: http://viper.unige.ch/other_systems
16. Jimmy R. Massey and James C. Murphy. Vascular Plant Systematics: Categorized Glossary. 1998. Online address: http://www.ibiblio.org/botnet/glossary
17. A Morphometrics website. http://www.mat.univie.as.at/ neum/morph.html.
18. J.Z. Pan. *Description Logics: Reasoning Support for the Semantic Web.* PhD Thesis, School of Computer Science, The University of Manchester.
19. P. Pantel and D. Lin. Discovering Word Senses from Text. *Conference on Knowledge Discovery In Data.* Proc. 8th ACM SIGKDD Conference. pp 613-619. 2002.
20. M. Sarifuddin and R. Missaoui. A New Perceptually Uniform Colour Space and Associated Similarity Measure for Content-Based Image and Video Retrieval. Multimedia Information Retrieval Workshop. Salvador, Brazil. 2005.
21. Schwering, A. and Raubal, M. Spatial Relationships for Semantic Similarity Measurement. 2nd Intn. Workshop on Conceptual Modeling for GIS. Klagenfurt, Austria. 2005
22. Clive Stace. *New Flora of the British Isles.* Cambridge University Press. 1997.
23. Shenghui Wang and Jeff Z. Pan. Ontology-based Representation and Query of Colour Descriptions from Botanical Documents. in *Proc. of the 4th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE-2005).* 2005.
24. Wood, M., Lydon, S., Tablan, V., Maynard, D. and Cunningham, H. Populating a database from parallel texts using ontology-based information extraction. In Meziane, F. and Métais, E., (editors) *Proceedings of Natural Language Processing and Information Systems, 9th International Conference on Applications of Natural Languages to Information Systems.* 254-264. Springer 2004.
25. Wood, M., and Wang, S. Motivation for "ontology" in parallel-text information extraction. In *Proceedings of ECAI-2004 Workshop on Ontology Learning and Population. (ECAI-OLP).* Poster. Valencia, Spain. 2004.
26. Wood, M.M., Lydon, S.J., Tablan, V., Maynard, D., and Cunningham, H. Using parallel texts to improve recall in ie. In *Proc. of Recent Advances in Natural Language Processing (RANLP-2003).* 506-512. Borovetz, Bulgaria. 2003.