On the semantics of continuous quantities in natural languages: An overview

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

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 complexity and 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 and is the same over a wide range of such quantities. 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 investigate 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. These written descriptions have a degree of precision and detail not normally encountered in texts. Expressions for continuous quantities form a major component of these descriptions, for example, expressions for flower colour, leaf shape, leaf margins, petal texture, stem cross-sections, branching patterns, and also a wealth of temporal and spatial expressions. Because of the 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 computational tools

*School of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, U.K. emails: shenghui.wang@cs.man.ac.uk, david.rydeheard@manchester.ac.uk, m.m.wood@manchester.ac.uk and david.s.bree@manchester.ac.uk.

for information organisation and extraction. 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 the 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.

It is to the mathematics of metrics and topology that we turn to provide what appears to be a workable basis for the semantics of continuous quantities. We will show how to apply this semantics to colour descriptions and to simple shapes. Using botanical texts, we briefly illustrate how we may assess the semantics against nature and against textual sources of semantic information. We then describe applications of the modelling to text-based information retrieval and integration, and to the interrogation and organisation of databases. All of this is very much experimental and we see this paper as a series of experiments in practical semantics.

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 short preliminary report on the methods, implementation and results of the project.

As for other work in the area, there has been considerable interest in the way that spoken languages handle colour terms and the linguistic implications of this (see the bibliography [4] for details). There has also been work on the automated processing of text, handling colour [7] 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, [18]). Metrics have been proposed in various approaches to semantics, including in the work on 'conceptual spaces' (see e.g. [11], which uses some techniques similar to those in this paper but applied to the broader problem of the relationship between concepts and language), spatial relationships (see e.g. [24], [25] and [9]) and programming language semantics. The relatively new field of morphometrics [20] 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 [17], [30], [31], and [32]).

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 [14], [6], or the summary in the Appendix to this paper). 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 [2], [3], [15] and for leaf shapes [27], 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. This observation is not 'compositionality' of semantics but merely a statement that $[\![pale]\!]$ and $[\![broadly]\!]$ transform colours to colours and shapes to shapes.

We interpret modifiers 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 following semantic objects: $[\![pink\ to\ purple]\!]$ or $[\![ovate\ to\ elliptic]\!]$? 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, $[\![A\ to\ B]\!]$ 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

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.

2.5 Logical Operators

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.

2.6 Equality, Degrees of Semantic Equivalence, and Ambiguity

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 then is: Can we quantify the amount or range of agreement and disagreement? We would expect a semantics to be able to distinguish actual disagreement from different but equivalent descriptions. One question is: What is the relevant equality on semantic objects? However, 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 the semantic considerations just listed in order to capture the overall information without redundancy and detecting any inconsistency that is present.

The key to this is the representation of regions of the modelling space and determining distances between such regions. We have experimented with simple computational representations of regions as delimited ranges of parameters in multidimensional spaces. In general more complex representations are required, depending upon both the space and its metric.

Related to the question of equivalence of interpretations is the identification and resolution of ambiguity. Descriptions of continuous quantities are often complex because of the variation that we wish to describe. Thus even *pink to purple or blue* is open to two interpretations and more complex combinations of values, ranges and logical connectives are often a source of hidden ambiguity. Another aspect of ambiguity is that a numerical semantics of continuous quantities may be too fine-grained, that is, it may distinguish semantic objects that we wish to consider equivalent, and thus introduce a spurious ambiguity which is an artifice of the modelling technique. Since the intended interpretations are not always precise, it may not be clear whether a phrase is inherently ambiguous or whether it is an artifice of the modelling. Introducing a metric allows us to compare the granularity of the semantics with the intended meaning therefore to quantify the 'degree of ambiguity'.

2.7 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. This involves not simply devising a parametric model, but also defining a metric between points that reflects semantic proximity. 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.

Metric spaces may not be the only setting for the semantics of continuous quantities. However, we attempt to show that they provide workable models of sufficient generality, and that the notions of proximity and geodesic they provide correspond well with the semantics.

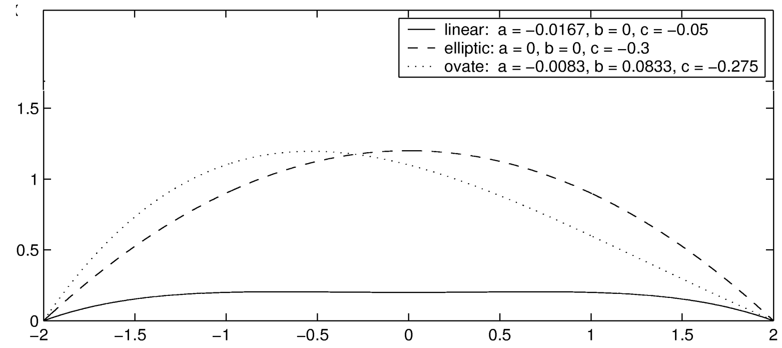

Figure 1: Examples of leaf outlines generated by polynomials $(x + 2)(x - 2)(ax^2 + bx + c)$.

3 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 [29], [19]. As a simple exercise in modelling plant shapes, we consider leaf shape.

There are numerous possibilities for modelling shape:

1. Interpolation techniques and curve-fitting. For example, polygonal fitting (i.e. selecting points on the margin); simple interpolation using, say, polynomials, through to sophisticated curve-fitting techniques, e.g. using (elliptic) Fourier series, or using splines,

2. Formulae for generating shapes,

3. Transformational methods e.g. conformal maps [26] or computer image morphing techniques, or compositional methods [8]. Included here are also so-called ‘landmark-based’ methods which identify key features of a shape and use the location and values of these to match and compare shapes.

3.1 Polynomial Curves

We begin with 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.

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. For polynomials $p(x)$ and $p'(x)$, the Euclidean metric is:

$$\mu(p, p') = \sqrt[2]{\int_{-2}^{2} (p(x) - p'(x))^2 \, \mathrm{d}x}.$$

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 for each value of $x \in [-2, 2]$.

Thus, in this metric, the answer to: ‘is *lanceolate* in the range *linear to elliptic*?’ is no, as *lanceolate* requires us to inflate in some regions more that the 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 x in $[-2, 2]$. Thus values of $p(x)$ can move at different rates between their end-points, and in this metric *lanceolate* is in the range *linear to elliptic*.

Note that in the above examples, proximity is a rather crude determiner of an appropriate metric, whereas the form of the geodesic paths is a more discriminating characteristic. We find this a general phenomenon in determining appropriate metrics for modelling continuous quantities.

Regions for interpreting shape names are determined by standard definitions [27]. 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. This may be expanded with higher-degree polynomials, but more complex leaf shapes are difficult to define. We now turn to a more expressive model, but one in which the semantics of shape expressions is more difficult to formulate.

3.2 A Formula for Generating Shapes

Following in a long tradition of devising formulae for generating shapes in nature, Gielis has recently proposed a formula (he calls it 'the superformula') which generates an impressive collection of shapes (see [12] and [13]). The formula is, in polar co-ordinates (r, θ):

$$r = \frac{1}{\sqrt[n_1]{(|\frac{1}{a}\cos(\frac{m\theta}{4})|)^{n_2} + (|\frac{1}{b}\sin(\frac{m\theta}{4})|)^{n_3}}} \qquad (1)$$

Though the formula is very expressive for its number of parameters, it is mathematically fairly intractible, in that many calculations using the formula, including curve-fitting problems, cannot in general be expressed in simple closed form.

In Figure 2, we depict leaf shapes that arise from the formula. Notice that with just 6 parameters we can generate many realistic leaf shapes, including fairly complex combinations of leaf base (at join with leaf stalk), overall leaf shape and leaf apex (or tip), for example *reniform* and *cordate*. This is much more expressive than simple polynomials or other general approximation methods with so few parameters. However, not all the representations of shape above are good matches to the actual meanings of the terms. For example, the entry for *linear* ought to be thinner and with more parallel sides. It general, the choice of parameters was guided by some general principles combined with a good deal of trail-and-error.

Parameters determining a shape are not unique: for example, if $n_2 = 0$ then any value of a will suffice, likewise n_3 and b. This means that pseudometrics, rather than metrics, are required.

Let us now turn to various phrasal forms for shape and their expression through this formula. For modifiers, such as *broadly* and *narrowly* we have defined transformations of the parameters which are suitable for a range of shapes, but no simple transformation appears to be a fully general interpretation of the modifiers.

For further phrasal forms, we need to determine an appropriate (pseudo)metric on the parameter space. Clearly, standard metrics on $\mathbf{R}^6$ are not suitable as the parameters have very different roles in the formula. Nevertheless, we need to define pseudometrics which reflect proximity and distance of shapes. To do so we consider a series of attributes of shapes each of which is important in determining proximity. An example of such a series is:

> (1) length to width ratio, (2) position along length of widest part, apex angle, and (4) angle at base with petiole (leaf stalk).

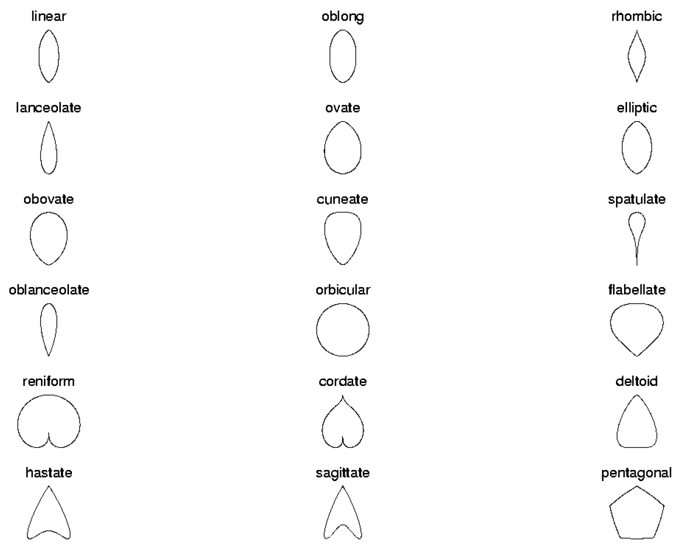

Figure 2: Some generated leaf shapes.

These attributes may be combined into a pseudometric in various ways, the most obvious being a (weighted or unweighted) Euclidean pseudometric.

We have extended the representation to provide a more flexible approach to shape description, in which we combine the use of the formula with transformations of the plane. We may define metrics, using the shape-attributes above, on these pairs. This representation provides a suitable basis for a semantics of leaf shape expressions for a useful variety of shapes.

In this world of leaf shapes, there is clearly much more work to do in developing models. There are many forms of leaf which we haven't deal with. Even the move to, say, lobed leaves shows that new definitions of metrics between shapes are necessary, not simply to describe more complex shapes but also to allow more complex variation of shape. Some of the techniques of morphometric modelling appear to be particularly relevant here for developing appropriate spaces and metrics.

3.3 Analysis of Real Leaves

A key aspect of this semantic framework is that we may assess it against natural phenomena. We briefly indicate here how this may be undertaken and the type of results that arise. We have collected a variety of plant samples, extracted data from these samples and undertaken statistical analysis to perform a three-way comparison: comparing nature with it description in botanical texts and with the semantic modelling.

We have space to describe only one such set of measurements: the leaf shapes of *Fagus sylvatica* (Beech), a European deciduous tree. Leaf shapes are described by different authors as: *oval or obovate; oval to elliptical; elliptical; oval; ovate to elliptic.*

In Figure 3, for a sample of leaves, we plot the ratio of quarter-way width of each leaf to its length (marked with $+$) and the ratio of the three-quarter-way width to its length (marked with $*$) (both along the y-axis) against the ratio of the half-way width to the length (along the x-axis).

We see here part of the variation of leaf shape for this species. Notice that there is considerable variation along the x-axis, along the y-axis and also between the two data points for each leaf. Some leaves are *elliptic*, some *elliptic-ovate* or *ovate*, whilst others are *elliptic-obovate* or *obovate*, with a fairly even spread amongst these. For both data sets illustrated, the x-y dependency approximates to a straight line, which is the geodesic for the Euclidean metric in the first of the models above, saying that the proportionate increase in each of the ratios is constant when moving between

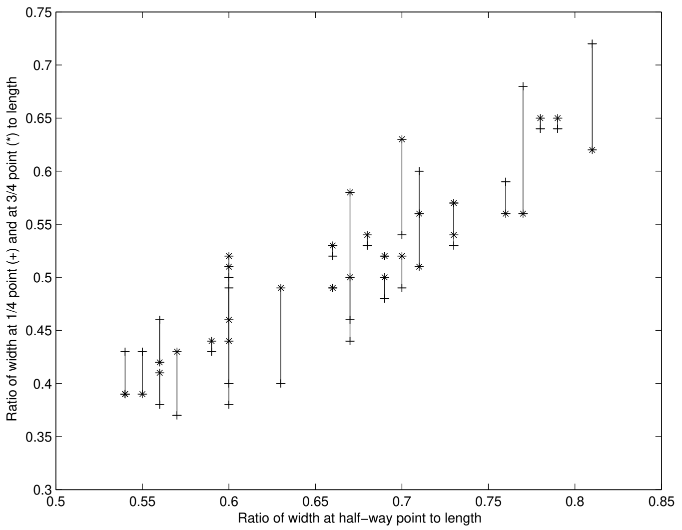

Figure 3: Ratios of leaf dimensions for a sample of *Fagus sylvatica* (Beech).

shapes. How good is this approximation for this sample? The best least-squares fit using a linear regression model for the quarter-way ratio is a line with gradient 0.95 and 95% confidence limits at approx. 0.09 above and below this line with a fairly uniform distribution of data points around the line.

This is but one set of measurements in a more extensive investigation. Even so, the results provide some support for the polynomial models of leaf shapes using the Euclidean metric. By extending this analysis to a wider variety of leaf shapes and measuring the attributes used in the metric, we may provide a similar assessment of the model based on Gielis' formula, and other modelling techniques too.

What is also clear from this preliminary analysis is the lack of unanimity in the textual descriptions, even for this simple leaf shape, and the corresponding variation in nature. This is a point we return to later when discussing information integration and descriptions of complex variation.

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 (www.eFloras.org). 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 [16], 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 both taking into account the parallel nature of the descriptions and ignoring it. Both analyses given 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
spatulate: oblong ovate linear lanceolate obovate
oblanceolate: oblong lanceolate linear ovate elliptic

We can read a good deal from these results, for instance, the proximity of narrow shapes, and of intermediate (*elliptic*-like) shapes, and the isolation of shapes such as *rhombic* and *spatulate*.

Such analysis provides a means of assessing models to see to what extent they are in accord with this contextual notion of semantics. It also provides a means of building metric models (see [16]) which themselves may be compared with other models.

5 Colour

We now turn briefly to the semantics of expressions describing colour. The direct modelling of colour is via light spectra, however for computational purposes we choose one of the standard three-parameter models [10], [23].

The simplest perhaps is the Red-Green-Blue model of additive colours. A model which appears to be closer to human perception is the Hue-Saturation-Lightness (HSL) model. This classifies colours not by contributory colour components, but by three aspects which are meant to represent how we analyse and match colours. The space consists of two polar co-ordinates (s, h) with the hue h representing the angle around the 'colour-wheel', and the saturation s the distance from the centre of the wheel. Lightness is orthogonal to this and makes the colour-wheel into a double cone with apices being black (lightness is 0) and white (lightness is 1). All the three-parameter models are equivalent in the sense that there are bijective conversions between them.

The appropriate metric for the HSL space is the polar expression for Euclidean distance:

$$\mu((h, s, l), (h', s', l')) = \sqrt{(l - l')^2 + (s^2 + s'^2 - 2ss' \cos(h - h'))}.$$

The geodesics of this metric are straight lines, and when depicted do appear to be in approximate accord with our expectations of colours in colour ranges.

We now explore the possibility of using this model to give a semantics of colour expressions as they occur in botany. Colour phrases are interpreted using standard interpretations for basic colour terms *red*, *blue* etc. and for colour modifiers *pale*, *deep* etc. (see, for example, [2], [3], [5], [15]) 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 provides some useful results.

To illustrate this, consider an example of a plant whose flowers exhibit a range of colours, *Pulmonaria longifolia* (Narrow-leaved Lungwort). This is described by various authors as: *red at first, soon turning to violet or violet-blue; violet to violet-blue; blue to violet when open*. Notice that the color range is here mixed with flower maturity in a product metric. In Figure 4, we depict the flower colour in HSL for flowers of differing ages. The first plot is the the projection of the double cone onto its base, i.e. the variation of Hue with Saturation. The second plot is the projection of the HSL double cone on the red-cyan axis, with red on the right. The first plot shows that the hues are indeed in a range from fairly pure blue through to red (on the blue side of reds) and the second plot shows that the red pixels (on the right) are predominantly in the darkish red area (bottom right), whilst the blue pixels (on the left) range more widely, especially

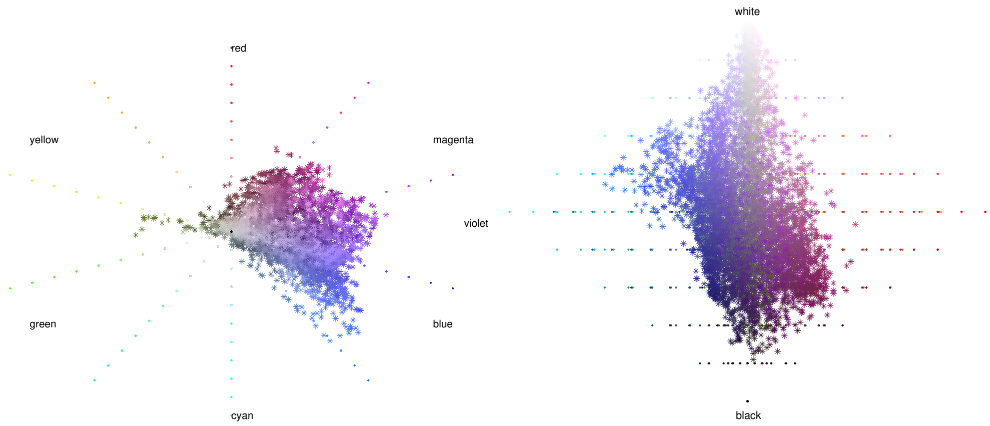

Figure 4: HSL plots of the flower colour pixels for *Pulmonaria longifolia*.

in the lighter blues where they extend out to a saturation of nearly 60%. Further analysis shows that pixel frequency peaks sharply in the blue area and again in the purplish red area with few intermediate points. As expected, the botanical descriptions concentrate on hue and less on the variation in saturation and lightness which is evident here.

As with leaf shapes, this is but a small example of the analysis of modelling colour in nature and its description in botanical texts, but the conclusions we draw about the variation in textual description and that in nature and the way the semantic modelling represents this are similar.

6 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.

With such a semantics, we may address questions such as: how do we detect inconsistency in knowledge bases, especially, in the applications to botany, inconsistency in multiple parallel descriptions? Moreover, a semantics is necessary for proper data integration: combining parallel descriptions into single descriptions without redundancy 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 and storing it in a suitable electronic form is described in [17], [30], [31] and [32]. 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 [28] 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 [1] and [21]).

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 then 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 [28] 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.

7 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 investigate 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. For the small range of models we have described, a variety of computational behaviours have been displayed. One can enumerate such computational aspects of models: for example the efficiency of computing semantic values, the computation of proximity and geodesic paths, and the ease or difficulty of incorporating these computations into ontologies and into automated reasoning systems.

What has become very clear during this investigation is the real variability of nature and the rather loose coupling of this with botanical description. Botanical descriptions are as precise as is needed for the purpose in hand. 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. In this way, the 'descriptive' part of 'descriptive science' would come to the fore and it would be interesting to see the phraseology needed to capture fully nature's variation.

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 the modelling some kinds of quantities, for example temporal aspects of plant development and spatial aspects of the arrangment 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] 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.

[2] 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.

[3] 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.

[4] Alex Byrne and David Hilbert. A Bibliography of Color and Philosophy. *Readings on Color, Vol. 1: The Philosophy of Color*. MIT Press. 1997 (See also: http://web.mit.edu/philos/www/color-biblio.html)

[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] E.T. Copson. *Metric Spaces*. Cambridge Tracts in Mathematics 57, Cambridge University Press. 1968.

[7] Mike Dowman. A Bayesian Approach To Colour Term Semantics. Lingscene, Volume 1. 2001.

[8] Shimon Edelman. Representation, Similarity, and the Chorus of Prototypes. *Minds and Machines* 5:45-68. 1995.

[9] 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.

[10] James D. Foley, Andries van Dam, Steven K. Feiner and John F. Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley, 1996.

[11] Peter Gärdenfors. *Conceptual Spaces*. MIT Press. Cambridge, MA. 2000.

[12] Johan Gielis. A Generic Transformation that Unifies a Wide Range of Natural and Abstract Shapes. *American Journal of Botany*, 90(3), 333-338. 2003.

[13] Johan Gielis and Tom Gerats. A botanical perspective on modeling plants and plant shapes in computer graphics. Radboud University, Nijmegen, Netherlands. http://www.pg.science.ru.nl/en/plantmodeling.html.

[14] J.R. Giles. *Introduction to the Analysis of Metric Spaces*. Australian Mathematical Society, Lecture Series 3. Cambridge University Press.1987.

[15] J.M. Lammens. *A computational model of color perception and color naming*. PhD Thesis. State University of New York at Buffalo. 1994.

[16] 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.

[17] 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.

[18] S. Marchand-Maillet et al. Viper's CBIRS webpage. Online address: http://viper.unige.ch/other_systems

[19] Jimmy R. Massey and James C. Murphy. Vascular Plant Systematics: Categorized Glossary. 1998. Online address: http://www.ibiblio.org/botnet/glossary

[20] A Morphometrics resources website. http://www.mat.univie.as.at/ neum/morph.html.

[21] J.Z. Pan. *Description Logics: Reasoning Support for the Semantic Web.* PhD Thesis, School of Computer Science, The University of Manchester.

[22] 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.

[23] 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.

[24] Schwering, Angela. Hybrid Model for Semantic Similarity Measurement. University of M unster. http://ifgi.uni-muenster.de/ eidueidu.

[25] Schwering, A. and Raubal, M. Spatial Relationships for Semantic Similarity Measurement. 2nd Intn. Workshop on Conceptual Modeling for GIS. Klagenfurt, Austria. 2005

[26] E. Sharon and D. Mumford. 2D-Shape Analysis using Conformal Mapping. CVPR (2) 2004: 350-357.

[27] Clive Stace. *New Flora of the British Isles.* Cambridge University Press. 1997.

[28] 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.

[29] William T. Stearn *Botanical Latin.* Timber Press, 4th edition, 2004.

[30] 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.

[31] 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.

[32] 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.

Appendix: Metric Spaces

We describe some of the mathematical concepts relating to metric spaces. This is a brief overview of those topics which appear in this paper.

Definition 1. A *metric space* $(\mathbb{S}, \mu)$ is a set $\mathbb{S}$ together with a binary operation $\mu : \mathbb{S} \times \mathbb{S} \to \mathbf{R}^+$, where $\mathbf{R}^+$ is the set of non-negative real numbers, such that, for all x, y, z in $\mathbb{S}$,

1. $\mu(x, y) = 0 \iff x = y$,
2. $\mu(x, y) = \mu(y, x)$,
3. $\mu(x, z) \leq \mu(x, y) + \mu(y, z)$.

Thus μ is symmetric and satisfies the triangle inequality (the third axiom). If the weaker form of the first axiom

$$\forall x \in \mathbb{S}. \quad \mu(x, x) = 0$$

holds instead then we call μ a pseudometric instead of a metric.

Example 1. Consider the set $\mathbb{S} = \mathbf{R} \times \mathbf{R}$, where $\mathbf{R}$ is the set of real numbers, and the series μ_i for $i \geq 1$:

$$\mu_i((x_1, y_1), (x_2, y_2)) = \sqrt[i]{(|x_1 - x_2|)^i + (|y_1 - y_2|)^i}.$$

Each μ_i is a metric on $\mathbf{R} \times \mathbf{R}$.

For $i = 1$ the metric becomes

$$\mu_1((x_1, y_1), (x_2, y_2)) = |x_1 - x_2| + |y_1 - y_2|$$

and is sometimes called the *city-block* metric.

For $i = 2$, we have the familiar Euclidean metric:

$$\mu_2((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

For an example of these two metrics, see Figure 5.

Example 2. Consider bounded functions integrable over an interval $[a, b]$. We may define several simple metrics on the space of these functions:

1. The series of metrics μ_i, $i \geq 0$:

$$\mu_i(f, g) = \sqrt[i]{\int_b^a (|f(x) - g(x)|)^i \, \mathrm{d}x}.$$

2. The *maximum distance metric*, defined as:

$$\mu(f, g) = sup(\{|f(x) - g(x)| \ : \ x \in [a, b]\}).$$

The first are the functional analogues of the metrics in Example 1. The second is the limit of the first as i tends to infinity.

Functions into a pseudometric space induce a pseudometric: Given pseudometric space $(\mathbb{S}, \mu)$ and arbitrary function $f : \mathbb{S}' \to \mathbb{S}$, we define a pseudometric on $\mathbb{S}'$ as $\mu'(x, y) = \mu(f(x), f(y))$. For μ a metric and f injective, then μ' is a metric. A common example of this is that any function $f : \mathbb{X} \to \mathbf{R}$ induces a pseudometric μ on $\mathbb{X}$ by $\mu(x, y) = |f(x) - f(y)|$.

Each metric space determines a topological space:

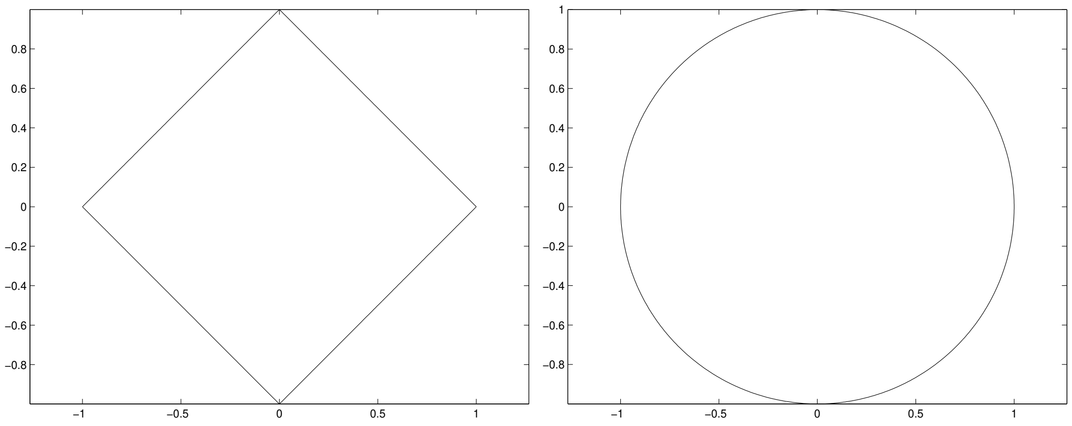

Figure 5: $\{(x, y) : \mu((x, y), (0, 0)) = 1\}$ for (left) the city-block metric and (right) Euclidean metric

Definition 2. For metric space $(\mathbb{S}, \mu)$, define the topological space $T(\mathbb{S}, \mu)$ to have the set of points $\mathbb{S}$. The open sets are generated from the following base:

$$\{\{y \in \mathbb{S} : \mu(x, y) < d\} : x \in \mathbb{S}, d \in \mathbf{R}^+\}$$

These sets are the open 'spheres', and the open sets $\mathcal{O}(\mathbb{S}, \mu)$ of the space are generated as arbitrary unions of these.

The same construction yields a topological space from a pseudometric. Spaces arising from metrics are Hausdorff. Notice that different metrics may yield the same topology.

There are several useful types of maps between metric spaces:

Definition 3. For metric spaces $(\mathbb{S}, \mu)$ and $(\mathbb{S}', \mu')$, a function $f : \mathbb{S} \to \mathbb{S}'$ is

1. *continuous* if the function induced on the topological spaces $f : T(\mathbb{S}, \mu) \to T(\mathbb{S}', \mu')$ is a continuous map,
2. an *isometry* if $\forall x, y \in \mathbb{S}.\ \mu'(f(x), f(y)) = \mu(x, y)$,
3. a *scaling* if there is a constant $k \in \mathbf{R}$ such that $\forall x, y \in \mathbb{S}.\ \mu'(f(x), f(y)) = k\mu(x, y)$. A bijective scaling is a *similarity*,
4. *partially contractive* if $\forall x, y \in \mathbb{S}.\ \mu'(f(x), f(y)) \leqslant \mu(x, y)$ (*contractive* if the inequality is strict).

We now consider the notion of paths and *shortest paths* (sometimes called *geodesics*) in metric space.

A *path* in a space $(\mathbb{S}, \mu)$ is a continuous function $u : [a, b] \to \mathbb{S}$ where $[a, b]$ $(a \leqslant b)$ is the closed interval of reals between a and b. Sometimes conditions stronger than continuity are imposed, but this suffices for the following definitions.

Shortest paths ('geodesics') are defined as paths parameterisable by arc-length:

Definition 4. A path $u : [a, b] \to \mathbb{S}$ (with $a \leqslant b$) in a metric space $(\mathbb{S}, \mu)$ is a *shortest path* or *geodesic*, if for all $t, t' \in [a, b]$,

$$\mu(u(t), u(t')) = |t - t'|.$$

This is a limited notion of geodesic applicable to metric spaces which have at least one such geodesic path between every pair of points. The models we consider are of this form. Geodesics may then be unique or there may be a multiplicity (sometimes an infinity) of geodesics between a pair of points.

To calculate geodesic paths, often the metric space is of such a form that the length of a path can be expressed as an integral and then the shortest path is the path (in a certain set of paths) that minimizes the integral, and this can be determined by variational methods in calculus.

Example 3. For the city-block metric in $\mathbf{R}^n$, geodesics are paths which monotonically advance towards the end-points. Thus, a continuous path $u : [0, 1] \to \mathbf{R}^n$ is a geodesic iff for all $s, t \in [0, 1]$, $s \leqslant t \implies u(s)_i \leqslant u(t)_i$ when $u(0)_i \leqslant u(1)_i$ and $u(s)_i \geqslant u(t)_i$ when $u(0)_i \geqslant u(1)_i$, where x_i denotes the i-th projection, $1 \leqslant i \leqslant n$.

For the Euclidean metric in $\mathbf{R}^n$, geodesics are straight lines.

We now turn to subsets of metric spaces:

Definition 5. A *region* of a metric space is a non-empty path-connected subset, i.e. a subset $S \subseteq \mathbb{S}$, such that all pairs of points in S are connected by a path.

Other conditions on 'regions' may be imposed, for example they may be closed and/or bounded. What is the distance between two regions? One possible definition of such a distance is:

Definition 6. The *Hausdorff distance* between two subsets X, Y of metric space $(\mathbb{S}, \mu)$, is given by the expression

$$\mu_H(X, Y) = \max\{\sup_{x \in X} \inf_{y \in Y} (\mu(x, y)), \sup_{y \in Y} \inf_{x \in X} (\mu(x, y))\}$$

and is defined when both sup's and inf's are defined.

On the set of closed, bounded subsets of $\mathbb{S}$, the Hausdorff distance is defined and is a metric.

Example 4. If X and Y are two spheres with radii r and s (with $r \geqslant s$), and distance between centres of d, then the $\mu_H(X, Y) = d + r - s$.

Products of metric spaces: Metric spaces, with isometries as morphisms, do not have cartesian products. However, given metric spaces $(\mathbb{S}, \mu)$ and $(\mathbb{S}', \mu')$, we can define a pseudometric on the product of the spaces $\mathbb{S} \times \mathbb{S}'$, if we are given a function $\nu : \mathbf{R}^+ \times \mathbf{R}^+ \to \mathbf{R}^+$, satisfying the following, for all $u, v, w, x \in \mathbf{R}^+$:

$$\nu(0, 0) = 0$$
$$\nu(u + v, w + x) \leqslant \nu(u, w) + \nu(v, x)$$
$$u \leqslant v \text{ and } w \leqslant x \implies \nu(u, w) \leqslant \nu(v, x).$$

The pseudometric $\overline{\mu}$ on $\mathbb{S} \times \mathbb{S}'$ is defined as

$$\overline{\mu}((x, x'), (y, y')) = \nu(\mu(x, y), \mu'(x', y')).$$

If further:

$$\nu(u, v) = 0 \implies u = v = 0$$

then this pseudometric is a metric.

The function ν in the above defines how the two metrics are to be composed. Examples of ν are

1. a linear combination: $\nu(u, v) = \lambda_1 u + \lambda_2 v$ with $\lambda_1, \lambda_2 \geqslant 0$,
2. maximum distance: $\nu(u, v) = \max(u, v)$, and
3. the Euclidean combination: $\nu(u, v) = \sqrt{u^2 + v^2}$.

Products have two roles: allowing us to combine metric spaces modelling different quantities, and also in the definition of metrics for a single quantity based on a combination of separate aspects of the quantity each of which is itself a metric.