

# HOW MANY COLORS SHOULD BE IN THE RAINBOW? REPRESENTING COLOR NAMES IN KNOWLEDGE BASES

Ronaldo Martins<sup>1</sup>

**Abstract** — This paper addresses the color categorization problem from the multicultural knowledge representation perspective. Given the fact that languages employ different color naming strategies, two questions are examined: 1) which colors should be represented in a multilingual and cross-cultural knowledge base? and 2) how to represent them to assure reciprocal translatability? Three different, although network-like structured, knowledge representation formalisms are considered as candidate alternatives: Conceptual Graphs (CG), Resource Description Framework (RDF) and Universal Networking Language (UNL). It is claimed that UNL, better than the others, can cope with multilingualism and cross-cultural color representation schemes.

**Index Terms** — Knowledge Representation, Reciprocal Knowledge, Universal and Local Knowledge, UNL System.

## INTRODUCTION

It is been widely observed that languages contain different numbers of color names and carve up the color spectrum in different ways [1,2]. In English, for instance, Sir Isaac Newton identified seven different colors in the rainbow: red, orange, yellow, green, blue, indigo and violet (or purple). In Shona, a language spoken by nearly 80% of people in Zimbabwe, there seems to be only three: [Cips<sup>w</sup>uka], which roughly corresponds to the English ‘purple’, ‘orange’ and ‘red’ (i.e., the borders of the color spectrum); [Citema], which covers part of both English ‘blue’ and ‘green’; and [Cicema], part of the ‘green’ and most of the ‘yellow’ [3]. For the Dani people of Papua New Guinea, only two basic color terms have been reported: [mili], for “cold, dark colors”; and [mola], for “warm, light colors” [4].

Such diversity has been often addressed as a proof for either cultural relativism [1,5] or biological determinism [3,4,6]. In the former case, language is believed to govern perception and determine world structure; in the latter, there would be color universals, provided that regularities in the pattern of color naming would have been noticed across different languages.

In what follows, the so-called lexical color categorization problem will be referred to in a rather different perspective: given the fact that there is no possible one-to-one mapping between color names of English, Shona and Dani, two questions must be addressed in the field of Knowledge Representation: 1) which colors should be

represented (i.e., named) in a multilingual and cross-cultural knowledge base?; and 2) how to represent them, to assure reciprocal translatability between different color naming strategies?

In order to answer these two questions, this paper will consider and compare three different knowledge representation formalisms: Sowa’s Conceptual Graphs (CG); W3C’s Resource Description Framework (RDF); and the Universal Networking Language (UNL). Although all of them are semantic networks, they are supposed to adopt different strategies for representing knowledge.

This paper is divided in two different parts: in the first, the color representation problem will be addressed inside each analyzed formalism: CG (Section 1), RDF (Section 2) and UNL (Section 3). The second part provides a very brief comparative analysis (Section 4) and draws a partial conclusion (Section 5).

## CONCEPTUAL GRAPHS

Conceptual graphs (CGs) [7,8] are said to be “a system of logic based on the existential graphs of Charles Sanders Peirce and the semantic networks of artificial intelligence”. They would express meaning in form that is “logically precise, humanly readable, and computationally tractable”. CGs have been implemented in a number of projects for information retrieval, database design, expert systems, and natural language processing.

CGs represent meaning sentence by sentence, as a bipartite graph where every arc links 1) a concept node, represented by rectangles (or square brackets), and 2) a conceptual relation node, represented by circles or ovals (parenthesis). In CGs linear form, the English ‘the sky is blue’ would be represented either as (1), (2) or (3) below, depending on the type hierarchy defined by the user:

- (1) [sky] – (Att) → [blue]
- (2) [sky] – (Att) → [color: blue]
- (3) [sky] – (Color) → [blue]

Given that CGs are mathematical structures, they impose no commitments to any concrete notation or implementation. In this sense, the set of conceptual relation nodes (‘att’), as well as the set of concept nodes (‘sky’ and ‘blue’), are not a part of the formalism itself, which is rather a very abstract syntax for knowledge representation.

<sup>1</sup> UNL Center/UNDL Foundation, martins@undl.org

If we consider the cross-linguistic lexical matching problem referred to in the Introduction, the CG formalism, because of its excessive power, will not be of much help. Neither (1), (2) nor (3) brings any clue to the fact that, in this context, but maybe not in others, ‘blue’ is to be translated as [Citema] and [mili] in Shona and Dani, respectively.

The user can obviously refine the semantic granularity of ‘blue’, as to refer to different instances of it, according to the wavelength, for instance: [blue:#436nm], [blue:#437nm], [blue:#438nm], etc. As the Shona and Dani vocabularies would also be indexed to the same scale, one could easily precise the intersection between any language reference, to provide precise translations.

Nevertheless, there is no clear way how to get to such a precise reference, i.e., how to define the wavelength intended by the English ‘blue’ in a phrase like ‘the sky is blue’. It can range from 430nm to 490nm, and still fall out the range intended by [Citema], which goes from 440nm to 510nm. The problem, thus, is not only a matter of lexical matching, but mainly a problem of crossing world references, which are supposed to be different to an English and a Shona speaker. Given the same circumstances, a Shona speaker may come across the idea that the sky is rather ‘green’.

## RESOURCE DESCRIPTION FRAMEWORK

Resource Description Framework (RDF) [9] is a framework for describing and interchanging metadata. It is a general-purpose language for representing (meta-)information in the Web, and it has been the backbone of the Semantic Web initiative, devised by the W3C to improve information retrieval and knowledge processing. An expression in RDF is a directed labeled graph, which consists of nodes (resources, literals or blanks) and labeled directed arcs (properties) that link pairs of nodes. A resource in RDF is anything that can have a URI (Uniform Resource Identifier). This includes web pages, as well as individual elements of an XML document. Formally, as in the case for URL, it is a compact string of characters for identifying an abstract or physical resource. A property is a resource that has a name and can be used to identify the relationship between the nodes connected by the arc. As the arc label may also be a node in the graph, any property can have its own properties. Any statement in RDF consists of a resource (the ‘subject’), a property (the ‘predicate’) and a value (the ‘object’). The value can just be a string, or it can be another resource as well. There is a straightforward method for expressing this in XML:

```
<rdf:Description about=SUBJECT>
<PREDICATE>OBJECT</PREDICATE>
</rdf:Description>
```

In the case for ‘The sky is blue’, the use of RDF can be somewhat misleading, in the sense it is a language for

representing meta-knowledge rather than knowledge itself. However, it should be stressed that, in many circumstances, ‘The sky is blue’ can be said to share the structure of ‘The author of ‘http://www.unl.org’ is ‘UNDL Foundation’’, which seems to stand for a kind of knowledge more frequently represented by means of RDF. Accordingly, there should be possible to take both ‘sky’ and ‘blue’ as different resources (with specific URIs), to be linked by a property of ‘color’. In this case, the knowledge conveyed by ‘The sky is blue’ could be expressed by (4):

```
(4)
<rdf:Description about='sky'>
<color>blue</color>
</rdf:Description>
```

If we take the predicate ‘color’ to be a class in a RDF vocabulary, and the value ‘blue’ to be another resource, rather than a string, the RDF statement would be (5):

```
(5)
<rdf:Description about='sky'>
<color rdf: resource='blue'/>
</rdf:Description>
```

As we consider the lexical matching problem between English, Shona and Dani, we could both say (6) and (7) below:

```
(6)
<rdf:Description about='blue'>
<name>[Citema]</name>
</rdf:Description>
```

```
(7)
<rdf:Description about='blue'>
<name rdf: resource = '[Citema]' />
</rdf:Description>
```

These solutions obviously do not allow for solving any possible mismatching as referred to above. Yet one can associate some hues of ‘blue’ to some hues of [Citema] in some web ontology language, such as OWL, there is no possible way of representing the information that ‘blue’ can be either [citema] or [Cips<sup>w</sup>uka], and that [Citema] can be either ‘blue’ or ‘green’, depending on the context.

Additionally, as RDF properties and classes are not built-in and must be defined by vocabularies (such as the Dublin Core, for instance), even the conception that ‘blue’ and [citema] are ‘colors’ may be defied.

## UNIVERSAL NETWORKING LANGUAGE

The Universal Networking Language (UNL) [10] is an “electronic language for computers to express and exchange every kind of information”. In UNL, information is also

represented sentence by sentence, as a hyper-graph defined by a set of directed binary labeled links (relations) between nodes (Universal Words, or simply UW), which stand for concepts. UWs can also be annotated with attributes representing subjective, mainly deictic, information.

In UNL, the English sentence ‘The sky is blue’ would be represented as a single binary relation:

(8) `aoj(blue(icl>color).@entry, sky(icl>natural world))`

In this expression, ‘aoj’ is a relation (standing for ‘thing with attribute’), ‘blue(icl>color)’ and ‘sky(icl>natural world)’ are UWs, and ‘@entry’ is an attribute. Differently from CG and RDF, UNL relations are built in the very formalism. They constitute a fixed 44-relation set and convey information on ontology structure (such as hyponym and synonym), on logic relations (such as conjunction and condition) and on semantic case (such as agent, object, instrument, etc). The attribute set, which is subject to increase, currently consists of 72 elements, which cope with speaker’s focus, attitudes, viewpoints, etc., towards the event. The set of UWs, which is open, can be extended by the user, but any UW should be also defined in the UNL KB through a master definition (MD).

In the UNL approach, the solution for the lexical color categorization is rather different from the others. There could be concurrently ‘blue(icl>color)’, ‘green(icl>color)’, ‘citema(icl>color)’, ‘cispwuka(icl>color)’, ‘mili(icl>color)’, ‘mola(icl>color)’, etc. These UWs could belong to the UW Dictionary all together, as long as they were defined in the UNL KB. In this sense, a Shona and a Dani speaker may use (9) or (10) below instead of (8):

(9) `aoj(citema(icl>color).@entry, sky(icl>natural world))`

(10) `aoj(mili(icl>color).@entry, sky(icl>natural world))`

Accordingly, there is no need to reduce ‘blue’ to ‘citema’ or vice-versa. Both can be kept, because they represent different concepts.

Translating English into Shona or into Dani may seem, at this extent, rather impossible, at least inside the UNL approach. But UNL is claimed not to be a mere interlingua; it should rather be placed either at the target or at the source position in a bilingual MT system. For that reason, there is no commitment for (8), (9) and (10) to be the same.

The Shona version of ‘The sky is blue’ would be translated (enconverted) into UNL as (9) above because, in the UNL approach, analysis is supposed to be totally independent from any generation process or target language other than UNL itself. To translate [Citema] as ‘blue’ is to perform an English-driven translation movement, i.e., to carry out a Shona analysis from the English perspective. In a Shona-to-Dani translation, such correspondence would be not only pointless but even harmful, as [Citema] is completely comprised under [mili]. Therefore, the UNL expressions (8), (9) and (10), although provoked by very

similar settings, are not planned to be the same, because their reference is not exactly the same, otherwise UNL would impose, over Shona and Dani, an English bias, and cultural differences, as well as non-English-mediated similarities, would be lost.

As a result of this difference-preserving analysis strategy, the English generation out of (8), (9) and (10) may lead to (8a), (9a) and (10a) referred to below:

(8a) The sky is blue.

(9a) The sky is citema.

(10a) The sky is mili.

One may claim that (9a) and (10a) are not English yet and could not be understood by English speakers. This is true, but from (11) and (12) below, which stand for entries in the UNL KB, (9b) and (10b) can be automatically derived:

(11) ‘citema(icl>color)’ definition in the UNL KB

`icl(citema(icl>color), color(icl>property)) = 1;`  
`aoj(between(aoj>thing,obj>thing), citema(icl>color)) = 1;`  
`obj(between(aoj>thing,obj>thing), green(icl>color)) = 1;`  
`fmt(green(icl>color), blue(icl>color)) = 1;`

(12) ‘mili(icl>color)’ definition in the UNL KB

`icl(mili(icl>color), color(icl>property)) = 1;`  
`aoj(dark(aoj>color), color(icl>property)) = 1;`  
`aoj(cold(aoj>color), color(icl>property)) = 1;`

(8b) The sky is of a color in between green and blue.

(9b) The sky is of a cold, dark color.

This is certainly English, and much better translation than just ‘The sky is blue’, which may not convey the meaning intended by the Shona and Dani speaker.

## A BRIEF COMPARATIVE ANALYSIS

The main differences between CG, RDF and UNL seem to concern substance rather than form. All of them are semantic networks represented by hyper-graphs, i.e., a set of links between nodes. In RDF and UNL, these links are labeled; in CG, they are not, but there can be relational nodes. Yet they are very similar, there are at least three striking differences in the UNL perspective: 1) the set of labeled links (or relational nodes) is part of the very formalism; 2) nodes can be annotated by attributes; and 3) in spite of the name ‘‘Universal Word’’, nodes are not supposed to be universal, and can be language-dependent, as long as they are associated to each other in the UNL KB. At least in the case for color naming, this latter commitment seems to be decisive. Color names cannot be conflated, and difference-preserving strategies should be an asset in knowledge representation formalisms. Although this solution may place a heavy burden on the generation (deconverting) process, it should be stressed that, in the UNL System, 1) the UW

definition in the UNL KB is supposed to be registered by the own UW author (the Shona and the Dani speaker, for instance); and 2) the UNL KB is supposed to be a distributed knowledge base, to be hosted in a remote language server, to be accessed worldwide by the internet. This can be taken to alleviate the requirements for the UNL-to-natural language generation system.

It should be stressed that this does not mean that CG or RDF are not able to representing color naming in different cultures. Actually, as they constitute mere formalisms, rather than a theory of knowledge, they can be adapted to represent the same as UNL. The difference, thus, is not a matter of expressive power, but of use. Although they can be said to preserve cultural differences, they have been mainly used for truth-preserving purposes, which is rather difference-conflating. In this sense, the knowledge conveyed by CGs and RDFs is rather encapsulated and culture-dependent. Yet concepts might stand for semantic primitives whose validity is said to be universal, in both frameworks the language bias is somewhat unavoidable. Foreign concepts are useful if and only if they can be reduced to some conceptual primitive that is defined inside a given language (normally English). In UNL this seems not to be the case.

## TOWARDS A CONCLUSION

Irrespective of the nature of the color naming process, whether cultural or natural, whether evolutionary or not, there is no possible one-to-one mapping between color names of English, Shona and Dani. Consequently, the answer to the two questions addressed in the beginning is the paper would be the following: 1) any color name should be represented in a really multilingual and cross-cultural knowledge base, no matter how language-dependent the color name can be. This should be referred to as the comprehensiveness commitment of any KB; 2) in order to allow for reciprocal translatability between different color naming strategies, these color names should be associated to each other by means of complex (relational) definitions. This can be said to be the self-consistency commitment of any KB. As seen above, at least for the time being, it seems that only UNL, for its difference-preserving structure, can really cope with both the comprehensiveness and the self-consistency engagements.

## REFERENCES

- [1] Whorf, B. L. "Science and Linguistics", *Technology Review* 42: 229-231, 1940.
- [2] Lucy, J. A. "The linguistics of color", *Color, Categories in Thought and Language*, Hardin, C. L. and Maffi, L. (eds), Cambridge, England, Cambridge University Press, 1997.
- [3] Dowling, J.E., *The Retina: an approachable part of the brain*, The Belknap Press, Harvard University Press, Cambridge, Massachusetts, 1987.

- [4] Rosch, E. "Universals in color naming and memory", *Journal of Experimental Psychology* 93: 1-20
- [5] Sapir, E., *Language*, New York, Harcourt, Brace, 1921.
- [6] Brent, B. and Kay, Paul, *Basic Color Terms: Their Universality and Evolution*, Berkeley, University of California, 1969.
- [7] Sowa, J. F., *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley, Reading, MA, 1984.
- [8] Sowa, J. F., *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA, 2000.
- [9] Lassila, O. and Swick, R. R. (eds.), *Resource Description Framework (RDF): model and syntax specification*. W3C Recommendation 22 February 1999.
- [10] Uchida, H., Zhu, M., and Della Senta, T. *A gift for a millenium*, Tokyo, IAS/UNU, 1999.