How blue is azzurro? Representing probabilistic equivalency of colour terms in a dictionary

[1,3] **Arvi Tavast**
arvi@tavast.ee
[1,2] **Mari Uusküla**
muuskyla@tlu.ee

[1] Dept. of Translation Studies, School of Humanities, Tallinn University
[2] Institute of the Estonian Language, Tallinn
[3] Dept. of Estonian, University of Tartu

1. INTRODUCTION

Semasiological dictionaries both store and present information about source language words, including their meanings and target-language equivalents. Since this is what dictionary users are consulting a dictionary for, it is the natural way of presenting the dictionary. For compiling and storing dictionary data, however, the solution is less than optimal. The core of the semasiological dictionary data structure is a one-to-many (1:n) relation between words and meanings, i.e. one word can have several meanings, while every meaning has exactly one (source-language) word. In situations of synonymy, information (e.g. definitions and equivalents) must be repeated in each synonym entry, or synonym entries must refer to each other. For both methods or any combinations thereof, semasiological compilation has been shown to cause problems: broken references, synonym conflicts, circularity and inconsistency [1]–[3]. Many of these problems can be avoided using the onomasiological approach for compilation, even if published semasiologically. There is still a 1:n relation, but in the opposite direction: one concept can have multiple designations, while each word has exactly one concept it refers to. Of the problems listed above, the onomasiological data structure makes broken or circular references and synonym conflicts impossible. It still allows inconsistent information to be entered for similar concepts, though. Inconsistencies could only be avoided by systematic terminology work [4], [5].
In terminology, onomasiology is the preferred data structure both in the classical theory [6], [7] and many contemporary approaches [8]–[11]. Due to its limited scalability [12], systematic terminology work is less universally recommended, but has still been successfully used in specialised dictionaries [13], [14]. As a more workable alternative, a partially systematic approach has been used for smaller groups of concepts within a dictionary (many dictionaries of the TSK, e.g. [15]).

Onomasiology is much less known and understood in general lexicography (a notable exception being the Wordnet lexical database [16], [17], which is both onomasiological and partially systematic). In what follows, we argue that taking concepts into account is as feasible and beneficial for the compilation of general language dictionaries as it is in terminology. We start by analysing the equivalents in bilingual dictionaries between Spanish, Italian, English and Estonian, finding that dictionary entries are inconsistent, circular and lacking discriminative information. Moreover, they also contradict the results of our experiments in Castilian Spanish [18] and Standard Italian [19] using the empirical-cognitive field method [20], [21], which we briefly describe in Section 4.

To conclude, we present extracts from the results of our fieldwork as probabilistic conceptual graphs, representing an n:m relation between words and concepts and encoding the likelihood of a word designating a concept. We propose this data structure as an alternative to the 1:n structures of both semasiology and onomasiology, arguing that it is more robust than the former and more intuitive for the lexicographer than the latter.

2. METHOD OF THE DICTIONARY STUDY

We analysed the dictionary equivalents for the Castilian Spanish terms *violeta, morado* and *lila* designating purplish colours, and the Italian terms *blu, azzurro* and *celeste* designating bluish colours, which are well known for their lack of direct equivalents in other languages (we purposefully avoid glossing the example terms throughout the article.). The second languages of the bilingual dictionaries were English and Estonian. The dictionary sources are listed in [22]. The procedure was the following:

1. Look up the headwords in Spanish and Italian, getting the English and Estonian equivalents for each.

2. Look up these equivalents in the opposite language direction, getting their back-translations into Spanish and Italian.

3. Present the results as a directed graph of word equivalence relations.

4. Weigh the edges of the graph according to the number of dictionaries that contained this particular relation.

Variation of hyphenation and parentheses was ignored.

3. RESULTS OF DICTIONARY STUDY

Figures 1-4 present the results of the dictionary study. Estonian dictionary data is more sparse due to the smaller number and volume of dictionaries with this language. Dictionaries between English and Spanish (Figure 1) stand out in terms of having a clear convergence of equivalent pairs across dictionaries. The remaining three language pairs exhibit much more variation. The following can be observed to some degree in all four language pairs:

- Equivalents tend to follow orthographic similarities across languages.
- Headword selection is not comprehensive, with even the most frequent colour terms sometimes missing from dictionaries.
- Dictionaries of a single language direction contradict each other and do not justify their choice of equivalents, leaving the user with a seemingly random set of equivalent candidates.
- Dictionaries of two opposite language directions contradict each other.

In Spanish-Estonian dictionaries, the term *morado* could be either *punakasvioletne* 'reddish purple', *tume punakaslilla* 'dark reddish purple', *tumelilla* 'dark purple' or the Estonian basic term *lilla* 'purple'. The term violet could be *violett, violetne* or *lilla* (the difference between *violett/violetne* and *lilla* in Estonian could mainly be accounted for through people's idiolects or individual preferences (see more on Estonian purple terms in [23]). The term *lila* could be *(hele)lilla* or *lilla*. Each of the three terms was absent from at least one of the dictionaries. According to the dictionaries, all three Spanish terms could correspond to the Estonian basic term *lilla*. In Estonian-Spanish dictionaries the term *lilla* had also a many different counterparts

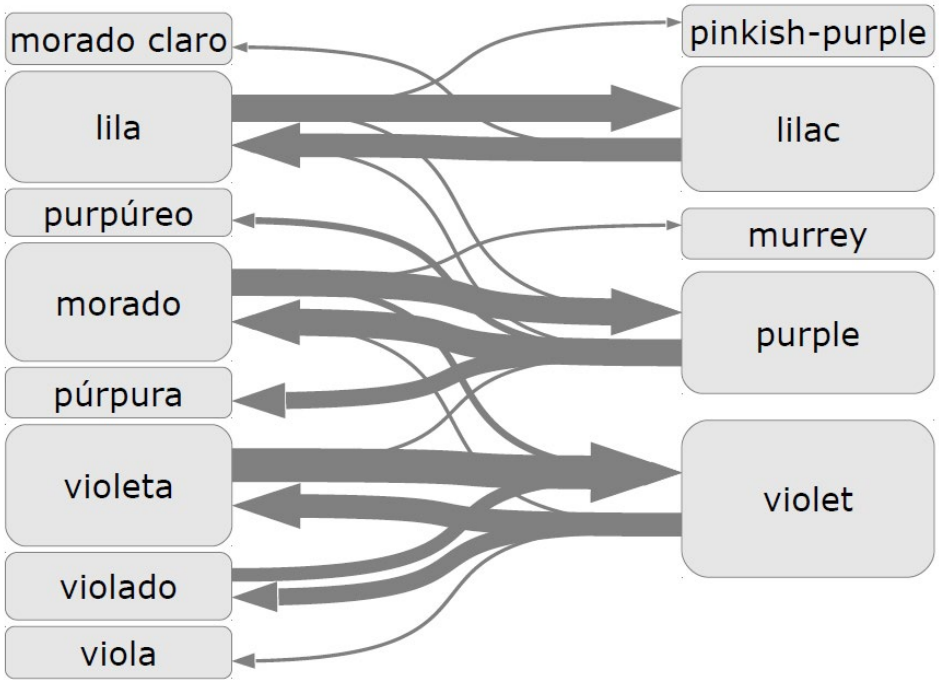

Figure 1 – Purplish colours in Spanish-English and English-Spanish dictionaries

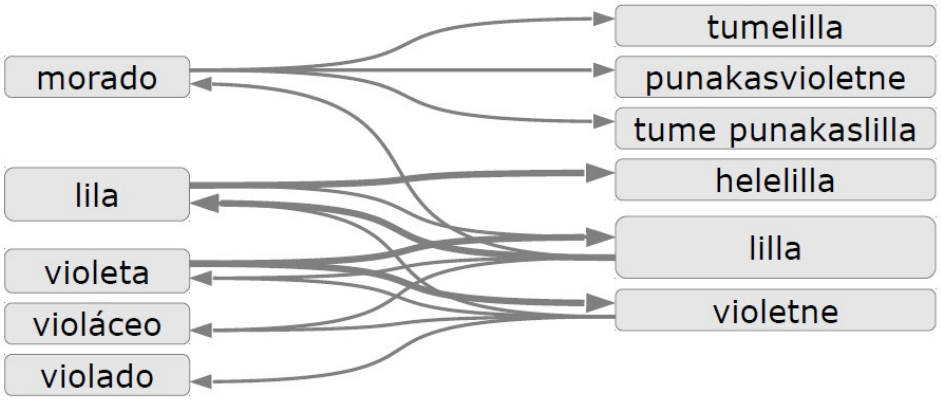

Figure 2 - Purplish colours in Spanish-Estonian and Estonian-Spanish dictionaries

(*violado, violáceo, de color lila, violeta, lila, morado*), without any explanation about the conceptual differences. The terms tend to get counterparts through homography, perhaps neglecting the spectrum that the colour terms represent. We could think that the similar word shape is due to the same etymological background and that assures the counterparts, but languages develop differently and words, though from the same source, develop different meanings through cultural impact [24]. For example, Spanish violeta and *lila* have a very transparent source. However, *morad*o, being a loan from Latin *mōrum* 'mulberry' has a very deep cultural meaning. Perhaps the cultural importance and a different origin of morado has made the terms for purple divide differently from other languages.

4. METHOD OF THE FIELDWORK

The field data was obtained using an empirical-cognitive field method following [20], [21]. The method consists of two tasks. In the list task, the participants were asked to name all the colour terms they could think of. In the colour-naming task, 65 matt-surfaced coloured stimuli from the Color-aid Corporation 220 selection were presented to the participants one by one in a random sequence and they were asked to name the perceived stimuli with the appropriate colour term . 65 stimuli were 5x5 cm plywood squares constituting a "coarse, but evenly spread sample of colour space" ([21]; for seletion criteria see [25]). Color-aid is based on Ostwald colour system: each colour can be described by CIE coordinates available in [25]. The tiles were shown to participants in a random order in natural daylight on a neutral gray surface (comparable to Munsell N2). Lighting conditions were similar for all participants. The participants were allowed to use simple words, compounds or even phrases.

The Standard Italian data was collected between 2006 and 2008 in Florence (102 participants, 56 female, age range 11-80, mean 38.6). The Castilian Spanish data was collected in 2012 in Madrid (38 participants, 20 female, age range 22-85, mean 42.7). Participants were all

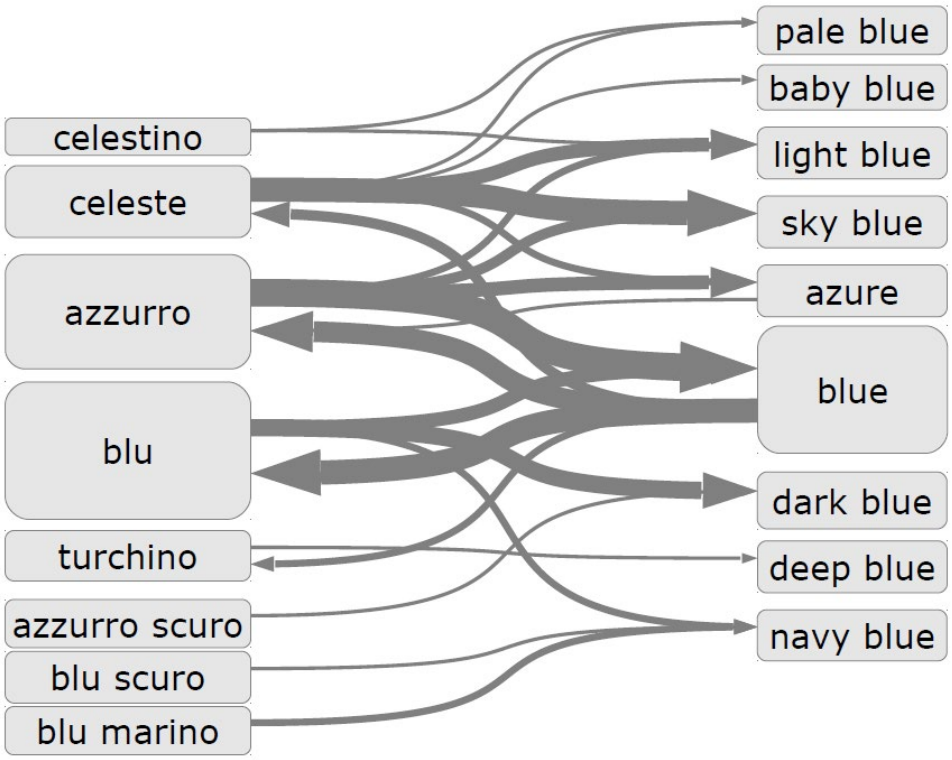

Figure 3 – Bluish colours in Italian-English and English-Italian dictionaries

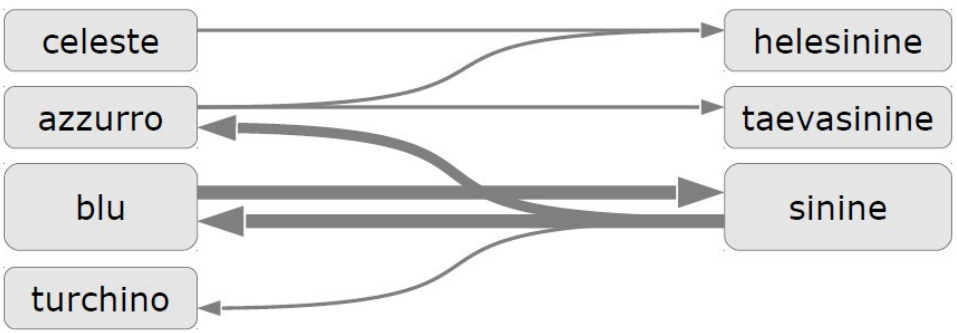

Figure4 - Bluish colours in Italian-Estonian and Estonian-Italian dictionaries

volunteers with different dialectal, educational and occupational backgrounds (for details see [19]). The interviews were carried out in Castilian Spanish or Standard Italian by a proficient L2 speaker. Estonian data was collected in 2014 (20 participants, 12 female, age range 25-48, mean 31.7). All participants had normal colour vision, as ascertained using either the City University [26] or Ishihara's [27] colour vision tests

5. FIELDWORK RESULTS

In the list task, we calculated the naming frequency of a colour term, its mean position in the list and the cognitive salience index which unifies these two parameters [28]. In the colour naming task, we took into account the term frequency, the number of tiles assigned to each colour term, and calculated dominance and specificity indices [20] to examine the consensus rate among the participants [18], [29]. We observe which Color-aid tiles the Castilian Spanish terms for purple and Standard Italian terms for blue are attached to. Spanish *morado* was given with the highest frequency to tiles VBV (60%) and VRV (58%), *violeta* to tile VBV-T4 (42%) and *lila* to tile VRV-S3 (37%). Italian *blu* was most frequently attached to colour tile BVB (named by 54% of participants), while BGB-T3 was regarded as *celeste* by 57% of respondents. *Azzurro* was used to describe colour tile BGB by 44% of participants.

Figures 5 and 6 represent extracts from our fieldwork results. The graphs differ from the bilingual graphs of Figures 1-4 by the addition of concepts (here represented by codes of colour stimuli) between the languages. Edge weights encode the percentage of respondents naming this stimulus with this term and were cut off at 10% (which is why the total of weights for each stimulus is generally less than 100). The graphs contain all stimuli and terms within 3 hops from the original terms, resulting in the inclusion of some terms that intuitively would not belong there. Italian *verde*, for instance, is included among the bluish colours on Figure 6 because 10% of Estonian respondents called the stimulus BG-S2 *sinine* 'blue' and 24% of the Italian respondents called the same stimulus *verde*.

6. DISCUSSION

The observed overlaps across categories and shifts across languages illustrate the inadequacy of direct univocal equivalences postulated in dictionaries. Colour terms are an exceptionally easy semantic domain to perform such analysis on, due to the relative ease of presenting colour stimuli to participants - a similar experiment with modal verbs or abstract nouns would be quite complex if not impossible.

However, the same overlaps and shifts are still there regardless of how much is known about the concept, causing the same dangers of misrepresenting linguistic reality in the dictionary. This paper suggests that words should be related to each other only through concepts, since direct relations (synonymy and equivalence) are not flexible enough to

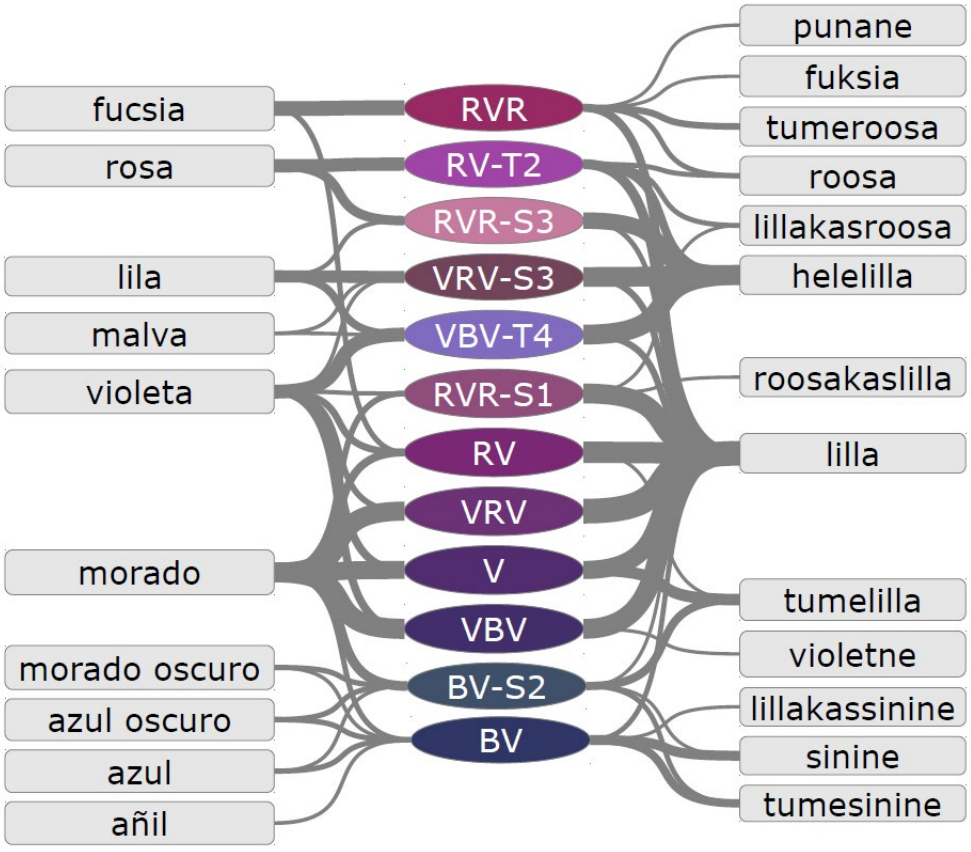

Figure 5 – Purplish colours in Spanish and Estonian fieldwork results

represent the complexity of natural language. We hope that our solution to using an n:m relation between words and concepts instead of the 1:n of traditional onomasiology will make the approach less daunting to lexicographers by removing the need to have separate headword entries for each meaning of a word.

We used probabilistic weighting on our graphs to account for the fact that one meaning of a word can be more likely than another. Here the probabilities were obtained from elicited performance experiments, but other sources could be used as well, some of which do not depend on the semantic class, e.g. vector semantics or word-level alignment of parallel corpora.

Our fieldwork results are consistent with previous studies that underline the special status of blue category in Italian (e.g. [19], [30]–[32]). The status of the purple category has been discussed by [34] and [35]. While [35] regard morado and violeta as synonyms, our field data indicates that these colour terms have different conceptual references.

Our probabilistic conceptual graphs are similar to Dyvik's semantic mirroring [36], [37] and the workflow of the EFNILEX project [38], [39], differing from them by the explicit addition of concepts into the graph instead of relying on isolation of subgraphs to identify synsets. The objectives of [37] do include obtaining relations between the explicit concepts of Wordnet from parallel corpora; what we add is retention of the probability information rather than reducing it to discrete relations. Finally, from systematic terminology work ([4], [5]) we differ by the use of probabilities, but also the n:m relations between words and concepts.

ACKNOWLEDGEMENTS

We are indebted to all our Castilian Spanish, Standard Italian and Estonian test participants, and to Triin Kalda who helped to collect the Estonian data.

NOTES

[1]To be precise, concepts are subjective abstractions of indvidually perceived or imagined objects, dependent on a wide range of variables from medical issues to life experience to cultural norms, not the physical objects (colour cards) themselves. The experimental stimuli merely invoke the processes of perception and categorisation, which can then be followed by finding and uttering a name for the resulting concept. The variation in naming includes the variation in perception, and since the experimenter has no access to subjective processes of the participant, these two can not be separated in the current experiment. In any reference to colour (or any other phenomena) in this paper, perception is always implied to be present as an additional degree of freedom. The use of colour tiles as stimuli in this study is only motivated by the fact that there is much less room for variation in the perception of colours compared to the perception of e.g. kindness or running or even bird or table.

BIBLIOGRAPHY

[1] A. Tavast, "Eesti oskussõnastikud 1996-2000," Keel ja Kirjandus, vol. 6, 7, pp. 401–414, 489–503, 2002.

[2] A. Tavast, The Translator is Human Too: A Case

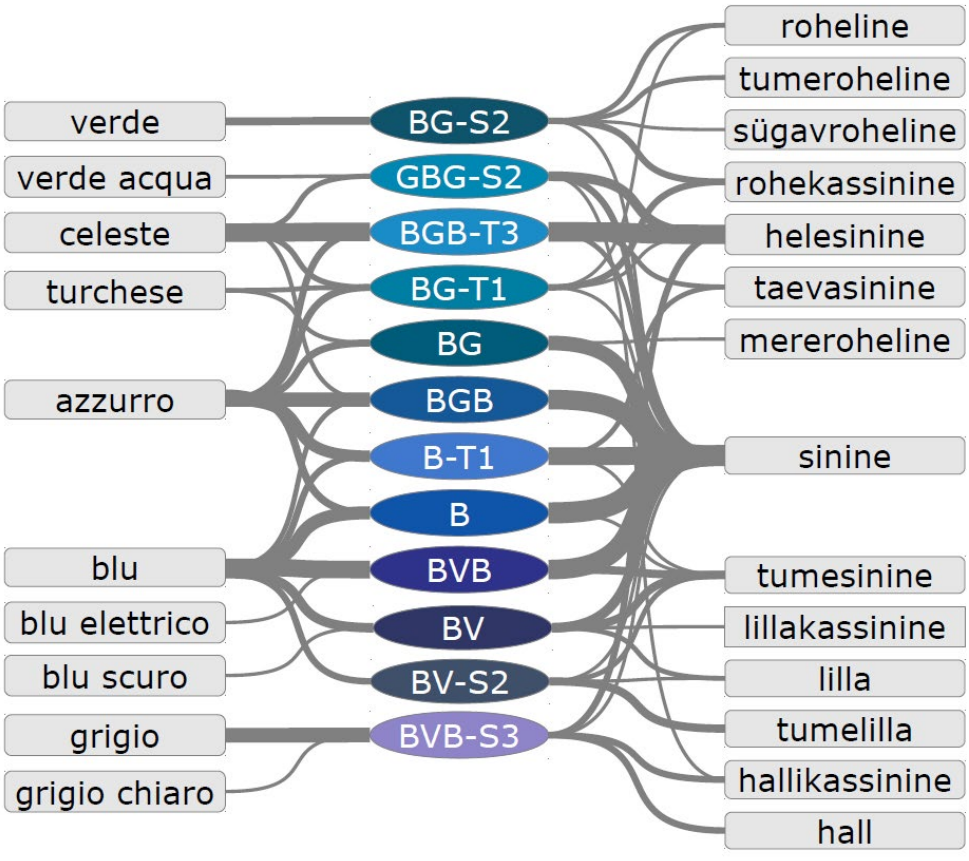

Figure 6 – Bluish colours in Italian and Estonian fieldwork results

for Instrumentalism in Multilingual Specialised Communication. Tartu: Tartu Ülikooli Kirjastus, 2008.

[3] A. Tavast, "Eesti oskussõnastikud 2001–2010," Keel ja Kirjandus, vol. 4, pp. 255–276, 2011.

[4] S. E. Wright and G. Budin, Handbook of Terminology Management: Application-oriented Terminology Management. Amsterdam: John Benjamins, 2001.

[5] H. Suonuuti, Guide to Terminology. Helsinki: Tekniikan Sanastokeskus, 2001.

[6] E. Wüster, Einführung in die allgemeine Terminologielehre und terminologische Lexikographie. Dordrecht: Springer, 1979.

[7] H. Felber, Terminology Manual. Paris: UNESCO, 1984.

[8] R. Temmerman, Towards New Ways of Terminology Description: The Sociocognitive Approach. Amsterdam: John Benjamins, 2000.

[9] M. T. Cabré, La terminología: representación y comunicación : elementos para una teoría de base comunicativa y otros artículos. Girona: Documenta Universitaria, 1999.

[10] P. Faber Benítez, "The cognitive shift in terminology and specialized translation," MonTI. Monografías de Traducción e Interpretación, no. 1, pp. 107–134, 2009.

[11] P. Faber Benítez, A Cognitive Linguistics View of Terminology and Specialized Language. Berlin: Walter de Gruyter, 2012.

[12] H. Picht and J. Draskau, Terminology: An Introduction. Guildford: University of Surrey, 1985.

[13] E. Wüster, The Machine Tool: An Interlingual Dictionary of Basic Concepts: Comprising an Alphabetical Dictionary and a Classified Vocabulary with Definitions and Illustrations. London: Technical Press, 1967.

[14] Eduskuntasanasto = Riksdagsordlista = Finnish Parliamentary Glossary = Soome Parlamendi Seletussõnastik. Helsinki: Eduskunta, 2008.

[15] T. S. K. Sanastokeskus, Geoinformatiikan sanasto. Sanastokeskus TSK, 2011.

[16] G. A. Miller, "WordNet: a lexical database for English," Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[17] C. Fellbaum, WordNet. Dordrecht: Springer, 2010.

[18] K. Parker, "Hispaania värvinimed sõnastikes ja mentaalses leksikonis," MA thesis, Tallinn University, Tallinn, 2013.

[19] M. Uusküla, "Linguistic categorization of blue in Standard Italian," in Colour Studies: A broad spectrum, John Benjamins Publishing Company, 2014, pp. 67–78.

[20] I. Davies and G. Corbett, "The basic color terms of Russian," Linguistics, vol. 32, no. 1, pp. 65–90, 1994.

[21] I. R. L. Davies and G. G. Corbett, "A practical field method for identifying probable basic colour terms," Languages of the World, vol. 9, no. 1, pp. 25–36, 1995.

[22] A. Tavast, M. Uusküla, K. Parker and U. Sutrop, "Using probabilistic conceptual graphs for representing colour terms in dictionaries," in Color and Colorimetry Multidsciplinary Contributions, Maggioli Editore, 2013, pp. 455-466.

[23] V. Oja and M. Uusküla, "Mõnest värvinimetusest ja nende tähendusvahekordadest eesti ja soome keeles," Eesti Rakenduslingvistika Ühingu aastaraamat, no. 6, pp. 195–205, 2010.

[24] C. P. Biggam, The semantics of colour: a historical approach. Cambridge University Press, 2012.

[25] I. R. Davies, C. Macdermid, G. G. Corbett, H. McGurk, D. Jerrett, T. Jerrett, and P. Sowden, "Color terms in Setswana: a linguistic and perceptual approach," Linguistics, vol. 30, no. 6, pp. 1065–1104, 1992.

[26] R. Fletcher, The City University Colour Vision Test, 2nd ed. 1980.

[27] S. Ishihara, Ishihara's tests for colour-deficiency. Kanehara & Company, 1996.

[28] U. Sutrop, "List task and a cognitive salience index," Field Methods, vol. 13, no. 3, pp. 263–276, 2001.

[29] M. Uusküla, "Mediterranean Ecology and the colour blue in Standard Italian.," in The Language of Color in the Mediterranean, 2nd ed., forthcoming.

[30] G. Paggetti, G. Menegaz, and G. V. Paramei, "Color naming in Italian language," Color Res. Appl., p. n/a–n/a, Feb. 2015.

[31] D. Bimler and M. Uusküla, "'Clothed in triple blues': sorting out the Italian blues," JOSA A, vol. 31, no. 4, pp. A332–A340, 2014.

[32] G. V. Paramei, M. D'Orsi, and G. Menegaz, "'Italian blues': A challenge to the universal inventory of basic colour terms," JAIC - Journal of the International Colour Association, vol. 13, 2014.

[33] J. Sandford, "Blu, Azzurro, Celeste: what color is blue for Italian speakers compared to English speakers?" Colour and Colorimetry: Multidisciplinary Contributions, pp. 281-288, 2012.

[34] L. Rello, "Términos de color en español: semántica, morfología y análisis lexicográfico. Definiciones y matices semánticos de sus afijos."

[35] J. Lillo, H. Moreira, I. Vitini, and J. Martín, "Locating basic Spanish colour categories in CIE L* u* v* space: Identification, lightness segregation and correspondence with English equivalents," Psicológica, vol. 28, no. 1, pp. 21–54, 2007.

[36] H. Dyvik, "A translational basis for semantics," Language and Computers, vol. 24, pp. 51–86, 1998.

[37] H. Dyvik, "Translations as semantic mirrors: from parallel corpus to wordnet," Language and computers, vol. 49, no. 1, pp. 311–326, 2004.

[38] E. Héja and D. Takács, "Automatically generated customizable online dictionaries," in Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 51–57.

[39] E. Héja and D. Takács, "An online dictionary browser for automatically generated bilingual dictionaries," in Proceedings of the 15th EURALEX International Congress, 2012, pp. 468–477.