A Feature Type Classification for Therapeutic Purposes: a preliminary evaluation with non-expert speakers

Gianluca E. Lebani
University of Trento
gianluca.lebani@unitn.it

Emanuele Pianta
Fondazione Bruno Kessler
pianta@fbk.eu

Abstract

We propose a feature type classification thought to be used in a therapeutic context. Such a scenario lays behind our need for a easily usable and cognitively plausible classification. Nevertheless, our proposal has both a practical and a theoretical outcome, and its applications range from computational linguistics to psycholinguistics. An evaluation through inter-coder agreement has been performed to highlight the strength of our proposal and to conceive some improvements for the future.

1 Introduction

Most common therapeutic practices for anomia rehabilitation rely either on the therapist's intuitive linguistic knowledge or on different kinds of resources that have to be consulted manually (Semenza, 1999; Raymer and Gonzalez-Rothi, 2002; Springer, 2008). STaRS.sys (Semantic Task Rehabilitation Support system) is a tool thought for supporting the therapist in the preparation of a semantic task (cfr. Nickels, 2002).

To be effective, such a system must lean on a knowledge base in which every concept is associated with different kinds of featural descriptions. The notion of feature refers to the linguistic descriptions of a property that can be obtained by asking a subject to describe a concept. Examples of concept-feature pairings will be represented here as <*concept*> *feature*[1] couples such as <*dog*> *has a tail* or <*dog*> *barks*.

As a consequence of this scenario, an intuitive and cognitively plausible classification of the feature types that can be associated with a concept is a vital component of our tool. In this paper, we present a classification that meets such criteria, built by moving from an analysis of the relevant proposals available in the literature.

We evaluated our classification by asking to a group of naive Italian speakers to annotate a test set by using our categories. The resulting agreement has been interpreted both as an index of reliability and as a measure of ease of learning and use by non-expert speakers. In these preliminary phases we focus on Italian, leaving to future evaluations whether or how to extend the domain of our tool to other languages.

These pages are organized as follows: in Section 2 we briefly review the relevant works for the following discussion. In Section 3 we introduce our classification and in the remaining part we evaluate its reliability and usability.

2 Related Works

2.1 Feature Norms

In the psychological tradition, a collection of feature norms is typically built by asking to a group of speakers to generate short phrases (i.e. features) to describe a given set of concepts.

Even if normative data have been collected and employed for addressing a wide range of issues on the nature of the semantic memory, the only freely available resources are, to our knowledge, those by Garrard et al (2001), those by McRae et al (2005), those by Vinson and Vigliocco (2008), all in English, and the Dutch norms available in the Leuven database (De Deyne et al, 2008).

Moving out of the psychological domain, the only collection built in the lexicographic tradition is that by Kremer et al (2008), collected from Italian and German speakers

[1] Typographical conventions: concepts, categories and features will be printed in *italics courier new* font. When reporting a concept-feature pair, the concept will be further enclosed by <*angled brackets*>. Feature types and classes of types will be both reported in times roman, but while the formers will be written in *italics*, type classes will be in SMALL CAPITALS.

2.2 Related Classifications

The proposals that constitute our theoretical framework have been chosen for their being either implemented in an extensive semantic resource, motivated by well specified theoretical explanations (on which there is consensus) or effectively used in a specific therapeutic context. They have originated in research fields as distant as lexicography, theoretical linguistics, ontology building, (clinical) neuropsychology and cognitive psychology. Specifically, the works we moved from have been:

- a type classification adopted for clinical purposes in the CIMeC's Center for Neurocognitive Rehabilitation (personal communication);
- the knowledge-type taxonomy proposed by Wu & Barsalou (2009), and the modified version adopted by Cree & McRae (2003);
- the brain region taxonomy proposed by Cree & McRae (2003);
- the semantic (but not lexical) relations implemented in WordNet 3.0 (Fellbaum, 1998) and in EuroWordNet (Alonge et al, 1998);
- the classification of part/whole relations by Winston et al (1987);
- the SIMPLE-PAROLE-CLIPS Extended Qualia Structures (Ruimy et al, 2002).

3 STaRS.sys feature types classification

The properties of our classification follow from the practical use scenario of STaRS.sys. In details, the fact that it's thought to be used in a therapeutic context motivates our need for a classification that has to be: (1) intuitive enough to be easily used by therapist and (2) robust and (3) cognitively plausible so as to be used for preparing the relevant kinds of therapeutic tasks.

Furthermore, being focused on features produced by human speakers, the classification applies to the linguistic description of a property, rather than to the property itself. Accordingly, then, pairings like the following:

`<plane> carries stuff`
`<plane> is used for carrying stuff`

are though as instances of different types (respectively, *is involved in* and *is used for*).

Starting from an analysis of the relevant proposals available in the literature, we identified a set of 26 feature types, most of which have been organized into the following six classes:

TAXONOMIC PROPERTIES: Two types related to the belonging of a concept to a category have been isolated: the *is-a* and the *coordinate* types.

PART-OF RELATIONS: We mainly followed Winston et al's (1987) taxonomy in distinguishing six types describing a relation between a concept and its part(s): *has component*, *has member*, *has portion*, *made-of*, *has phase* and *has geographical part*.

PERCEPTUAL PROPERTIES: Inspired by the Cree and McRae's (2003) brain region taxonomy, we isolated six types of perceivable properties: *has size*, *has shape*, *has texture*, *has taste*, *has smell*, *has sound*, *has colour*.

USAGE PROPERTIES: This class is composed by three types of an object's use descriptions: *is used for*, *is used by* and *is used with*.

LOCATIONAL PROPERTIES: We identified three types describing the typical *situation*, *space* and *time* associated to an object.

ASSOCIATED EVENTS AND ATTRIBUTES: This class encompasses three kinds of information that can be associated to an object: its emotive property (*has affective property*), one of its permanent properties (*has attribute*) and the role it plays in an action or in a process (*is involved in*). As a matter of fact, each of the other classes is a specification of one of the two latter types, to which particular relevance has been accorded due to their status from a cognitive point of view.

Others: Two feature types fall out of this classification, and constitute two distinct classes on their own. These are the *has domain* type, that specifies the semantic field of a concept, and the dummy *is associated with*, used for classifying all those features that falls out of any other label.

Comparison and final remarks: A quick comparison between our types and the other classifications reveals that, apart from the *is used with* type, we didn't introduce any new opposition. Any type of ours, indeed, has a parallel type or relation in at least one of the other proposals. Such a remark shows what is the third major advantage of our classification, together with its usability and its cognitive plausibility: its compatibility with a wide range of well known theoretical and experimental frameworks, that allows it to serve as a common ground for the interplay of theories, insights and ideas originated from the above mentioned research areas.

4 Evaluation

Given the aims of our classification, and of STaRS.sys in general, we choose to evaluate our coding scheme by asking to a group of non experts to label a subset of the non-normalized Kremer et al's (2008) norms and measuring the

inter-coder agreement between them (Artstein and Poesio, 2008), adhering to the Krippendorff's (2004, 2008) recommendations.

The choice to recruit only naive subjects has the positive consequence of allowing us to draw inferences also on the usability of our proposal. That is, such an evaluation can be additionally seen as a measure of how easily a minimally trained user can understand the oppositions isolated in our classification.

4.1 Experimental Setup

Participants: 5 Italian speakers with a university degree were recruited for this evaluation. None of them had any previous experience in lexicography, nor any education in lexical semantics.

Materials: 300 concept-feature pairs were selected mainly from a non-normalized version of the Kremer et al's (2008) norms. We choose this dataset because (1) it's a collection of descriptions generated by Italian speakers and (2) we wanted to avoid any bias due to a normalization procedure, so as to provide our subjects with descriptions that were as plausible as possible.

The experimental concept-attribute pairs have been chosen so to have the more balanced distribution of concepts and feature types as possible, by not allowing duplicated pairs. As for the concepts, an uniform distribution of features per category (30 feature for all the ten categories of the original dataset) and of features per concept (i.e. between 4 and 7) has been easily obtained.

The attempt to balance feature types, however, has revealed impracticable, mainly due to the nature of the concepts of the Kremer's collection and to the skewness of its type distribution. Therefore, we fixed an arbitrary minimum threshold of ten plausible features per type. Plausible features have been obtained from a pilot annotation experiment performed by one author and an additional subject. We further translated 23 concept-feature pairs from the McRae (11 cases) and from the Leuven (12 cases) datasets for balancing types as much as possible.

Still, it has not been possible to find ten features for the following types: *has Geographical Part*, *has Phase* and *has Member* (no features at all: this is a consequence of the kind of concept represented the dataset), *has Portion* (only four cases, again, this is a consequence of the source dataset), *has Domain* (5) and *has Sound* (6). We nevertheless decided to include these types in the instructions and the relevant features in the test set. Our decision has been motivated by the results of the pilot experiment, in which the subjects made reference to such types as a secondary interpretation in more than ten cases.

Procedure: The participants were asked to label every concept-feature pair with the appropriate type label, relying primarily on the linguistic form of the feature. They received a 17-pages booklet providing an explanation of the annotation goals, a definition and some examples for every type class and for every type, a decision flowchart and a reference table.

Every participant was asked to read carefully the instructions, to complete a training set of 30 concept-feature pairs and to discuss his/her decisions with one of the two authors before starting the experimental session. The test set was presented as a unique excel sheet. On the average, labeling the 300 experimental pairs took 2 hours.

4.2 Results

The annotations collected from the participants have been normalized by conflating direct (e.g. *is-a*) and reverse (e.g. *is the Category of*) relation labels, and the agreement between their choice has been measured adopting Fleiss' Kappa. The "Kappa: annotators" column of Table 1 reports the general and the type-wise kappa scores[2] for the annotations of the participants.

Feature Type	***Kappa: annotators***	***Kappa: gold/majority***
is-a	0.900	0.956
coordination	0.788	0.913
has component	0.786	0.864
has portion	0.558	0.747
made of	0.918	0.955
has size	0.912	1
has shape	0.812	1
has texture	0.456	0.793
has taste	0.852	1
has smell	0.865	1
has sound	0.582	0.795
has colour	0.958	1
is used for	0.831	0.727
is used by	0.964	1
is used with	0.801	0.939
situation located	0.578	0.854
space located	0.808	0.898
time located	0.910	0.946
is involved in	0.406	0.721
has attribute	0.460	0.746
has affective property	0.448	0.855
has domain	0.069	0.277
is associated with	0.141	0.415
General	**0.73**	**0.866**

Table 1: Type-wise agreement values

[2] All reported Kappa values are associated with $p < 0.001$.

Even if there is no consensus on how to interpret Kappa values in isolation, and despite the fact that, to our knowledge, this is the first work of this kind, we can nevertheless draw interesting conclusions from the pattern in table 1. The general Kappa score has a value of 0.73, and the agreement values are above 0.8 for 12 types, not so distant in 2 cases, and well above 0.67 for 9 types, 5 of which are our "residual" categories, that is, those that are more "general" that at least one of the other types[3].

Such a contrast between the residual and the other types is even more pronounced in the class-wise analysis, where the only Kappa value below the 0.8 threshold is the one obtained for the ASSOCIATED EVENTS AND ATTRIBUTES class (κ = 0.766) [4]. Furthermore, the distribution of false positives in a confusion matrix between the performance of the annotators and the "majority" vote[5] shows that part of the low agreement for the residual types is due to the "summation" of the disagreement on the other categories. Obviously, part of this variance is due also to the fact that such types have fuzzier boundaries, and so are more difficult to handle.

As for the remaining four low agreement types, two of them (*has affective property*, *has domain*) have been signaled by the annotators to be difficult to handle, while the remaining two (*has sound*, *has portion*) have been frequently confused with one of the ASSOCIATED EVENTS AND ATTRIBUTES types and with the *has component* type, respectively. Such results are not very puzzling for the *has domain* and *has portion* types, given the technicality of the former and, for the latter, the nature of the described concepts. They do point, however, to a better definition of the remaining two types, the *has sound* and *has affective property* ones, in that most difficulties seem to arise from an unclear definition of their respective scopes.

As pointed out by Artstein and Poesio (2008), agreement doesn't ensure validity. In trying to evaluate how our annotators "did it right", we measured the exact Kappa coefficient (Conger, 1980) between the majority annotation (i.e. what annotators should have done to agree) and the annotation of the same set by one of the two authors. With some approximation, we see this last performance as the "right" one.

Results are reported in the "Kappa: gold/majority" column of Table 1. The general Kappa value is well above 0.8, and so it is for 15 of the 23 types. Only two types (*has domain* and *is associated with*) are below the 0.67 minimal threshold. These data further confirm the difficulties in handling residual types, but, more importantly, seem to suggest that our "gold standard" annotator have been able to learn the classification in a fairly correct way (at least, it did in a way similar as one of the two authors of this classification).

4.3 Discussion

We interpret the results of our evaluation as a demonstration of the reliability of our coding scheme as well as of the usability of our classification, at least as the non residual types are concerned. For the future, many improvements are suggested by our data. In particular, they showed the need of the annotators to receive a better training on some relations and distinctions.

This points in the direction of both a more deep training on the types we've dubbed as "residuals", and of a better definition of poorly understood types such as *has domain* and *has affective property* and puzzling distinctions such as the *has smell*/*is Involved in* ones.

5 Conclusions and Future Directions

In this paper we introduced a classification of the information types that can be expressed to describe a concrete concept. Even if we thought this classification mainly for therapeutic purposes, its use can be broadened to include a wide range of possible NLP tasks.

We evaluated our proposal by asking a group of naive speakers to annotate a list of concept-feature pairs with the appropriate label. Even if our results can't be interpreted as absolutely positive, we consider them promising, in that (1) the skeleton of the classification seems to have been validated by the performance of our participants and (2) a great part of the disagreement seems to be solvable through major care in the training phase. In the near future we are going to test our (improved) coding scheme with annotators from the population of the STaRS.sys final users, i.e. therapist with experience in semantic therapy. Finally, further research is needed to assess if and to what extent the semantic model underlying our classification is compatible with those of existing lexical and/or semantic resources.

[3] Our residual labels are *has Attribute*, *has Texture*, *is Associated with*, *is Involved in* and *Situation Located*.

[4] The general Fleiss' Kappa value for the class-wise comparison is 0.766.

[5] That is, the performance obtained by assigning the label chosen by the majority of the annotators.

Acknowledgments

We are grateful to the annotators who gave us the data reported in Section 4 and to all the CLICers that commented our classification. In particular, we would like to thank dr. Federica Cavicchio for their statistical advice and Gerhard Kremer for providing us with a non-normalized version of his dataset.

Reference

Antonietta Alonge, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Maria A. Marti and Wim Peters. 1998. The linguistic design of the EuroWordNet database. *Computer and the Humanities*, 32: 91-115.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34 (4): 555-596.

Anthony J. Conger. 1980. Integration and generalisation of Kappas for multiple raters. *Psychological Bulletin*, 88: 322-328.

George S. Cree and Ken MCrae. 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132 (2): 163-201

Simon De Deyne, Steven Verheyen, Eef Amel, Wolf Vanpaemel, Matthew J. Dry, Wouter Voorspoels and Gert Storm. 2008. Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, 40 (4): 1030-1048.

Christiane Fellbaum. 1998. *WordNet. An electronic lexical database*. The MIT Press. Cambridge, MA.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5): 378-382.

Peter Garrard, Matthew A. Lambon Ralph, John R. Hodges and Karalyn Patterson. 2001. Prototypicality, distinctiveness and intercorrelation: analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18 (2): 125-174.

Gerhard Kremer, Andrea Abel and Marco Baroni. 2008. Cognitively salient relations for multilingual lexicography. *Proceedings of COLING-CogALex Workshop 2008*: 94-101.

Klaus Krippendorff. 2004. Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research*, 30 (3): 411-433.

Klaus Krippendorff. 2008. Testing the reliability of content analysis data: what is involved and why. In K. Krippendorff and M.A. Bock (eds.). *The Content Analysis Reader*. Sage, Thousand Oaks, CA: 350-357.

Ken McRae, George S. Cree, Mark S. Seidenberg and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments & Computers*, 37 (4): 547-559.

Gregory L. Murphy. 2002. *The big book of concepts*. The MIT Press, Cambridge, MA.

Lyndsey Nickels.2002. Therapy for naming disorders: revisiting, revising, and reviewing. *Aphasiology*, 16 (10/11): 935-979

Anastasia M. Raymer and Leslie J. Gonzalez-Rothi. 2002. Clinical diagnosis and treatment of naming disorders. In A.E. Hillis (ed.). *Handbook of Adult Language Disorders*. Psychology Press: 163-182.

Eleanor Rosch and Carolyn B. Mervis. 1975. Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, 7: 573-605.

Nilda Ruimy, Monica Monachini, Raffaella Distante, Elisabetta Guazzini, Stefano Molino, Marisa Ulivieri, Nicoletta Calzolari and Antonio Zampolli. 2002. Clips, a multi-level Italian computational lexicon: a glimpse to data. *Proceedings LREC 2002*: 792-799.

Carlo Semenza. 1999. Lexical-semantic disorders in aphasia. In G. Denes and L. Pizzamiglio (eds.). *Handbook of clinical and experimental neuropsychology*. Psychology Press, Hove: 215-244.

Luise Springer. 2008. Therapeutic approaches in aphasia rehabilitation. In B. Stemmer and H. Whitaker (eds.) *Handbook of the Neuroscience of Language*. Elsevier Science, : 397-406.

David P. Vinson and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40 (1): 183-190.

Morton E. Winston, Roger Chaffin and Douglas Herrman. 1987. A taxonomy of part-whole relation. *Cognitive Science*, 11:417-444.

Ling-ling Wu and Lawrence W. Barsalou. 2009. Perceptual Simulation in conceptual combination: evidence from property generation. *Acta Psychologica*, 132: 173-189.