Migrating Psycholinguistic Semantic Feature Norms into Linked Data in Linguistics

Yoshihiko Hayashi
Graduate School of Language and Culture, Osaka University
1-8 Machikaneyma, Toyonaka 5600043, Japan
hayashi@lang.osaka-u.ac.jp

Abstract

Semantic feature norms, originally utilized in the field of psycholinguistics as a tool for studying human semantic representation and computation, have recently attracted the attention of some NLP/IR researchers who wish to improve their task performances. However, currently available semantic feature norms are, by nature, not well-structured, making them difficult to integrate into existing resources of various kinds. In this paper, by examining an actual set of semantic feature norms, we investigate which types of semantic features should be migrated into Linked Data in Linguistics (LDL) and how the migration could be done.

1 Introduction

Recently, some NLP/IR researchers have become interested in incorporating psycholinguistic features into their applications to improve task performance (Kwong, 2012; Tanaka et al., 2013). Among a range of psycholinguistic features, such as imageability, concreteness, and familiarity (Paivio et al., 1968), the most attractive is a set of semantic feature norms introduced by McRae et al. (2005). It captures prominent associative knowledge about a concept possessed by humans. Silberer and Lapata (2012), for example, employ semantic feature norms as a proxy for human sensorimotor experiences in their semantic representation model, and report improved performance in word association and word similarity computation tasks. However, currently available semantic feature norms are, by nature, not well-structured, making them difficult to integrate into existing resources of various kinds.

Given this background, in this paper, we extract a tentative set of *psycholinguistically significant* semantic feature types, and draw a technical map to structurize corresponding semantic feature norms by observing the Linked Data paradigm. Note that psycholinguistically significant semantic feature types, in particular, dictate semantic relations that amply observe associations by humans; however, those are usually *not* considered in existing lexico-ontological resources.

Semantic feature	BR Label
a_reptile	*taxonomic*
beh_-_eats_people	*visual-motion*
beh_-_swims	*visual-motion*
has_a_mouth	*visual-form_and_surface*
has_jaws	*visual-form_and_surface*
has_scales	*visual-form_and_surface*
is_dangerous	*encyclopaedic*
is_long	*visual-form_and_surface*
lives_in_swamps	*encyclopaedic*

Table 1: Semantic feature norms and the BR Labels for describing alligator.

2 Semantic Feature Norms

2.1 Overview of McRae's Database

In this paper, we take the well-known set of semantic feature norms provided by McRae et al. (2005) (henceforth, McRae's database) as an actual example. This database provides a total of 7,526 semantic feature norms assigned to 541 living and nonliving basic-level concepts, each organized on the basis of experimental data collected from a large number of participants. McRae's database also presents a range of supplementary information, including statistical data about the semantic features.

Table 1 displays some of the semantic feature norms given to describe alligator. Although not fully shown in the table, more than ten features are used to describe several aspects of alligator. In Table 1, Brain Region (BR) Labels are also shown, each of which roughly classifies semantic features from the perspective

of brain function localization (Cree and McRae, 2003). See Appendix-A for more details.

2.2 Semantic Feature Keywords

As exemplified in Table 1, all of the semantic features are prefixed by predefined keywords or key phrases (e.g., `beh_-` in "`beh_-` eats_people"; "`lives_in` swamps"). These keywords and key phrases (henceforth, semantic-feature keywords) can be utilized to classify semantic features into basic types.

Semantic-feature keyword	# of variations
used_for	469
has	257
is	247
has_a	192
a	139
beh_-	138
used_by	113
made_of	70
requires	66
inbeh_-	64
lives_in	57
found_in	52
associated_with	44
worn_for	43
eg_-	40

Table 2: Productive semantic-feature keywords.

Although McRae et al. (2005) described around twenty semantic-feature keywords, the database actually classifies almost one hundred semantic-feature keywords, including presumably erroneous ones. Table 2 lists fifteen of the most productive semantic-feature keywords, in the sense of how many variations they have in the semantic feature norm instances. Most of the semantic feature keywords are self-descriptive; however, note that `beh_-` signifies behavior exhibited by animate beings (e.g., "`alligator beh_-` eats_people"), while `inbeh_-` denotes that an inanimate being does something seemingly on its own (e.g., "`airplane inbeh_-` crashes").

3 Structurizing Semantic Feature Norms

Figure 1, which corresponds to the `alligator` example shown in Table 1, illustrates a fundamental method of structurizing the semantic feature norms in McRae's database into a Linked Data graph[1]. The graph is constructed as follows: (1) A subject node is created for the target concept; (2) the subject node is linked with a set of triple objects, each representing a semantic feature; (3) a residual feature expression[2] is analyzed where necessary; and (4) each of the triple predicates carries a corresponding semantic feature type. In addition, the constructs of the graph should be linked with existing external Linked Data constructs whenever possible. In Fig. 1, word nodes are assumed to be linked with corresponding WordNet synset nodes by semantically disambiguating them. We may further need to resolve named entities, if we are to link them, for example, with DBPedia nodes.

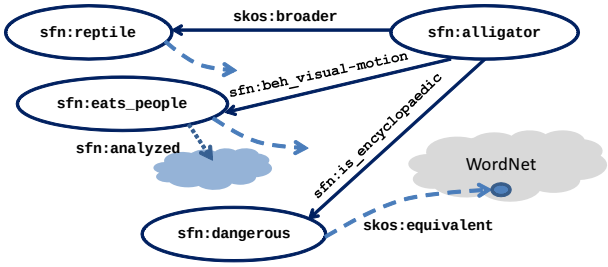

Figure 1: Linked Data graph structurizing a set of semantic features.

To actualize this illustration, we first need to create an inventory of triple predicates by identifying a reasonable set of semantic feature types, and then derive the sub-types where necessary.

4 Case Studies

We conducted our investigations by first extracting the tentative set of psycholinguistically significant semantic feature types shown in Table 3 from the ones already listed in Table 2 by performing the following actions:

- Excluding semantic feature types thought to be typical ontological constructs: these include, hyponymy (`a`), meronymy (`has_a, made_of, part_of`), telic/functional (`used_for, used_by`), exemplary (`eg_-`), causal (`causes`), and their subtypes (e.g., `worn_for`).
- Putting off semantic feature types whose semantics are clear and relatively restricted, such as `lives_in` and `found_in`, which both specify concrete/abstract places.

[1] In this paper, `sfn` denotes an imaginary prefix for representing constructs of a Linked Data graph. A more detailed modeling example using *lemon* (McCrae et al., 2010) is shown in Appendix-B.

[2] A *residual feature expression* denotes the natural language expression that follows a semantic-feature keyword: for example, "eats people" in "`alligator beh_-` eats people."

Semantic feature type	Example feature expressions
associated_with	cape associated_with Batman
is	apple is crunchy
requires	bread requires baking
beh_ -	alligator beh_- eats_people
inbeh_-	airplane inbeh_- crashes

Table 3: Psycholinguistically significant semantic feature types (tentative).

BR Label	Token frequency
visual-form_and_surface	546
visual-color	350
encyclopaedic	111
tactile	238
function	108
visual-motion	40
sound	34
smell	20

Table 4: Distribution of BR Labels for is.

The following subsections examine these nominated semantic feature types in turn.

4.1 associated_with

The "associated_with" semantic feature type associates a target concept with something associated with it, without specifying any particular semantic restrictions. The fact that all of the instances are labeled with *encyclopaedic* BR Labels endorses this action. Furthermore, this semantic-feature keyword exhibits a very high type/token ratio (TTR) of 0.96, asserting that an associated object is highly specific to the target concept, as exemplified by the "Batman" example shown in Table 3. Recall here that a type refers to a distinct semantic feature expression (word/phrase) succeeding a semantic-feature keyword, while a token dictates an occurrence of a semantic feature expression type.

The only thing we can do to structurize this semantic feature type is introduce a triple predicate such as, sfn:associated_with, as asserted in the above discussion.

4.2 is

The "is" semantic feature type in essence dictates several aspects/characteristics of a target concept from a variety of perspectives. In contrast to associated_with, this semantic feature type computed a very low TTR of 0.15: where the number of feature expression types was 247, while that of tokens amounted to 1,651. This situation forced us to further classify the feature expression types.

Here, we propose to classify this semantic feature type into a subclass by referring to the BR Labels. For example, by introducing the corresponding BR Label, "alligator is long" can be triplized as follows:

```
sfn:alligator
    sfn:is_visual-form_and_surface
sfn:long .
```

Table 4 summarizes the distribution of BR Labels for the is semantic feature type, where all but *function* and *encyclopaedic* are perceptual categories.

4.3 requires

The "requires" semantic feature type primarily specifies a typical object or entity that is somehow required by a nonliving target concept[3]. In contrast to the is semantic feature type, we cannot introduce BR Labels to further classify this semantic feature type into a subclass, as many of them (80/93 = 86.0%) are annotated with *encyclopaedic*, and the rest with *function*.

Therefore, we decided to investigate the semantic types of the *required* things by ourselves, and induced a set of sub-categories to combine with requires. Table 5 lists the sub-categories and the corresponding instance frequencies. Note that we in essence adopted semantic criteria from the Princeton WordNet for distinguishing physical/abstract entities: We however added *human* and *operation* to adequately classify the required things. With this in mind, "bread requires baking," for example, can be triplized as follows:

```
sfn:bread
    sfn:requires_operation
sfn:baking .
```

4.4 beh_-/inbeh_-

The "beh-" and "inbeh-" semantic feature types should intrinsically be considered *meta* feature types, only signaling typical or salient behavior/movement described in the residual feature expression, as seen in the examples introduced above: "alligator beh_- eats people" and "airplane inbeh_- crashes." Furthermore, as each of these expressions, in general, form a verb phrase, we would need to linguistically analyze the verb phrase to extract its semantic content.

[3]We observed 93 instances of the requires type in McRae's database, of which only two described living things.

Semantic type	Token frequency	Example feature expression
physical entity	55	`balloon requires` helium
human	19	`bus requires` driver
operation	13	`bread requires` baking
abstract entity	6	`unicycle requires` balance

Table 5: Semantic types of *required* things.

Types	*encyclopedia*	*sound*	*visual-motion*
`beh_-`	95	56	267
`inbeh_-`	33	50	32

Table 6: Distribution of BR Labels for `beh`/`inbeh`.

Further specification of such a linguistic analysis and the representation of the analysis results, however, are beyond the scope of this paper. We here focus instead on the sub-typing of these semantic feature types. As done earlier, we first checked the TTRs: `beh_-` computed 0.33, while `inbeh_-` exhibited 0.55, showing that some of the semantic-feature expression types are moderately productive. We then checked the distribution of the BR Labels, shown in Table 6[4]. The table clearly shows that only a few BR Labels are actually employed. Therefore, we decided to combine the BR Labels with these meta semantic feature types. Following this rationale, "`alligator beh_-_` eats people," for example, can be triplized as follows:

```
sfn:alligator
    sfn:beh_visual-motion
sfn:eats_people .
```

Intriguingly, while the majority of the behaviors taken by animate beings (`beh`-type) are classified as *visual-motion* (267/419 = 63.7%), the behaviors taken by inanimate beings (`inbeh`-type) are distributed across three categories: *encyclopaedic*, *sound*, and *visual-motion*, implying that the visibility of a behavior plays a psychologically prominent role in the characterization of living things.

5 Discussion

Psycholinguistic semantic features, in general, can improve the performance of semantic tasks in NLP, as demonstrated by Silberer and Lapata (2012). In other words, semantic features that are focused more on human perception should be combined with linguistic features. In this sense, migration of psycholinguistic semantic feature norms into a Linked Data cloud could provide an opportunity for a range of NLP applications to exploit psycholinguistic semantic features in combination with linguistic features acquirable from existing lexico-ontological resources.

The true benefits to be derived from publishing them as Linked Data, in particular, should be underpinned by concrete NLP applications. They are unfortunately not very clear at the moment, but the key to success is to employ the structurized set of psycholinguistic semantic features as a gateway to accessing existing resources of various kinds: including not only lexical/encyclopaedic resources such as WordNet, Wiktionary, and DBPedia, but also domain-specific ontologies such as GeoSpecies[5]. In this scenario, enabling proper linking with external resources is quite important.

Another crucial issue that has to be addressed in order to achieve the goal is the fact that the coverage of semantic feature norms needs to be significantly widened because currently available psycholinguistic resources, such as McRae's database, provide semantic features only for a limited number of concepts, notably, concrete concepts. Therefore, the development of a method to infer semantic features even for concepts not yet covered by existing resources (Johns and Jones, 2012) or, more importantly, a mechanism to mine useful properties from corpora (Baroni et al., 2010) would be highly appreciated.

6 Concluding Remarks

By examining the well-known McRae's database (McRae et al., 2005), we organized a reasonable set of psycholinguistically significant semantic feature types, and sketched a scenario for migrating them into the LDL.

For short-to-medium-term future work, we plan to (1) investigate other less-frequent/less-prominent semantic features observed in McRae's database; and (2) implement a computational process to actually convert the semantic feature norms into a set of Linked Data graphs.

[4] Labels with less than two occurrences have been omitted.

[5] `http://lod.geospecies.org/`

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 258201170.

References

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34:222–254.

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*, 132:163–201.

Bredan T. Johns and Michael N. Jones. 2012. Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, 4:103–120.

Oi Yee Kwong. 2012. *New Perspectives on Computational and Cognitive Strategies for Word Sense Disambiguation*, Springer.

John McCrae, et al. 2010. The *lemon* cookbook, `http://lexinfo.net/lemon-cookbook.pdf`

John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. 2012. Integrating WordNet and Wiktionary with *lemon*, In Chiristian Chiarcos et al. (eds.) *Linked Data in Linguistics*, Springer-Verlag, pp.25–29.

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

Allan Paivio, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns, *Journal of Experimental Psychology*, 76 (1, Part 2):1–25.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation, *Proceedings of the 2012 Joint Conference on EMNLP*, pp.1423–1433.

Sinya Tanaka, Adam Jatowt, Makoto P. Kato, and Katsumi Tanaka. 2013. Estimating content concreteness for finding comprehensible documents, *Proceedings of The Sixth ACM WSDM Conference*, pp.475–484.

Appendix-A: Brain Region Labels

Each of the BR Labels assigned to a semantic feature norm in the database is based on a taxonomy called *Brain Region Taxonomy* (Cree and McRae, 2003). Table A-1 classifies the nine (plus one: *taxonomic*) categories defined by the BR taxonomy, and the corresponding token frequencies in the database. Cree and McRae (2003) argue that these categories represent knowledge types that are closely associated with corresponding brain regions.

BR Label	Frequency
visual-form-and-surface	2,336
visual-color	424
visual-motion	339
tactile	245
sound	142
taste	84
smell	24
function	1,517
encyclopaedic	1,417
taxonomic	730

Table A-1: Distribution of the BR Labels.

As displayed in Table A-1, seven of the nine categories are linked with sensory channels/modes, of which three are associated with visual perception. In particular, the category *visual-form-and-surface* exhibits substantially high frequency, highlighting the fact that *visibility* plays a significant role in characterizing a concrete object psycholinguistically. The category *function*, on the other hand, organizes feature types, such as `used_for` and `used_by`, describing functional aspects of a target concept. Semantic features encoding other types of miscellaneous knowledge were labeled as *encyclopaedic*.

Appendix-B: Modeling with *lemon*

Figure B-1 exemplifies a more detailed modeling of the Linked Data graph presented in Fig. 1. In this modeling, McRae's entire database is modeled as a *lemon* lexicon. That is, every content word in McRae's database is modeled as a lexical entry, and the semantic feature types, derived in this paper, are modeled as sub-properties of `lemon:senseRelation`, which connects `lemon:sense` instances. In addition, linking to WordNet is represented by using `lemon:reference`, as in (McCrae et al., 2012), meaning that WordNet is treated as an external ontological resource.

Notice also that the residual semantic feature expression, such as "eats people," is modeled as a phrasal lexical entry, whose internal linguistic structure is meanwhile represented by a syntactic dependency structure, represented by the blue cloud in the figure.

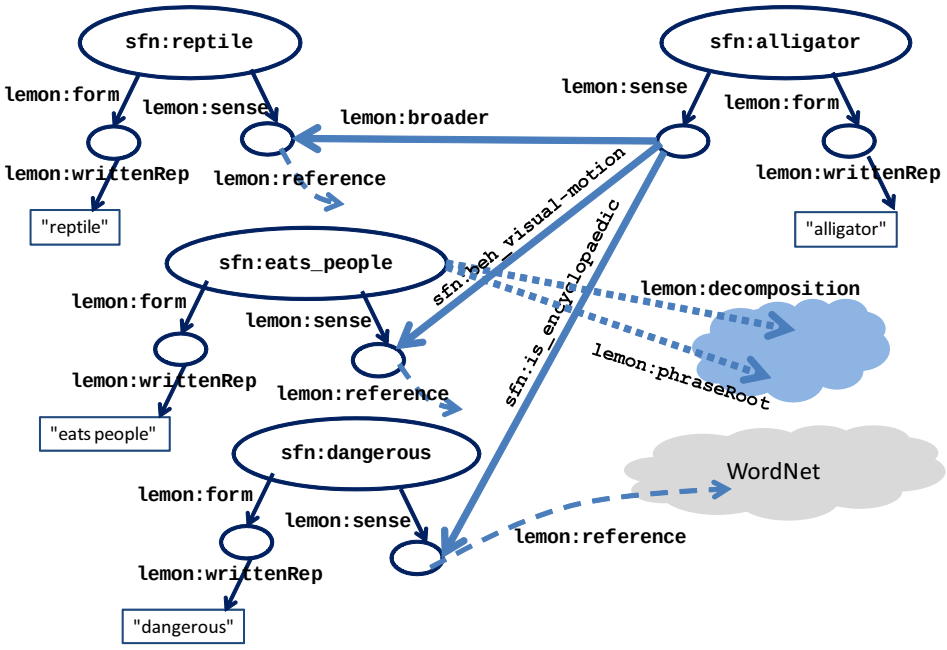

Figure B-1: Modeling using *lemon*.