When nouns need co-arguments
A case study of semantically unsaturated nouns

Kow KURODA Masaki MURATA Kentaro TORISAWA
National Institute of Information and Communications Technology (NICT), Japan
3-5 Hikari-dai, Seika-cho, Sooraku-gun, Kyoto, 619-, Japan

1 Introduction

Recently, we finished the manual cleaning of approximately 67,000 Japanese hypernym hierarchies paired with roughly 900,000 hyponyms. The original data, comprised of roughly 2,400,000 hypernym-hyponym pairs, were automatically acquired from the Japanese Wikipedia (Sumida et al., 2008). The data we processed resemble the following:

(1) a. $h_1$: singer, $h_2$: rock singer, $h_3$: British rock singer, $h_4$: famous British rock singer; $I$: Peter Gabriel
b. $h_1$: member, $h_2$: former member, $h_3$: former member of Floyd, $h_4$: former member of Pink Floyd; $I$: Syd Barrett

Pairs $(H; I)$, where $H = h_1, \ldots, h_n$, are automatically generated from such pairs $(h_{\text{max}}; I)$. We refer to $H$ as the “hypernym path” for $I$, and to units like $h_1, h_2, \ldots, h_n$ as the “path elements” of $H$.

A hypernym path may contain: (i) bare nouns (e.g., singer), (ii) modified nouns (e.g., famous British rock singer, former member), or (iii) noun phrases. Such a path is constructed by systematically removing modifiers from $h_n$ one by one. This operation is not error-free. Manual cleaning was performed to discard unacceptable units like former member of Floyd, $h_4$ of (1b) yielded in the automatic generation. It soon became apparent that checking for the conventionality alone was unworkable. We also needed a systematic treatment of units like (former) member, nouns that cannot be used independently and seem to require “arguments” at least semantically. Terms like singer make good hypernyms, but not terms like (former) member due to their semantic unsaturatedness. By contrast, (former) member of Pink Floyd creates good hypernyms; yet no shorter terms can be accepted as good hypernyms. Thus, semantically unsaturated nouns probably do not make good hypernyms. If correct, however, it poses another question: exactly what terms behave semantically unsaturated? The answer, which is far from obvious, motivated us to formulate and develop the descriptive model presented below.

The paper is organized as follows: In §2, we examine crucial cases of argument-taking nouns of (i) deverbal nouns, (ii) nondeverbal relational nouns, and (iii) nondeverbal quasi-relational nouns, and try to reinterpret their behavior in terms of their “frame-evoking” property, adopting a useful notion from FrameNet (Baker et al., 1998; Fontenelle, 2003). In the attempt, we introduce the notion of “co-arguments” that helps us to identify a class of “(semantically) unsaturated nouns” as a generalization over the relevant classes, and then we suggest a method for automatically measuring unsaturatedness of nouns using information about their “qualia structures” (Pustejovsky, 1995). In §3, we describe a few interesting properties of unsaturated nouns in this generalized sense.

---

1) This paper benefited from the comments by three anonymous reviewers. Any remaining errors, however, are the responsibility of the authors — mainly of the first.
2) The data we processed are special because nearly 2/3 of the hyponyms are proper names. These data are expected to complement such traditional thesauri as WordNet (Fellbaum, 1998) in which upper ontologies are specified. We previously linked the roots of the hypernym hierarchies thus cleaned to the leaves of WordNet-Ja (WN-Ja) (Bond et al., 2009). Currently, 70% of the roots of the hypernym hierarchies are linked to WN-Ja (the coverage was only 8% in the original $(h_n; I)$ pairs).
3) We worked on Japanese data, but we present the English counterparts in this paper.
4) For both practical and theoretical purposes, we did not distinguish “instance-of” and “subclass-of” relations.
5) We did not use Named Entity tagger to filter them out.
2 Description of Phenomena

2.1 Argument structures of nouns

2.1.1 Case of deverbal nouns

Let us begin with simpler cases like singer, inventor, creation, and invention. These are quite frequent deverbal hypernyms in the data: singer.n is derived from sing.v, inventor.n is from invent.v, and creation.n is derived from create.v. Due to their nature, these nouns clearly have argument structures, inherited from base verbs, which are realized in nominal forms in different ways:

(2) a. The inventor of WWW is Tim Berners-Lee.
   b. Linux is a creation { i. by; ii. of } Linus Torvalds.

2.1.2 Case of nondeverbal nouns

Are deverbal nouns the only nouns with argument structures? It seems not. Consider cases like au-

(3) a. The author of “Gargantua and Panta-gruel” is François Rabelais.
   b. “The Creation” is a masterpiece of Joseph Haydn.

The parallelism between the cases in (2) and (3) is obvious, but it is hard if not impossible to claim that nouns like the ones in (3) are derived form verbs. For example, author is not deverbal, even if ending with agentivity-marking morpheme -or.6)

2.1.3 Case of so-called “relational” nouns

The relevant property of nouns like author is related to the notion of relational nouns exemplified by sister, mother, and father which arguably have their own arguments, despite not being derived from any verbs. Many linguists have concentrated on such nouns of this class (de Bruin and Scha, 1988; Nishiyama, 1990; Partee and Borschev, 2003) mainly due to the peculiarities in their semantics. Let us cite a relevant definition of relational nouns:

(4) Relational nouns are semantically unsaturated. They are normally used in combination with an implicit or explicit argument: John’s brother. The argument of a relational noun, if overtly realized in the sentence, is connected to the nouns by means of a relation-denoting lexical element: the verb have or one of its semantic equivalents (the genitive and the prepositions of and with): John has a sister, John’s sister, a sister of John’s, a boy with a sister.

The definition of relational nouns cited from de Bruin and Scha (1988) represents the notion of them widely accepted in linguistics.

A similar notion is adopted in cognitive psychology (Gentner, 2005; Gentner and Kurtz, 2005), but there seems to be a subtle difference. Most linguists would be reluctant to admit that masterpiece, letter, victim, part, member, and (vice-)president, are relational nouns, yet cognitive psychologists would not; they do not seem to hesitate to consider them as relational nouns, hypothesizing that relational nouns are any nouns that denote (components of) relational categories (Gentner, 2005; Gentner and Kurtz, 2005). It is not trivial to tell what nouns count as relational and what nouns as not.

Let us pose a crucial question, Is “former member” in (1b) a relational noun? It is likely, but confirmation is not trivial. Differently put, it is not trivial to differentiate true relational nouns from “quasi-relational” nouns such as (former) member, masterpiece and letter. The reason is two-fold: First, former adds unsaturatedness to member, which itself is unsaturated, suggesting former has unsaturatedness of its own. It is not clear at all how they interact, however. Second, as far as we know, nobody has ever tried to exhaustively list relational nouns for the lack of reliable criteria to identify relational nouns.

Under this, this paper proposes a set of prerequisites for effective identification of relational nouns. Before starting it, however, let us first specify the conceptual problem that needs be solved. If non-deverbal nouns including relational nouns can have argument structures, what nouns cannot? Does masterpiece in (3), for example, have an argument structure or not? The answer is not obvious. More specifically, are all nouns with argument structures relational nouns? To answer this successfully, we need to extend the notion of relational nouns to (semantically) unsaturated nouns.

2.1.4 Co-argument structures of nouns

Note that nouns like inventor, invention, author, and masterpiece are either names for frame elements (FEs) in the sense of Berkeley FrameNet...
2.2 Semantically unsaturated nouns

Based on the set of issues described so far, we define the “semantically unsaturated nouns” as a cover term for the relevant classes of nouns:

(5) (Semantically) unsaturated nouns\(^{11}\) are nouns that has either (proper) arguments or co-arguments. This class of nouns include: (i) nouns derived from predicates such as verbs and adjectivals, (ii) relational nouns, and (iii) (quasi-relational) nouns with (strong) co-arguments.

The definition in (5) is rather informal. We will give a more formal definition under Formal Concept Analysis (FCA) (Ganter et al., 2005) in §3. But let us address an unsolved issue before proceeding. Note that the definition in (5) relies on the assumption that there is a procedure with which we can identify a given noun’s arguments and co-arguments. Admittedly, effective identification of co-arguments is hard to achieve. Let us specify our idea for it, though it is not implemented yet.

2.2.1 Identifying co-arguments through qualia structure

A closer examination of nouns with “strong” co-arguments (e.g., masterpiece and letter) suggests that they tend to have “clear” and “distinct” qualia structures (Pustejovsky, 1995). We say that nouns have clear qualia structures if they have well-defined purposes (clear and determined telic roles) and well-understood origins (clear and determined agentive roles).

It is not a coincident that clearly unsaturated nouns tend to have distinct qualia structures, although it is unclear if the aspects of unsaturatedness are all derivatives of the qualia structure. But the contrary does not seem to be true: Beer, for example, has a clear qualia structure that defines brewer or brewery bearing agentive role, but it is not clear whether this implies that beer is an unsaturated noun. For evidence, the by-phrase in beer by Budweiser is more likely to be treated as a modifier than as an argument. But the notion of co-argument allows us to say that by-phrase is a co-argument of beer.

This implies that so-called “modifiers” on a noun \(N\) fall into three types: (i) \(N\)’s arguments, \(N\)’s co-arguments (often licensed via qualia structure) and (iii) \(N\)’s pure modifiers. Typical example of pure modifiers are specifiers of time and place. This provides a new insight into the classification of modifiers.

\(^{11}\)Note that the term “unsaturated nouns” was first coined by Yuji Nishiyama in his work (Nishiyama, 1990) to characterize the behavior of Japanese relational nouns in the sense of de Bruin and Schä (1988) presented in (4). For the lack of a better term, we redefined Nishiyama's term to denote a larger set of nouns and hope this does not mar his contribution to Japanese theoretical linguistics.
Based on this typology of modifiers, we hypothesized as follows: the clearer qualia structure a noun has, the stronger its co-arguments are (smaller variance in types of qualia structure contributes to better identification). If this hypothesis is correct, it follows that we will be able to measure unsaturatedness automatically if we have resources that enable us to separate nouns with stronger co-arguments from ones with weaker ones. Resources of the relevant kinds are being constructed through recent advances in automatic acquisition of qualia structures (Cimiano and Wenderoth, 2005; Torisawa, 2005; Yamada et al., 2007) and case frames of nouns (Sasano et al., 2004; Murata et al., 1999). This line of investigation is left for future work.

3 Behavior of unsaturated nouns

3.1 FCA of Japanese predicates

We exploited FCA to classify saturated and unsaturated nouns and verbs of Japanese, trying to capture the complex relationship for requirements of semantic (co-)arguments. The result is presented in Fig. 1. We expect the result to be applicable to languages other than Japanese.

The following are the attributes used as input to FCA to yield the lattice in diagram 1:

(6) a. [takes-more-than-1-arg: ±] (Column B of Table 2)
b. [mark-arg-differently: ±] (Column C)
c. [takes-no-arg: ±] (Column D)
d. [is-a-verb: ±] (Column E)
e. [derived-from-V: ±] (Column F)
f. [allow-more-than-2-args: ±] (Column G)
g. [denote-individual: ±] (Column H)
h. [takes-only-2-args: ±] (Column I)
i. [inflects: ±] (Column J)
j. [needs-support-to-inflect: ±] (Column K)
k. [needs-aux-to-inflect: ±] (Column L)
l. [is-a-noun: ±] (Column M)

The crucial point here is that [is-a-noun(\(x\))] is independent of [takes-no-arg(\(x\))]: in other words, the arity of a predicate is independent of lexical types. This poses a new question about a noun’s definition, suggesting that a noun is a distributional category rather than a conceptual one. This
is the most straightforward explanation of why nouns can have arguments and co-arguments.

3.1.1 Note on Japanese verb morphology
As mentioned above, we do expect the classification to be naturally applicable to languages other than Japanese, but there is a peculiar fact about Japanese morphosyntax that complicates it. We take a brief look into it.

In the analysis in Fig. 1, noun X is called a “SA-HEN noun” (サ変名詞 in Japanese) if X is comprised of a verb of the form X-suru, with no case marker intervening between X and suru, a kind of light verb.\(^{12}\) Let us explain the relevant facts using an example. In (7), nominative NP sakusha-ga is marked as (Creator), and accusative NP sakuhin-wo as (Created_entity) of (Creating).

(7) Clausal form:
Sakusha-ga sakuhin-wo sousaku-suru.
author-NOM work-ACC create-does
“An author creates (a piece of) work.”

(8) Nominal(ized) forms of (7) meaning “creation of work by an author . . . ”

a. Sakusha-no sakuhin-no sousaku(-ga . . . ).
author-OF work-OF creation(-NOM . . . )

b. Sakusha-nyoru sakuhin-no sousaku(-ga . . . ).
bauthor-BY work-OF creation(-NOM . . . )

12) Kageyama (1993) called X a “verbal noun.” This terminology is quite influential, both in and out of Japan, in the tradition of Generative Grammar (GG), but we decided not to use it because it is less accepted outside the narrow circle of GG; most researchers in Natural Language Processing (NLP) of Japanese, who know what SA-HEN nouns are, would not know what verbal nouns are.

Also, the proper characterization of SA-HEN nouns remains somewhat controversial. It is not clear whether they are zero-derived from corresponding verbs that have -suru underlyingly. Our position is that those nouns are underlyingly underspecified for the verb/noun distinction, and -suru is required when overt inflection is needed. This analysis is compatible with the lattice diagrammed in Fig. 1.

In this case, \(X = \text{sousaku} \). In (7), \(\text{sousaku-suru} \) is a verb. In (8), \(\text{sousaku} \) is an unsaturated noun that require arguments semantically, sakusha and sakuhin, marked for adjuncts (by niyoru = “by,” no = “of”). (9) illustrates the English counterparts of the Japanese classes of (a) deverbals, (b) SA-HEN nouns, and (c) non-deverbal, non-SA-HEN nouns:

(9) a. deverbal nouns (A): invention.n, inventor.n, inventory.n (from invent.v), creation.n, creator.n (from create.v)
b. SA-HEN nouns (B):\(^{13}\) walk.n (= walk.v), move.n (= move.v)
c. non-deverbal, non-SA-HEN nouns (C): (counter)measure.n (in devise a (counter)measure against), antipathy.n (*antipathize.v; cf. sympathy.n, sympathize.v)

3.1.2 Support verb constructions in Japanese
In Japanese, the syntax and semantics of nouns of type A and B nouns are regular, but this is not true of type C nouns: taisaku.n (“countermeasure”), hankan.n (“antipathy”) and kugen.n (“expostulation”) are of type C. They have the same kind of argument structures as type C nouns, which differ from type B nouns because they cannot be used as verbs by adding -suru: *taisaku-suru, *hankan-suru, and *kugen-suru are all unacceptable. Instead, they form support verb constructions, as shown in the following examples:

(10) a. X-ga Y’s { i. ni; ii. he(no) } taisaku-wo koujiru (“X devise a (counter)measure against Y”)

b. X-ga Y-ni(taishite) hankan-wo { i. motsu; ii. idaku; iii. kanjiru } (“X { i. have; ii. feel } antipathize against Y”)

13) Nouns like song and product could be classified into this.
where *taisaku*, *hankan* and *kugen* are used as accusative nouns (marked by *-wo*), and *kouji-ru*.*v* ("devise"), *motsu*.*v* ("have"), *idaku*.*v* ("hold"), *kanji-ru*.*v* ("feel") and *tei-suru*.*v* ("show") are support verbs for them.\(^{14}\)

### 3.2 Findings and Motivations

This section addresses a few important findings in the cleaning task that motivated the classification.

#### 3.2.1 Modifiers as sources of unsaturatedness

It is tempting to assume that unsaturatedness is a lexical property. But the reality of the data we processed suggests otherwise. We have seen many cases in which compound nouns acquire unsaturatedness by modifiers. For example, *city* is arguably a saturated noun, but *sister city* is unsaturated because *sister* adds unsaturatedness to it.

Note also that unsaturatedness accumulates through modification. Take "*X is a fellow disciple of Y*" for example. In this case, we can infer that both *X* and *Y* are *disciples* of *Z*, the same master. Interestingly, the unsaturatedness for *fellow* and *disciple* can co-exist, though the surface realization of *Z* seems to be suppressed.\(^{15}\) Note that *former* in (1b) behaves exactly like *fellow* here.

Unsaturatedness is composable, allowing the unsaturatedness of one noun to get “bridged” to another. In cases like *secretary of the Minister of Foreign Affairs*, unsaturatedness is reduced through variable-binding, because *secretary of X*, *X* is bound to *the Minister of Y*, and when *Y* is bound to *Foreign Affairs* (with the aid of *of*), it gets saturated; otherwise, it stays unsaturated.

#### 3.2.2 Unsaturated nouns are not rare

In this regards, it deserves a mention that linguists seem to tacitly assume that relational nouns are rare and exceptional, and that their set is closed. This assumption is far from well grounded because the sets of relational nouns and of nouns with strong co-argument structures are often indistinguishable.

We state this based on our experience in the cleaning task explained in the Introduction. The task was performed based on classification guidelines that incorporated the definition of unsaturated nouns presented in (5). As a result, we obtained roughly 118,000 types of manually rated path elements. The ratios of saturated, unsaturated, dubious, and discarded elements were 57.4% \(21.2\%\), 7.2%, and 14.2%, respectively. The net ratio of saturated against unsaturated was 73.0% against 27.0%, suggesting that nearly 1/5 of noun types are unsaturated. This does not seem to support the traditional view of unsaturated nouns.

#### 3.2.3 Degrees of unsaturatedness

The classification guidelines we prepared contain some heuristics to deal with the degrees of unsaturatedness that confused annotators. For example, the three Japanese nouns in (11) turned out to have different strengths of unsaturatedness and can be ordered in a specified way:

\[(11)\quad (X(sha)-no)\ sha-cho
\]

("president (of company) X"), *sha* means "company"

< (X(bu)-no) bu-cho

("manager (of section X)"

< (X(kyoku)-no) kyoku-cho

("director/chief (of department X)"

Literally, (X-no) *cho* means “person of prime importance/head (of X)."

What is puzzling is that *shacho, bucho* and *kyokucho*, all of which are relational and should therefore be equally unsaturated, show different degrees of unsaturatedness after human rating.

On a closer examination, the degrees of unsaturatedness, at least as measured through human judgments, correlated with distributional statistics: i) nouns with weaker word boundaries at the onset were judged to be more unsaturated; ii) nouns with stronger word boundaries at the onset were judged saturated even if they are theoretically unsaturated nouns: (X(sha)-no) *sha-cho* is one of such nouns.

Note that *shacho, bucho* and *kyokucho* have different degrees of “wordhood.” Three statistics are relevant: First, the frequency order of the three terms is: *shacho* \(\gg\) *bucho* > *kyokucho*. It is likely that high frequency reduces the (feeling of) unsaturatedness of a term. Second, letting *Y*={ *shacho, bucho, kyokucho* }, the varieties of *L* in the

---

\(^{14}\)In most constructions, support verbs are selected in complex ways influenced by subtle collocations, and achieving prediction with high precision is hard about what verbs can serve as support verbs for what type C nouns. This is probably true of similar cases in other languages.

\(^{15}\)Note here that a mutuality interpretation of relational nouns (Eschenbach, 1993) seems to have an interesting effect on the construction and interpretation of *sister city(ies)* and *fellow disciple(s)*.
context \(LYR\) follows the order of \(shacho \gg bucho \gg kyokuch\), irrespective of whether \(L\) ends with a case marker or not. Third, the relative frequencies of genitive marker \(-no\) against other markers follows the order of \(shacho \gg bucho \gg kyokuch\), given the context \(XMYR\) where \(X\) is a noun marked by marker \(M\) such as \(-no, \-ga, \-de,\) and \(-wo\). The second and third are natural consequences of unsaturation. This observation suggests that \(shacho\) has the strongest independence.\(^{16}\)

Taking these properties into account, we hypothesized that potentially unsaturated nouns were sometimes wrongly classified as saturated nouns when they were frequent. Based on this, human raters were advised not to follow candid intuitive judgments during classification but to scrutinize the semantics of target nouns combined with potential arguments or co-arguments.

### 3.2.4 Measuring unsaturatedness

As an extension of this theorizing, we derived two different heuristic measures to differentiate unsaturated nouns. They are based on different heuristic measures to differentiate unsaturated nouns and smaller for saturated nouns, with other things being equal.

We originally planned to test the validity of the heuristics, but no further details are reported in this paper, since it is incomplete for the time being. But if either turns out to be valid, it would offer a quantitative definition of unsaturatedness, as we briefly touched on in §3.2.1.4

### 3.3 Benefits of co-argument notion

Several projects (Meyers et al., 2004; Iida et al., 2007), some of which are ongoing, have investigated the annotation of argument-taking nouns in general domains. The typical annotation task marks the arguments of deverbal nouns like inventor and invention but it includes annotations for nondeverbal nouns like masterpiece and letter.\(^{18}\) Scrutiny of annotation guidelines (Meyers, 2007) reveals that no distinction is made between the proper arguments of deverbal nouns and the co-arguments of nondeverbal nouns. This does not look optimal.

Under the notion of co-arguments, letter, for example, can be marked for information of the kinds specified in (12),\(^{19}\) but such labels as \(\langle\text{Author}\rangle, \langle\text{Addressee}\rangle, \langle\text{Medium}\rangle,\) and \(\langle\text{Content}\rangle\) identify co-arguments rather than proper arguments:

\[
\text{(12) … \langle\text{Addressee: he}\rangle received a \langle\text{Medium: letter}\rangle from \langle\text{Author: Nissan Motor}\rangle stating \langle\text{Content: that the company had a concern over Nissan Computer’s use of that domain name}\rangle.} ^{20}
\]

For some readers, the distinction of co-arguments from proper arguments probably resembles a mere terminological difference. But it is not, at least if we consider the qualia structure of a noun for its relevance. Given that write(e, \(x, y, z\)) and inform(e, \(x, y, w\)) are a letter’s agentive and telic components, respectively, \(x=\text{Nissan}\)
4 Conclusion

This paper proposed the notions of a noun’s co-arguments differentiated from proper arguments and of semantically unsaturated nouns. We discussed some benefits from them. We hope they contribute to better understanding of nouns by unifying theories of semantic frames and qualia structures.

References


