Semantic relations in binominal lexemes: 
A cross-linguistic survey

Steve Pepper

A key feature of binominal lexemes is the unstated (or underspecified) relation, ℜ, that pertains between the two major constituents. Understanding the nature of ℜ – the kinds of relations and their frequency – has the potential to provide insights into the nature of concept combination and the associative character of human thought. In this chapter I approach the question from a cross-linguistic, onomasiological perspective. I develop a two-tier classification called the Hatcher-Bourque system based on work by earlier researchers. The classification operates at two levels of granularity, with 29 and five relations respectively. The classification is applied to a data set consisting of over 3,700 binominal lexemes from 106 languages. This permits a statistical analysis of the frequency and distribution of semantic relations at both an highly abstract level and a more granular level. The Hatcher-Bourque classification, and an accompanying tool called the Bourquifier, are offered to the research community in order to encourage more collaboration than has hitherto been the case in the field of semantic relations.¹

1 Introduction

The unstated (or underspecified) semantic relation, ℜ, is a defining feature of both binominal lexemes in general and the canonical subtype, noun-noun compounds, the latter defined here as binominals that use either the cmp or the jxt morphosyntactic strategy (see Introduction). Guevara & Scalise (2009: 107) suggest that the “inner essence” of a compound can be captured (in the prototypical case) with a “rough schema” (1), where X, Y and Z represent “major lexical categories” (in the case of binominals, these are, of course, nouns or other thing-morphs), and ℜ “represents an implicit relationship between the constituents (a relationship not spelled out by any lexical item)”.

(1) \([ X_N \, ℜ \, Y_N \, ] \, Z_N\)

Jackendoff (2016) provides a nice set of examples (2) to show that the kind of semantic relation can be “hugely varied”, even across compounds that share a common hyponym, such as cake.

¹ This chapter is based on my doctoral dissertation (Pepper 2020).
In N+N compounds there is no indication of the nature of the semantic relation, and the same applies to the types der and els that also belong to Level 0 in the formal typology. In types that belong to Levels 1 and 2 the semantic relation is underspecified rather than unstated: the presence of one or more additional morphemes indicates the presence of some kind of relation. However, the meaning inherent in those relational morphemes is extremely schematic. The fact that $R$ is not stated explicitly gives rise to at least two questions (Bauer 2017):

1. How can these semantic relationships best be classified?
2. What mechanism allows compounds to be generated (and understood) when the semantics is so variable?

The second of these questions is the focus of Hacken (2016), in which the process of compounding is examined in the context of three theoretical frameworks, viz. Jackendoff’s Parallel Architecture, Lieber’s Lexical Semantics and Štekauer’s Onomasiological theory. The present chapter concentrates on the first of the questions posed above and investigates the kinds of semantic relation found in binominals. To that end, the next section (§2) provides a brief overview of previous work on this topic. While most of this relates to noun-noun compounds, there is reason to believe that the same approach can be applied to other types of binominal, including denominal derivations.

Following this overview, §3 describes the onomasiological approach taken in the present study and the language sample. Then, in §4 I present the two classification schemes on which the present study is based: the one, rather granular, developed by Yves Bourque, the other, more schematic, by Anna Granville Hatcher and I show how the former can be mapped to the latter, resulting in a combined classification at two levels of granularity. §5 and §6 contain the meat of the chapter: a statistical analysis of the kinds of semantic relation found in the data – and §7 offers a summary and ideas for further research.

---

2 The cognitive approach (or more precisely, that of Cognitive Grammar) to word-formation is described by Tuggy (2005), and psycholinguistic aspects are explored by Libben & Jarema (2006).
2 A brief history

The history of research into semantic relations in compounding tends to be rehearsed by every researcher entering the field (e.g. Ryder 1994; Pepper 2010; Szubert 2012; Bourque 2014; Eiesland 2016; Toquero 2018). The reader is referred to these and similar works for more historical details. The aim of the present section is simply to present a brief overview and prepare the ground for a more detailed survey of work that is particularly relevant to my own work.

Interest in semantic relations can be traced at least as far back as Grimm (1826) on compounding in German, and Mätzner (1860) on English, but the starting point for modern studies of the topic is usually taken to be Jespersen (1942). This was followed by a large number of influential and widely cited studies, notably Hatcher (1960), Lees (1960; 1970), Marchand (1960; 1969), Brekle (1970), Li (1971), Adams (1973), Downing (1977), Allen (1978), Bauer (1978), Levi (1978), Warren (1978), Ryder (1994) and Jackendoff (2010). A useful way to gain an overview of the various approaches taken in these and other works is the classification of compounding theories suggested by Søgaard (2005), who distinguishes between transformational, slot-filler, reductionist, and pragmatic theories.

Transformational approaches attempt to derive semantic relations from underlying syntactic structures such as relative clauses. The most well-known studies of this type are Lees (1960; 1970) and Levi (1978), which are situated within the frameworks of early generative theory and Generative Semantics, respectively. This line of research was abandoned after the 1970s (Hacken 2009), but while the underlying theories are no longer regarded as viable, the actual results achieved, especially by Levi, have had a lasting effect.

In slot-filler theories of compounding constituents are seen as bundles of features with the modifying constituent supplying a value to one of the features of the head. Søgaard refers in particular to Johnston & Busa’s (1999) work within the framework of the Generative Lexicon (Pustejovsky 1995), but the slot-filler approach applies equally to Allen’s (1978) work in the lexicalist framework and that of Rochelle Lieber in lexical semantic analysis (Lieber 2009; 2016).

Most work on semantic relations, however, has followed what Søgaard calls the reductionist approach, in which the researcher attempts to enumerate a limited number of “primitive relationships”. These vary in number from four (in the case of Hatcher 1960) to well over 100, depending on whether one includes subtypes, i.e. the degree of granularity of the analysis. Bourque (2014: 167) provides a table listing 16 studies of this type and the number of relations that they posit. The following list offers a representative selection of such studies within a variety of theoretical frameworks:

- Jespersen (1942): Eight types of NN compound AB in which B is modified by A, along with the claim that “the number of possible logical relations … is endless”.
• **Hatcher (1960):** Reduces seven of Jespersen’s types and two of Mätzner’s to four “logical” types.

• **Levi (1978):** Nine “recoverably deletable predicates” (RDPs), of which three are reversible, for a total of 12.

• **Warren (1978):** 12 “semantic classes”, most of them named according to the role played by the constituents in the relation.

• **Ryder (1994):** 50 templates representing compound schemas that emerged from her psycholinguistic experiment.

• **Jackendoff (2016):** 13 “basic functions”, six of them reversible.

When Jespersen declared the set of possible semantic relations to be open-ended, he was echoing sentiments expressed earlier by Carr:

> Although an attempt may be made to classify the compounds from a semantic point of view, it would be impossible to state all the relationships which do occur, and to assign each compound to a particular class (Carr 1939: 319–320).

At the other end of the spectrum we find the fourth kind of theory of compounding identified by Søgaard, which he calls *pragmatic* theories. These claim that there is really only one, very general relationship between the constituents of a compound, and that the compound’s meaning is derived solely from pragmatic knowledge about the world. This position was advanced by Bauer (1979) and was couched in terms of the predicate deletion approach then current (cf. Levi 1978):

> I suggest that only one ‘verb’ (or more accurately ‘pro-verb’ since it is an abstract unit) should be deleted in the generation of compounds. If only one verb is to account for the range of semantic relations that exist between the two elements of compounds it will have to be very abstract and have a vague meaning. I suggest a gloss something like ‘there is a connection between’.

All of the above-mentioned works focus exclusively on compounding in English, with a couple of exceptions that investigate German (or early Germanic) and Chinese. In recent years, however, the topic of semantic relations has been explored in other languages, including Nizaa (Pepper 2010), Danish (Szubert 2012), French (Arnaud 2003; 2016; Bourque 2014), Norwegian (Eiesland 2016) and Spanish (Toquero 2018). It has also received a lot of attention in computational and corpus linguistics (e.g. Vanderwende 1994; Moldovan et al. 2004; Girju et al. 2005; Ó Séaghdha 2008; Tratz & Hovy 2010; Nakov 2013; Schäfer 2018) and was even the focus of an NAACL-HLT Workshop on Semantic Evaluations task on “the interpretation of noun compounds using paraphrasing verbs and prepositions” (Butnariu et al. 2009).

The position taken in the present study is that the number of relations one identifies will be a function of the degree of granularity one requires them to express. It can therefore be anything the researcher desires, from one (as suggested by Bauer) to unlimited (as opined by Carr and Jespersen). Furthermore, even if the total number turns out to be unbounded, “the vast majority”, as Tratz & Hovy (2010) point out, will fit within
“a relatively small set of categories.” As we shall see, in the present study I actually apply two classifications: the one fairly granular, consisting of 29 relations, the other much more abstract and consisting of just five relations. However, before presenting these I describe the research design of the present study.

3 The present study

Unlike previous studies, the present study focuses on a broad range of languages and adopts an onomasiological approach. The starting point is a set of 100 meanings (3) that were selected in a principled manner in order to give a maximum yield of binominals of the determinative rather than the coordinative type.

(3) ankle, arctic lights, backpack, bee, beehive, beeswax, bicycle, bicycle pump, blacksmith, boy, bracelet, breakfast, carpenter, chieftain, cock/rooster, collarbone, cookhouse, dairy cow, dinner, doorpost, earlobe, earring, earwax, eyebrow, eyelash, eyelid, farmer, fireplace, fisherman, fishing line, flame, flea market, foal or colt, footprint, girl, glove, gold ring, hand brake, handbag, handkerchief or rag, herdsman, hospital, host, keyword, kid, lamb, license plate, lunch, magic, mail box, mare, married woman, midday, milky way, mother-in-law (of a man), native country, neighbour, niece, nipple or teat, nostril, paddle wheel, palm of hand, postcard, potter, queen, railway, rainbow, rib, shoemaker, shore, shoulderblade, skull, sorcerer or witch, spectacles/glasses, spider web, spine, stable or stall, stone bridge, sugar cane, Sunday, supper, tear, thatch, thumb, toe, toilet, toilet paper, tool, toolbox, toothbrush, train, tree trunk, vein or artery, vine, water pump, Wednesday, widower, windmill, wrist, yolk

The meanings were assigned to the following 16 semantic fields using the scheme of the World Loanword Database (WOLD; Haspelmath & Tadmor 2009):

(4) Agriculture and vegetation, Animals, Basic actions and technology, Clothing and grooming, Emotions and values, Food and drink, Kinship, Modern world, Motion, Religion and belief, Social and political relations, The body, The house, The physical world, Time, Warfare and hunting

They were also classified into one of the following seven semantic types:

(5) Person, Body part, Animal, Location, Natural phenomenon, Basic technology (or concept), Advanced technology (or concept)

Both semantic field and semantic type will be explored in the analyses of distribution and frequency in §5 and §6.

Data was gathered from a broad sample of 106 languages distributed across 42 families and 72 genera. On average there are 1.47 languages per genera, 2.52 languages per family and 1.7 genera per family, but the distribution is rather uneven. Two families,
Indo-European and Afro-Asiatic, account for over a third (34%) of the languages in the sample and 17% of the genera. At the other end of the scale, 28 families are represented by a single language. Of the 72 genera, 18 are represented by more than one language. The areal distribution is shown in Figure 1. (The map includes ISO 693-3 language codes which are legible via zooming in the electronic version of this document.)

Figure 1: Areal distribution of language sample

The sources employed were the World Loanword Database, questionnaires and dictionaries. The resulting database contains over 10,000 data points, of which 3,738 are binominals. The latter were coded by morphosyntactic strategy according to the nine-way typology described in the Introduction, and by the low-level classification of semantic relations described in the next section.

4 The classification schemes

In this section I describe the two classification schemes used in the analysis, one low-level, consisting of 29 relations, the other high-level, consisting of just five relations.

4.1 Bourque2: low-level

For the low-level classification I turned to Yves Bourque’s dissertation entitled Toward a typology of semantic transparency: The case of French compounds (Bourque 2014). The core of this work is an examination of the semantic relations found in 1,048 French compounds of the type NN and N à N, based on a classification developed by the author. There were a number of reasons for selecting this classification for the present study:
1. The work is based on a detailed and comprehensive survey of some 20 earlier schemes from which Bourque synthesizes his own set of relations. Bourque thus stands on the shoulders of giants, instead of simply starting from scratch, as tends to be the case in this field.

2. Unlike some other researchers, Bourque recognizes that some relations work in two directions, e.g. part-whole (*wheelchair*) and whole-part (*table leg*); these are termed reversible.

3. The scheme consists of a moderate number of classes (15 + 10 inverse = 25), a size which was deemed appropriate for the present study. It is more granular than the 12-14 relations of Levi, Warren and Downing; this is important, because it is much more challenging to classify the data correctly and consistently using a very abstract system. On the other hand, it is not as fine-grained as Ryder’s system of 50 relations; applying the latter to over 3,700 binominals would have been prohibitive.

4. Each relation is accompanied by a full description, the presentation extending over 40 pages. This makes it easier for another researcher to understand just what is intended by each relation.

5. The metalanguage is English while the object language is French. As a result, the terminology Bourque employs is less Anglocentric than that encountered in studies based only on English. For example, it makes no sense to use the names of English verbs and prepositions (cf. Levi’s *HAVE*, *BE* and *FROM*) to describe relations in French. This is also reflected in Bourque’s use of *NON-HEAD* (or modifier) and *HEAD* instead of *N1* and *N2* (cf. Levi, Jackendoff and many others), or *A* and *B* (Hatcher), both of which imply head-final order.

6. English and French examples are provided for each relation. This makes the dissertation accessible to researchers who are not fluent in French.

7. Each relation is accompanied by a test frame or “template” consisting of both an English and a French paraphrase. The ensures much greater accuracy in the assignment of relation types to individual binominals.

8. Bourque includes discussion of overlaps between different relations. This is important because semantic relations, at least at Bourque’s level of granularity, are prototypical rather than Aristotelian categories. Because of the gradience they exhibit, some binominals could be seen as instances of multiple relations. While this is in the nature of the phenomenon, it is helpful to receive a third party’s confirmation that certain cases will be slightly ambiguous.

9. Finally, Bourque has made his data available on his website (Bourque 2016), thus facilitating further exploration of his classification.

In sum, these nine aspects of Bourque’s work provide a compelling argument for adopting his classification scheme. Three of Bourque’s reversible relations are shown in Table 1, along with the corresponding templates, linking material and examples.
Table 1: Bourquifier template for PRODUCTION

<table>
<thead>
<tr>
<th>Label</th>
<th>Type</th>
<th>Template</th>
<th>Linking material</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernymy</td>
<td>Basic</td>
<td>an H of kind M</td>
<td>kind of, type of</td>
<td>oak tree</td>
</tr>
<tr>
<td></td>
<td>Rev.</td>
<td>an H that M is a kind of</td>
<td></td>
<td>bear cub</td>
</tr>
<tr>
<td>Part</td>
<td>Basic</td>
<td>an H that is part of M</td>
<td>part of (have / of)</td>
<td>table leg</td>
</tr>
<tr>
<td></td>
<td>Rev.</td>
<td>an H that M is part of</td>
<td></td>
<td>wheelchair</td>
</tr>
<tr>
<td>Production</td>
<td>Basic</td>
<td>an H that makes M</td>
<td>makes, produces</td>
<td>honey bee</td>
</tr>
<tr>
<td></td>
<td>Rev.</td>
<td>an H that M makes</td>
<td></td>
<td>beeswax</td>
</tr>
</tbody>
</table>

In applying Bourque’s classification to my data I experienced the need to extend it slightly, refine some of the names and templates, and change some of the examples, for reasons of consistency and to correct some minor errors. For example, I renamed Hypernymy to TAXONOMY, on the grounds that the former is a lexical relation, not a conceptual relation. I also renamed Part to MERONOMY, since “part” names one of the roles in the relation (the other being “whole”), and this could cause confusion. Bourque’s examples for his Source relation, cane sugar and sugar cane, have the pleasing property of consisting of the same two elements in reverse order; where possible I changed other examples in order to have the same property, for example by substituting song bird and birdsong for Bourque’s honey bee and beeswax.

Two new reversible relations, CONTAINMENT and DIRECTION, were added at a late stage; these were not used in the data annotation but are part of the final version of what I call the Bourque2 classification, which thus consists of 29 relations all told: 12 of them reversible and five unidirectional (see Table 4).

In order to simplify the daunting task of classifying over 3,700 binominals, I created an Excel application called the Bourquifier, a screenshot of which is shown in Figure 2. To use it, one simply types in the binominal, modifier and head and all the templates are automatically populated at the same time. These can then be scanned in a matter of seconds to find the most appropriate relation.

I have illustrated this with Bourque’s example, oak tree. The 12 reversible relations and five unidirectional relations are listed under the heading Relation; paraphrases appear under the headings Basic template and Reversed template (note that the latter are blank for unidirectional relations). The column labelled Roles contains my proposals for naming the roles played by the participants in each relation; the column labelled B2 contains codes for each relation that are used in the database: these mnemonics will be used extensively in §5, so the reader is advised to get a feel for them immediately. The other parts of the Bourquifier interface will become clear after the two-tier classification has been presented in §4.3.
### Table 2: The Bourquifie – OAK TREE

Table 2 gives examples of each of the 24 relations that actually occur in the binominal data from 28 different languages, together with the gloss and paraphrase. Of the 29 relations in Bourque2, two (reversible) relations (CONTAINMENT and DIRECTION) were not used, and one (Basic COMPOSITION, i.e. COMP) was not found. These five are therefore exemplified using data from elsewhere.
Table 2: Cross-linguistic examples of Bourque2 relations

<table>
<thead>
<tr>
<th>relation</th>
<th>B2</th>
<th>H2</th>
<th>language</th>
<th>binomial [gloss]</th>
<th>meaning</th>
<th>type</th>
<th>paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAXONOMY</td>
<td>TAX</td>
<td>MisH</td>
<td>Trinitario</td>
<td>mopó.ṣi [bee_related.CLF(round)]</td>
<td>BEE</td>
<td>cls</td>
<td>a bee is a kind of CLF:round</td>
</tr>
<tr>
<td></td>
<td>TAX-R</td>
<td>MisH</td>
<td>Norwegian</td>
<td>hund.e.kvalp [dog.LE.puppy]</td>
<td>PUPPY</td>
<td>gen</td>
<td>a puppy is a kind of dog</td>
</tr>
<tr>
<td>COORDINATION</td>
<td>COOR</td>
<td>MisH</td>
<td>Murui Huítoto</td>
<td>jītaiːño [adolescent.CLF(fem)]</td>
<td>GIRL</td>
<td>cls</td>
<td>a female that is also an adolescent</td>
</tr>
<tr>
<td>SIMILARITY</td>
<td>SIM</td>
<td>MisH</td>
<td>Central Yupik</td>
<td>pacicaguq [gills.UAQ]</td>
<td>NOSTRIL</td>
<td>der</td>
<td>a NMLZ that is similar to a gills</td>
</tr>
<tr>
<td>TOPIC</td>
<td>TOP</td>
<td>HinM</td>
<td>Greek</td>
<td>pinakida kikloforia.s [sign traffic.GEN]</td>
<td>LICENSE PLATE</td>
<td>gen</td>
<td>a sign that is about a traffic</td>
</tr>
<tr>
<td>CONTAINMENT</td>
<td>CONT</td>
<td>HinM</td>
<td>Portuguese</td>
<td>semente de laranja [seed PREP DEF.orange]</td>
<td>ORANGE SEED</td>
<td>prp</td>
<td>a seed that is contained in an orange</td>
</tr>
<tr>
<td></td>
<td>CONT-R</td>
<td>MinH</td>
<td>Wolof</td>
<td>ndāh.u lem [container.PER honey]</td>
<td>BEEHIVE</td>
<td>con</td>
<td>a container that contains a honey</td>
</tr>
<tr>
<td>POSSESSION</td>
<td>POSS</td>
<td>HinM</td>
<td>Oroqen</td>
<td>dʒuː.gdu.ʒi dʒuː.n [bee.GEN house.3SG.POSS]</td>
<td>BEEHIVE</td>
<td>dbl</td>
<td>a house that is possessed by a bee</td>
</tr>
<tr>
<td></td>
<td>POSS-R</td>
<td>MinH</td>
<td>Swahili</td>
<td>fumo [CL9:spear]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MERONOMY</td>
<td>MER</td>
<td>HinM</td>
<td>Áiwoo</td>
<td>nagago nyike [digit foot]</td>
<td>TOE</td>
<td>jxt</td>
<td>a digit that is part of a foot</td>
</tr>
<tr>
<td></td>
<td>MER-R</td>
<td>MinH</td>
<td>Cabécar</td>
<td>uva kal.i [grape tree.SPEC]</td>
<td>VINE</td>
<td></td>
<td>a tree that is a part of</td>
</tr>
<tr>
<td>LOCATION</td>
<td>LOC</td>
<td>MinH</td>
<td>Western Mari</td>
<td>pöläß.tängä [ear.coin]</td>
<td>EARRING</td>
<td>cmp</td>
<td>a coin located at/near/in an ear</td>
</tr>
<tr>
<td></td>
<td>LOC-R</td>
<td>MinH</td>
<td>Anindilyakwa</td>
<td>a.mukwa a. nguara [CLF.source CLF.fire]</td>
<td>FIREPLACE</td>
<td>jxt</td>
<td>a source that a fire is located at/near/in</td>
</tr>
<tr>
<td>TEMPORALITY</td>
<td>TEMP</td>
<td>HinM</td>
<td>Akkadian</td>
<td>naptan šер.ım [meal:STC.morning.GEN]</td>
<td>BREAKFAST</td>
<td>dbl</td>
<td>a meal that occurs at/during a morning</td>
</tr>
<tr>
<td></td>
<td>TEMP-R</td>
<td>MinH</td>
<td>Puyuma</td>
<td>ka-lauk-an [TLP-lunch-LOC]</td>
<td>MIDDAY</td>
<td>der</td>
<td>a TEMP at/during which a lunch occurs</td>
</tr>
<tr>
<td>COMPOSITION</td>
<td>COMP</td>
<td>HinM</td>
<td>German</td>
<td>Würfel.zucker [cube.sugar]</td>
<td>CUBE SUGAR</td>
<td>cmp</td>
<td>a sugar that a cube is composed of</td>
</tr>
<tr>
<td></td>
<td>COMP-R</td>
<td>MinH</td>
<td>Navajo</td>
<td>tsé bee na’nt’á [stone with it bridge]</td>
<td>STONE BRIDGE</td>
<td>prp</td>
<td>a bridge that is composed of stone</td>
</tr>
<tr>
<td>DIRECTION</td>
<td>DIR</td>
<td>HioM</td>
<td>French</td>
<td>adoration de soleil [worship PREP sun]</td>
<td>SUN WORSHIP</td>
<td>prp</td>
<td>a worship whose goal is a sun</td>
</tr>
<tr>
<td></td>
<td>DIR-R</td>
<td>MtoH</td>
<td>English</td>
<td>sales target</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOURCE</td>
<td>SRC</td>
<td>HioM</td>
<td>Russian</td>
<td>saxar.ny trostnik [sugar.ADJ.Z cane]</td>
<td>SUGAR CANE</td>
<td>adj</td>
<td>a cane that is a source of a sugar</td>
</tr>
<tr>
<td></td>
<td>SRC-R</td>
<td>MtoH</td>
<td>Caijia</td>
<td>k’ŋý ’ja’ṱ’í [sky.OLD woman]</td>
<td>QUEEN</td>
<td></td>
<td>an old woman whose source is a sky</td>
</tr>
<tr>
<td>CAUSATION</td>
<td>CAUS</td>
<td>HioM</td>
<td>Bezhta</td>
<td>ʃ.erec’.madi [(onom).NMLZ]</td>
<td>WATER PUMP</td>
<td>der</td>
<td>a NMLZ that causes a ONOM</td>
</tr>
<tr>
<td></td>
<td>CAUS-R</td>
<td>MtoH</td>
<td>Takia</td>
<td>nə.n aba.n [leg.3SG.mark.3SG]</td>
<td>FOOTPRINT</td>
<td>dbl</td>
<td>a mark that a leg causes</td>
</tr>
<tr>
<td>PRODUCTION</td>
<td>PROD</td>
<td>HioM</td>
<td>Turkish</td>
<td>sitt ineq.i [milk cow.3SG]</td>
<td>DAIRY COW</td>
<td>con</td>
<td>a cow that produces a milk</td>
</tr>
<tr>
<td></td>
<td>PROD-R</td>
<td>MtoH</td>
<td>Polish</td>
<td>wosk pzech.eli [wax bee.ADJ.Z]</td>
<td>BEESWAX</td>
<td>adj</td>
<td>a wax that a bee produces</td>
</tr>
<tr>
<td>USAGE</td>
<td>USG</td>
<td>HioM</td>
<td>Harakmbut</td>
<td>amiko.en kinamah [foreigner.GEN bag]</td>
<td>BACKPACK</td>
<td>gen</td>
<td>a bag that a foreigner uses</td>
</tr>
<tr>
<td></td>
<td>USG-R</td>
<td>MtoH</td>
<td>Lower Sorbian</td>
<td>weth. nik [wind.NMLZ]</td>
<td>WINDMILL</td>
<td>der</td>
<td>a NMLZ that uses a wind</td>
</tr>
<tr>
<td>FUNCTION</td>
<td>FUNC</td>
<td>HioM</td>
<td>Slovak</td>
<td>klůć.ové slovo [key.ADJ.Z word]</td>
<td>KEYWORD</td>
<td>adj</td>
<td>a word that serves as a key</td>
</tr>
<tr>
<td>PURPOSE</td>
<td>PURP</td>
<td>HioM</td>
<td>Wik-Mungkan</td>
<td>yu mée’.akana [tree/thing.EY.DAT]</td>
<td>GLASSES</td>
<td>gen</td>
<td>a tree/thing intended for an eye</td>
</tr>
</tbody>
</table>
4.2 Hatcher2: high-level

The high-level classification used in this study takes Hatcher’s (1960) four-way system as its starting point. Her paper is a trenchant critique of Jespersen’s (1942) attempt to classify semantic relations, mentioned above. Jespersen conceded that his analysis was incomplete and that there were many compounds which “do not fit in anywhere”. He claims that his failure is due simply to the inherent unclassifiability of his material: “the number of possible logical relations between the two elements is endless”; “the analysis of the possible sense-relations can never be exhaustive”. But, says Hatcher:

it all too often happens that scholars in linguistics proclaim a given problem to be insoluble, when they themselves have not worked out the categories necessary for its solution; we should, then, examine the outline offered by Jespersen to see if some of the difficulty he encountered may not be explained by his method of classification. For example, was his set of categories constructed with logical rigor: and, before surrendering to the “difficult” types that he mentions, had he been able, at least, to account for all the “easy” compounds, subdividing these as carefully as his patience and his talent permitted? The subdivision of the obvious may lead to greater understanding of the less obvious, if one is guided by logically consistent criteria (p. 356).

Thereupon, Hatcher sets about tearing Jespersen’s system to pieces (her words, see p. 365). She starts by listing seven of Jespersen’s types, omitting one of original eight on the grounds that it more properly belongs to “apposition”, which she wants to keep separate. Examining each of these in turn, Hatcher notes a lack of careful subdivisions, an absence of any principle of symmetry, and mixing of two basic criteria, Reference and Relation, exclaiming:

Little wonder that to Jespersen the difficulty of classifying our compounds was insuperable. But how could he fail to see the inadequacy of his categories? How could any linguist today construct a system of classification based on two (or three) kinds of main criteria? (p. 361).

Jespersen’s blindness to the flaws of his outline is put down to his “consummate lack of interest in the problem”; as to the second question, the answer is that what Jespersen actually offers is no more than a simplification of the classification suggested by Mätzner in 1860. “It is sad,” writes Hatcher, “that a linguist of today cannot go beyond a linguist of the nineteenth century; it is sadder still if his work is inferior to his predecessor’s. And this is the case, here”.

Hatcher now sets about rearranging Jespersen’s scheme as depicted in Figure 3. For consistency she chooses to base her new scheme on Relation exclusively and to avoid Reference, so she starts by separating the first three of Jespersen’s types (1)-(3) – all of which are either based on reference or mixed – from the rest (4)-(7), which are all relational. The former are set to one side, and to the latter she adds the two relational types that Jespersen for some reason failed to adopt from Mätzner (α broomstick and β castor

---

3 Numbers (and Greek letters) in bold refer to the nine divisions in Figure 3.
Semantic relations in binominal lexemes: A cross-linguistic survey

She then proceeds to reorganize these six relational types into four abstract classes (6):

(6) (a) \( A \subseteq B \) “A is contained in B”
(b) \( A \supseteq B \) “B is contained in A”
(c) \( A \rightarrow B \) “A is the source of B”
(d) \( A \leftarrow B \) “A is the destination of B”

Having reduced the six relational categories of Jespersen/Mätzner to two pairs of mutually exclusive concepts (6), Hatcher now turns her attention to the referential types in order to see how they might be accommodated in her new scheme. She starts with Place (2), Time (3) and their subdivisions, which map neatly into her scheme, as shown in the diagram. Finally, the two verbal types (1) Subject and (2) Object are “easy”:

Sunshine and sun worship, these perfect opposites, fall under \( A \rightarrow B \) and \( A \leftarrow B \), respectively. Surely the subject is the “source” of its own activity (in putting sunshine under \( A \rightarrow B \), we are merely adding Agent to Agency); and in sun-worship (\( A \leftarrow B \)), the sun is that toward which the worship is directed.

Thus we see that both the referential and the relational types of Matzner-Jespersen can be included in our two pairs of relational criteria: the static Ⓡ and Ⓢ, and the dynamic \( A \rightarrow B \) and \( A \leftarrow B \) (p. 365).

Hatcher concludes this part of her analysis by pointing out that the scheme she has developed has two advantages over the one she has just “torn to pieces”. Firstly, it is logically conceived, and therefore neater and more pleasing aesthetically; and secondly, it is far more comprehensive, and thus may “be able to account for all possibilities of determinative, non-appositional compounding in the English language,” which she suggests are surely not “endless”. At the same time she expresses the hope that her work represents not a “result”, but rather a beginning, and that it will offer “a more spacious framework” within which research dedicated to the proposition that “all compounds are endowed by their creators with the right to belong somewhere” may proceed more profitably and hopefully than before.
Hatcher’s work is often cited (Citeseer, accessed 2018-05-30, counts 328 citations), but usually dismissed, often on less than scientific grounds. For example, Søgaard writes: such an account is by definition both arbitrary (Bauer 1978; van Santen 1979) and incomplete because of the infinite set of compounding relationships. For illustration, try to place a compound such as car thief in [Hatcher’s] four-way typology. Is a car thief a ‘car in a thief’, a ‘thief in a car’, a ‘thief as the goal of a car’ or a ‘thief as the source of a car’? (Søgaard 2005: 320).
Unfortunately for Søgaard the last two paraphrases are incorrect: He has muddled up the order of A and B. The head of the construction (B) is *thief*, not *car*, so these two paraphrases should read: a ‘*car* as the goal of a *thief*’ and a ‘*car* as the source of the *thief*’. With the correct paraphrase, it is obvious that the car is indeed the goal of the thief (i.e. A ← B). Søgaard’s objection must therefore be rejected.

One researcher who has taken Hatcher seriously is Arnaud (2016). Arnaud’s paper on categorizing the modification relations in French subordinative NN<sub>N</sub> compounds is full of interesting observations, examples and discussion, and I cannot do it justice here. The interested reader is encouraged to consult it directly. In the present context it is mainly noteworthy for the fact that Arnaud first develops his own highly granular classification, and then maps it onto Hatcher’s scheme, which is what I will do with the Bourque classification.

Arnaud’s classification is based on a database of 949 French binominals of type cmp, which he dubs “les composés timbre-poste” (‘postage stamp compounds’, Arnaud 2003). As none of the then-existing taxonomies of semantic relations seemed satisfactory, he decided to start from the data up, applying the principles of cognitive linguistics, “in particular the idea that relations are emergent phenomena which gain psychological existence” (p. 71). After the first stage of work based on 809 compounds, he ended up with 54 categories, each with a definition of the type. In the second stage, in which he analysed a further 140 compounds, Arnaud was obliged to adapt some of his categories slightly and to add four new ones, which confirmed for him “the frequently expressed opinion that a categorization of compounds cannot be exhaustive” (p. 82). However, he also notes that we are “clearly in a situation of diminishing returns, since most of the units in the second dataset were accounted for by already identified relations.”

Next Arnaud applies his taxonomy to a random sample of 200 compounds drawn from a dataset of 3000 lexicalized English NN compounds. Once more he is obliged to modify a few categories and add some new ones, but no more than six. This is remarkably few considering that compounds are considerably less numerous in French than in English, and that French has an entrenched competing N PREP N construction that is preferred for some types of relation (for example, habitat and part-whole), as Arnaud himself points out (p. 89). This suggests that the number of relations required at any given level of granularity will tend to flatten out and not increase indefinitely.

Of most interest to us is that Arnaud now proceeds to map his set of 58 empirically derived (low-level) relations to Hatcher’s set of four logically derived (high-level) relations. For the most part, this is plain sailing. Arnaud’s bottom-up deduction melds neatly with Hatcher’s top-down induction. Or at least, it almost does. Arnaud does have problems with some units that he feels do not correspond to one of the four abstract categories. One example is régime jockey [diet jockey] ‘jockey diet’, which denotes a diet that is typical of jockeys. To cater for these cases, Arnaud feels obliged to create a fifth category, ANALOG (not to be confused, he insists, with the attributive relation). In all, four “supplementary abstract categories” were necessary.
In Pepper (2020) I examine each of the 12 lower-level relations that seemed to Arnaud to justify the creation of his four new high-level relations and show that all but one of them can in fact be accommodated by Hatcher’s four-way system. The exception, one of the four subtypes ANALOG (7) presents a different and more interesting case.

(7) ANALOG (fourth subtype)

NON-HEAD names analogically a perceptual characteristic of HEAD
brasse papillon [breast_stroke butterfly] ‘butterfly stroke’

Here there can be no doubt that some kind of analogy is at work. But brasse papillon is not a non-appositional compound in Hatcher’s terms and therefore falls outside her scope (recall that she restricts herself to non-appositional compounds). If we want to extend Hatcher’s scheme to cover appositional compounds, then we do indeed need a new high-level relation. However, ANALOGY may not be the best term for that relation. Hatcher’s logically defined pair of reversible relations in (6) are both based on Contiguity, which is one of Aristotle’s “three principles of remembering”, the others being Similarity and Contrast. In Pepper (2020) I suggest that the relation underlying the types of appositional compound discussed in by Hatcher herself in an earlier paper (Hatcher 1952) – species-genus and cross-classification – as well as Arnaud’s brasse papillon (and incidentally also coordinative compounds) is Similarity. This is at about the right level of generality or abstraction as those of Hatcher and so I propose to extend Hatcher’s four-way system to a five-way system (Hatcher2) that also includes appositional compounds.

The new system is summarized in (8) and (9), in which I have taken the further step of replacing Hatcher’s A and B with M (for modifier) and H (for head), in order to make it more suitable for cross-linguistic comparison. Hatcher drew attention to the fact that her four relations comprised two pairs, which she characterized as “static” and “dynamic”. I have recast them in terms of two superordinate relations, CONTAINMENT and CAUSATION (or source/goal). The four-letter codes in italics are the codes used in the database.

(8) Contiguity-based

CONTAINMENT
(a) M ⊃ H HinM “H is contained in M” (orange seed)
(b) M ⊂ H MinH “M is contained in H” (seed orange)
CAUSATION (source/goal)
(c) M ← H HtoM “M is the destination of H” (sugar cane)
(d) M → H MtoH “M is the source of H” (cane sugar)

(9) Similarity-based

SIMILARITY
(e) M ≈ H MisH “H is similar or identical to M”

4.3 Hatcher-Bourque: multilevel

Once Hatcher’s four-way system covering non-appositional compounds was extended to a five-way system that also covers appositional compounds, mapping the revised set of
29 relations in Bourque2 posed no problems, confirming again the robustness of Hatcher’s system. Table 3 provides examples of the mapping.

*Table 3: Mapping Bourque2 to Hatcher2*

<table>
<thead>
<tr>
<th>Bourque2</th>
<th>Template</th>
<th>Example</th>
<th>Hatcher2</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAXONOMY</td>
<td>an M is a kind of H</td>
<td><em>oak tree</em></td>
<td>Similarity</td>
<td>A $\equiv$ B</td>
</tr>
<tr>
<td></td>
<td>an H is a kind of M</td>
<td><em>bear cub</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MERONOMY</td>
<td>an H that is part of M</td>
<td><em>car motor</em></td>
<td>Containment (“static”)</td>
<td>A $\supset$ B</td>
</tr>
<tr>
<td></td>
<td>an H that M is part of</td>
<td><em>motor car</em></td>
<td></td>
<td>A $\subseteq$ B</td>
</tr>
<tr>
<td>PRODUCTION</td>
<td>an H that produces M</td>
<td><em>song bird</em></td>
<td>Causation (“dynamic”)</td>
<td>A $\leftarrow$ B</td>
</tr>
<tr>
<td></td>
<td>an H that M produces</td>
<td><em>birdsong</em></td>
<td></td>
<td>A $\rightarrow$ B</td>
</tr>
</tbody>
</table>

(21 further low-level relations) (same five high-level relations)

Arnaud does not explain his motivation for trying to map his low-level system of 58 relations to Hatcher’s four; it was perhaps a purely intellectual exercise designed to test the robustness of Hatcher’s system. But in my own work there was a very practical reason for mapping Bourque2 to Hatcher2. The great advantage of the resulting articulated classification is that it *combines the strengths of high-level and low-level systems*: A highly abstract scheme like Hatcher’s is extremely difficult to apply consistently in practice, but is much more amenable to statistical analysis. A very granular classification, on the other hand, although easier to apply, often yields categories with insufficient data for statistical analysis. With the Hatcher-Bourque system one can achieve a more rigorous initial coding of the data with the 29 relations of Bourque2, and automatically map to the corresponding five of Hatcher2. The result is much greater overall consistency and a small number of more densely populated categories.

The two-tier Hatcher-Bourque scheme is given in its entirety in Table 4. The columns labelled *Bourque2, B2, Template* and *Example* are familiar from the presentation in §4.1. Those labelled *Hatcher2* and *H2* indicate the mappings to the five relations of Hatcher2 (cf. 8 and 9) and the codes used for them in the database, viz. *mīsh, hīm, mīn, hīm* and *mtoh*; the latter will be used extensively in §6.
### Table 4: Hatcher-Bourque classification of semantic relations

<table>
<thead>
<tr>
<th>Bourque2</th>
<th>B2</th>
<th>Hatcher2</th>
<th>H2</th>
<th>Template</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAXONOMY</td>
<td>TAX</td>
<td>M=H</td>
<td>MisH</td>
<td>an M is a kind of H</td>
<td>oak tree</td>
</tr>
<tr>
<td></td>
<td>TAX-R</td>
<td>M(\leq)H</td>
<td>MisH</td>
<td>an H that is contained in an M</td>
<td>orange seed</td>
</tr>
<tr>
<td>COORDINATION</td>
<td>COOR</td>
<td>M=H</td>
<td>MisH</td>
<td>an H that is also an M</td>
<td>boy king</td>
</tr>
<tr>
<td>SIMILARITY</td>
<td>SIM</td>
<td>M=H</td>
<td>MisH</td>
<td>an H that is similar to M</td>
<td>kidney bean</td>
</tr>
<tr>
<td>CONTAINMENT</td>
<td>CONT</td>
<td>M=H</td>
<td>HinM</td>
<td>M (\subseteq)H an H that is contained in an M</td>
<td>sunflower</td>
</tr>
<tr>
<td></td>
<td>CONT-R</td>
<td>M=H</td>
<td>MinH</td>
<td>M (\leq)H an H that contains an M</td>
<td>seed orange</td>
</tr>
<tr>
<td>POSSESSION</td>
<td>POSS</td>
<td>M=H</td>
<td>HinM</td>
<td>an H that is possessed by an M</td>
<td>family estate</td>
</tr>
<tr>
<td></td>
<td>POSS-R</td>
<td>M=H</td>
<td>MinH</td>
<td>an H that possesses an M</td>
<td>career girl</td>
</tr>
<tr>
<td>MERONOMY</td>
<td>MER</td>
<td>M=H</td>
<td>HinM</td>
<td>an H that is part of an M</td>
<td>car motor</td>
</tr>
<tr>
<td></td>
<td>MER-R</td>
<td>M=H</td>
<td>MinH</td>
<td>an H that M is part of</td>
<td>motor car</td>
</tr>
<tr>
<td>LOCATION</td>
<td>LOC</td>
<td>M=H</td>
<td>HinM</td>
<td>an H located at/near/in an M</td>
<td>house music</td>
</tr>
<tr>
<td></td>
<td>LOC-R</td>
<td>M=H</td>
<td>MinH</td>
<td>an H that is located at/near/in</td>
<td>music hall</td>
</tr>
<tr>
<td>TEMPORALITY</td>
<td>TEMP</td>
<td>M=H</td>
<td>HinM</td>
<td>an H that occurs at/during M</td>
<td>summer job</td>
</tr>
<tr>
<td></td>
<td>TEMP-R</td>
<td>M=H</td>
<td>MinH</td>
<td>an H at/during which M occurs</td>
<td>golf season</td>
</tr>
<tr>
<td>COMPOSITION</td>
<td>COMP</td>
<td>M=H</td>
<td>HinM</td>
<td>an H that M is composed of</td>
<td>cube sugar</td>
</tr>
<tr>
<td></td>
<td>COMP-R</td>
<td>M=H</td>
<td>MinH</td>
<td>an H composed of M</td>
<td>sugar cube</td>
</tr>
<tr>
<td>TOPIC</td>
<td>TOP</td>
<td>M=H</td>
<td>MinH</td>
<td>an H that is about M</td>
<td>history book</td>
</tr>
<tr>
<td>DIRECTION</td>
<td>DIR</td>
<td>M→H</td>
<td>HtoM</td>
<td>an H whose goal is an M</td>
<td>sun worship</td>
</tr>
<tr>
<td></td>
<td>DIR-R</td>
<td>M→H</td>
<td>MtoH</td>
<td>an H that is the goal of an M</td>
<td>sales target</td>
</tr>
<tr>
<td>SOURCE</td>
<td>SRC</td>
<td>M→H</td>
<td>HtoM</td>
<td>an H that M is sourced from</td>
<td>sugar cane</td>
</tr>
<tr>
<td></td>
<td>SRC-R</td>
<td>M→H</td>
<td>MtoH</td>
<td>an H that is sourced from</td>
<td>cane sugar</td>
</tr>
<tr>
<td>CAUSATION</td>
<td>CAUS</td>
<td>M→H</td>
<td>HtoM</td>
<td>an H that causes an M</td>
<td>tear gas</td>
</tr>
<tr>
<td></td>
<td>CAUS-R</td>
<td>M→H</td>
<td>MtoH</td>
<td>an H that an M causes</td>
<td>sunburn</td>
</tr>
<tr>
<td>PRODUCTION</td>
<td>PROD</td>
<td>M→H</td>
<td>HtoM</td>
<td>an H that produces an M</td>
<td>song bird</td>
</tr>
<tr>
<td></td>
<td>PROD-R</td>
<td>M→H</td>
<td>MtoH</td>
<td>an H that an M produces</td>
<td>birdsong</td>
</tr>
<tr>
<td>USAGE</td>
<td>USG</td>
<td>M→H</td>
<td>HtoM</td>
<td>an H that an M uses</td>
<td>lamp oil</td>
</tr>
<tr>
<td></td>
<td>USG-R</td>
<td>M→H</td>
<td>MtoH</td>
<td>an H that uses an M</td>
<td>oil lamp</td>
</tr>
<tr>
<td>FUNCTION</td>
<td>FUNC</td>
<td>M→H</td>
<td>HtoM</td>
<td>an H that serves as an M</td>
<td>buffer state</td>
</tr>
<tr>
<td>PURPOSE</td>
<td>PURP</td>
<td>M→H</td>
<td>HtoM</td>
<td>an H that is intended for an M</td>
<td>animal doctor</td>
</tr>
</tbody>
</table>

### 5 Low-level relations: frequency and distribution

In the following sections I analyse the distribution of both low-level and high-level semantic relations in my data, both in terms of their overall frequency and their distribution across languages, meanings, morphosyntactic strategies, semantic types and semantic fields. It will become apparent that some relations are far more frequent than
others, a fact that has the potential to provide insights into the way in which humans conceptualize the world. This is a topic which has hardly been addressed in the typological literature at all; to my knowledge, the only researcher to even approach the question from a cross-linguistic perspective is Bauer (2001), who has the following to say regarding *tatpurusa* (i.e. determinative) compounds:

In a detailed survey of just three languages, Bauer (1978: 147) points out that underlying semantic relationships of location appear to be the most common relationships in those languages. The same is true with the sample discussed here. Compounds in which the head is the location of the entity denoted in the modifier (e.g. English *furniture store*) or where the head denotes an entity located at the modifier (e.g. English *bone cancer*) are the types most frequently illustrated or commented on for the languages in my sample across all areas. The next most frequent type to be illustrated is the type where the head is made from the material in the modifier (e.g. English *sandcastle*). Other meanings are illustrated or commented on far more sporadically. While this does not show that other meanings are not also in common use, it does suggest that compounds may be used prototypically to indicate location or source (especially if ‘made from’, ‘made by’, ‘belonging to’ and ‘coming from’ are all interpreted as sources).

Under the Bourque scheme, Bauer’s three examples (*furniture store*, *bone cancer* and *sandcastle*) would be classified as LOC-R, LOC and COMP-R, respectively (“a store that (a) furniture is located at/near/in”, “a cancer located at/near/in a bone” and “a castle composed of sand”). The present data provide an opportunity to test Bauer’s conjecture and I do so in this section, which is concerned with the low-level relations of the Bourque2 scheme. Then, in the next section, I investigate the high-level relations of the Hatcher2 scheme.

5.1 Overall distribution

The overall frequency of low-level semantic relations in the database is shown in Figure 4 and is summarized in the following scale:

(10) \( \text{MER} \gg \text{PURP} > \text{COORD} > \text{LOC} > \text{COMP-R}, \text{POSS} > \text{USG-R} > \text{TEMP} > \ldots \)  

We note that MER (i.e. part-whole), a relation that Bauer does not consider in his discussion, is far and away the most frequent relation in absolute terms, at least in the present data. This is largely due to its extreme frequency in binominals that denote body parts, as we will see below. For this reason, Figure 4 also shows the values when body parts are excluded: apart from the greatly reduced frequency of MER and a slightly reduced frequency for LOC, the differences are minimal, so while it may be the case that method employed for selecting the set of meanings used in this study overstates the overall prevalence of MER, it is clearly among the most important relations and even with body parts excluded it is just as frequent as LOC.

Bauer’s suggestion that the next most frequent type is where the head is made from the material in the modifier (e.g. *sandcastle*) – this equates to the Reversed COMPOSITION relation, COMP-R – is also not supported out by the data, which put it at joint fifth in terms
of overall frequency. Instead, the next most frequent relation is PURPOSE (PURP), also not mentioned by Bauer. As we will see below, this relation is especially prevalent in binominals that belong to the domain Modern World and/or denote entities that fall into the semantic type Advanced technology (or concept).

Figure 4: Overall frequency of low-level semantic relations

The third most frequent relation is COORDINATION. This might seem surprising, given that the set of meanings was expressly designed to exclude binominals whose constituents are in a coordinate relation. However, this applied only to what Wälchli (2005) terms ‘co-compounds’: “word-like units consisting of two or more parts which express natural coordination”, such as Hmong Daw zaub.mov [vegetable.rice] FOOD and Vietnamese bô mê [father mother] PARENTS. The COORDINATION relation inherited from Bourque is broader than this: it is used for binominals that refer to “combinations that, from a semantic perspective, seem to involve both elements equally” (p.180). In our binominals data, the latter type is often encountered with items that denote animates of a certain age (Hawaiian kao keiki [goat child] KID), gender (Mbyá Guaraní kavaju kunha [horse woman] MARE), or both (Ket qim.dûl [woman.child] GIRL); cases such as these account for over 90% of binominals that exhibit the COORDINATION relation.

The LOCATION relation, in its LOC form as “(an) H located at/near/in (an) M” (e.g. TEAR < ‘eye’ + ‘water’ in several languages), is only the fourth most frequent. Together with its inverse, LOC-R, “(an) H that (an) M is located at/near/in” (e.g. Hupdê yɔ́h moṣ [medicine house] HOSPITAL), it is found in 428 binominals, i.e. 12% of the data. Thus Bauer’s suggestion that this is the most common kind of relation appears to be unfounded.

5.2 Languages

We can also look at relations in terms of the number of languages in which each relation is attested (Figure 5).

The frequency scale here is:

(11)  MER > LOC > COOR, POSS, PURP > COMP-R > USG-R, LOC-R, PROD > …
The same six relations predominate in both distributions, albeit with slightly different rankings. Note that one relation (COMP, “(an) H that (an) M is composed of”, e.g. cube sugar) is not attested at all in the database, and that a further four (CAUS, SRC-R, TAX-R and TEMP-R) are rare.

5.3 Meanings

The distribution across meanings (Figure 6) shows a generally similar scale, but now with the TAX relation displaying far greater prominence. USG now appears among the top six, with COMP-R and POSS relegated to joint 9th and 11th place:

(12)  MER > COOR, PURP, TAX, LOC > USG, SIM, USG-R, POSS, LOC-R, COMP-R…

This suggests that while TAX (“(an) M is a kind of H”, e.g. oak tree) is not especially common, it is rather versatile; conversely, while COMP-R and POSS are rather frequent, their scope of application is relatively limited. It is also worth noting that of the 45 binominals that exhibit the TAX relationship, 18 employ the der strategy. In many cases the gloss
indicates an (apparently redundant) nominalizer or diminutive affixed to a root whose meaning is the same as that of the derived form, as in Lithuanian *spēn.eiš* [nipple.DIM] NIPPLE OR TEAT.

Overall, the data indicate that the most frequent low-level semantic relations cross-linguistically, at least as far as binominal lexemes are concerned, are the following:

### Table 5: Most frequent low-level semantic relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Template</th>
<th>example</th>
<th>H2 code</th>
<th>H2 symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>MER</td>
<td>(an) H that is part of (an) M</td>
<td><em>table leg</em></td>
<td>HinM</td>
<td>M ⊃ H</td>
</tr>
<tr>
<td>PURP</td>
<td>(an) H intended for (an) M</td>
<td><em>animal doctor</em></td>
<td>HtoM</td>
<td>M← H</td>
</tr>
<tr>
<td>COOR</td>
<td>(an) H that is also (an) M</td>
<td><em>boy king</em></td>
<td>MisH</td>
<td>M ≈ H</td>
</tr>
<tr>
<td>LOC</td>
<td>an H that (an) M is located at/near/in</td>
<td><em>music hall</em></td>
<td>HinM</td>
<td>M ⊃ H</td>
</tr>
<tr>
<td>POSS</td>
<td>(an) H that (an) M possesses</td>
<td><em>family estate</em></td>
<td>HinM</td>
<td>M ⊃ H</td>
</tr>
<tr>
<td>COMP-R</td>
<td>(an) H composed of (an) M</td>
<td><em>sugar cube</em></td>
<td>MinH</td>
<td>M ⊂ H</td>
</tr>
</tbody>
</table>

#### 5.4 Morphosyntactic strategies

Figure 7 shows how many morphosyntactic strategies are used to express each kind of relation. Comparison with the overall frequency scale extracted from Figure 4 (above) shows that the most frequent relations can be expressed by any one of the nine binominal types; this suggests that – at this level of analysis – there is no overall correlation between morphosyntactic strategy and semantic relation. As the data become more sparse, the number of strategies declines; thus, at the lower end of the scale, we find TAX-R, TEMP-R and SRC-R, each of which is expressed by just two strategies. However, since each of these three relations is represented in the database by just two or three exemplars, this does not constitute evidence against the lack of overall correlation.

*Figure 7: Number of binominal types that exhibit each relation*
5.5 Semantic types

The frequency of different relations varies according to the semantic type of the referent. Figure 8 shows the proportional distribution of the six most common relations – MER, PURP, COOR, LOC, COMP-R and POSS – across the seven semantic types listed in (5). The results for Animal, Natural phenomenon and Location should be approached with caution, since these semantic types represent only 7, 5 and 12 of the 100 meanings, respectively, but the differences between the other four types are striking.

![Figure 8: Semantic relations and semantic types](image)

As noted above, in binominals that denote Body parts the MER relation accounts for 85% of the data; the only significant alternative is LOC, which is the preferred relation for naming bodily substances, such as EARWAX and TEAR. On the other hand, MER is rarely used to denote an Advanced technology (or concept), such as BICYCLE PUMP, KEYWORD or RAILWAY; instead, the PURP relation predominates, accounting for over 80% of the data, with COMP-R the most widely used alternative (as in many words for RAILWAY, which is often conceptualised as a road composed of iron). In short, there is a strong tendency to name (secondary) body parts/fluids in terms of the (primary) body parts they are a part of/located at, and to name advanced concepts in terms of either their intended function or the material they are made of.

The semantic type Basic technology (or concept) is more mixed: as with Advanced technology (or concept), PURP and COMP-R are the most widespread relations, but the two are now equally common; however, in contradistinction to the latter type, MER and LOC are also quite frequent. These are also the most widely used relations for Natural phenomena – together with POSS, which expresses the relation between a spider and its web, or bees and their hive, and phenomena viewed as belonging to supernatural beings, such as Ket Albara 곆 ‘Alba’s hunting trail’ MILKY WAY and Assamese ramdhenu ‘Lord Rama’s bow’ RAINBOW.
5.6 Semantic fields

A similar variation is found across semantic fields (4). Figure 9 shows the frequency of the six most common semantic relations across the nine most frequent semantic fields. We note again that MER plays the dominant role in The body, but also in Agriculture and vegetation and Food and drink; and, as expected, Modern world is dominated by the PURP relation. We see also that the patterning in Animals and Kinship is remarkably similar: binominals in these fields have an overwhelming preference for either COOR or POSS. The latter is also widely used in Social and political relations. Finally, the LOC relation that Bauer assumed to be most widespread is in fact largely confined to the fields of The physical world and Clothing and grooming.

![Figure 9: How low-level relations vary across semantic fields](image)

6 High-level relations: frequency and distribution

I turn now from the low-level semantic relations of Bourque2 to the high-level relations of Hatcher2. For ease of reference, Table 6 provides a summary of the five high-level relations and those 25 low-level relations used to annotate the data that map to them. The terms used in the column Role will sometimes be used in the following in order to simplify the discussion. They are in effect shorthand labels for the five Hatcher2 relations.
Semantic relations in binominal lexemes: A cross-linguistic survey

Table 6: Summary of high- and low-level semantic relations

<table>
<thead>
<tr>
<th>Role</th>
<th>Bourque2</th>
<th>Hatcher2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MisH</td>
<td>TAX, TAX-R, COOR, SIM</td>
<td>M ≡ H</td>
</tr>
<tr>
<td>HinM</td>
<td>POSS, MER, LOC, TEMP, COMP, TOP</td>
<td>M ⊃ H</td>
</tr>
<tr>
<td>HtoM</td>
<td>SRC, CAUS, PROD, USG, FUNC, PURP</td>
<td>M ← H</td>
</tr>
</tbody>
</table>

The first four plots in the previous section showed how the low-level relations distribute across the database as a whole (with and without body parts), and across languages, meanings and morphosyntactic strategies. Figure 10 provides similar information for the high-level relations. Predictably, the amount of information is considerably reduced; on the other hand, the categories are considerably more balanced and therefore more amenable to statistical analysis.

Figure 10: Binominals, languages, meanings and strategies by htype

The first thing to note is that every one of the nine morphosyntactic strategies is attested in the data as expressing each of the five high-level relations (plot d); this provides additional evidence that there is no overall, cross-linguistic correlation between morphosyntactic strategies and semantic relations.

The high-level relation HinM (Hatcher’s “B is contained in A”) accounts for nearly half of the data (a). This comes as no surprise, given that this relation subsumes MER; if body parts are excluded it has roughly the same frequency as HtoM (Hatcher’s “A is the destination of B”), which subsumes the PURP relation, among others. With body parts included, the overall scale is

---

4 Recall that Hatcher uses the letters A and B to denote the first and second constituents of a typical right-headed English compound, so A is the modifier (M) and B is the head (H). To avoid confusion, think of Hatcher’s letters as denoting the Attribute (modifier) and Base (head).
(13) \( \text{HinM} \gg \text{HtoM} > \text{MisH, MinH} > \text{MtoH} \)

and with body parts excluded, the scale is

(14) \( \text{HinM, HtoM} > \text{MisH, MinH} > \text{MtoH} \)

The two most frequent low-level relations (\( \text{HinM} \) and \( \text{HtoM} \)) account for two-thirds of the data and thus suggest a pronounced tendency for a non-basic meaning to be conceptualized in terms of either its container or its goal (i.e. destination) – both of which should be interpreted in Hatcher’s very broad sense.

Plot (b) tells us that \( \text{HinM} \) is ubiquitous, occurring in every language in the sample. However, the other four low-level relations are also widespread across languages and they are probably also ubiquitous. The fact that they are not attested in every language is almost certainly due to the paucity of data for some languages: it would be highly unlikely that a language that is represented by fewer than, say, 10 data points\(^5\) would exhibit all five high-level relations.

The distribution of relations across meanings (c) shows a scale is similar to the two preceding ones:

(15) \( \text{HinM} > \text{HtoM} > \text{MisH} > \text{MinH} > \text{MtoH} \)

but the values are more spread out: \( \text{HinM} \) is less dominant, while \( \text{MisH} \), the similarity-based relation added to Hatcher’s original four is higher up the scale (in the sense that it is significantly more widespread across meanings than \( \text{MinH} \)). This reflects what was referred to as the versatility of the \( \text{TAX} \) relation, shown in (12), above. More worthy of mention, though, is the fact that none of the high-level relations appears suited for conceptualizing anything like the full range of meanings. Even \( \text{HinM} \), which is found in every language and accounts for over 45\% of all binominals in the database, is used with only just over half of the 100 meanings: in other words, there are limits to the versatility of conceptualizations that are based on how an entity is (in the broadest sense) “contained”.

With regard to semantic types, Figure 11 shows clearly that such containment is central to the conceptualization of (secondary) Body parts and also important for concepts that express Location or denote Basic technologies (or concepts) or entities in the Natural world. On the other hand, it is marginal to the conceptualization of Persons and of almost no use when it comes to Animals and Advanced technologies (or concepts). With the semantic types Animal and Person (and only those) the similarity-based \( \text{HisM} \) relation is most important, whereas conceptualizations that are goal-oriented – indicated by the \( \text{HtoM} \) relation – are most frequent with Advanced technologies (and concepts), but also encountered with other semantic types (albeit only rarely with Body parts and Natural phenomena).

Conceptualization of an entity in terms of its contents (\( \text{MinH} \)) is considerably less common than the inverse and never the dominant form; it is found most often with

\[^5\] There are five of these in the database: Gurindji, Puyuma, Selice Romani, Datooga and Tuwari.
Semantic types that denote Basic and Advanced technologies (and concepts) and Locations, rarely with Body parts and never with Animals. As for source-based conceptualizations, they are mostly found with Persons (in particular, professions), Natural phenomena, and Advanced technologies (and concepts).

Figure 11: How high-level relations vary across semantic types

Similar patterns emerge with respect to semantic fields (Figure 12, the high-level equivalent of Figure 9). Whereas the low-level plot highlights similarities between Animals and Kinship, the new one reveals additional commonalities, in particular between The body, The physical world and Food and drink. In all of these, Hatcher’s ® predominates: there is a tendency for conceptualizations where (to quote Hatcher) the target concept, B, “is somehow, to some extent, contained, comprehended in” the modifying concept, A.

In sum, and referring back to the notion of roles introduced in Table 6, we see that containment (HinM) is particularly important for The body, Food and drink and The physical world; goal-oriented conceptualization (HtoM) for the Modern world; similarity (MisH) for Kinship and Animals; contents (MinH) for both Clothing and grooming and Social and political relations; and source-orientation (MtoH) for Agriculture and vegetation.
Semantic relations in binominal lexemes: A cross-linguistic survey

7 Summary and further work

In this chapter I started out by providing a brief overview of previous studies on semantic relations in compounding. I then described in some detail the systems developed by Bourque and Hatcher that I have harnessed in the present study. Bourque’s classification was revised and extended with two new relations, CONTAINMENT and DIRECTION, for a total of 29 relations, 12 reversible and five unidirectional. Hatcher’s classification was also revised by the addition of a fifth high-level relation, SIMILARITY, in order to cover appositional and coordinate compounds. The two revised classifications were then unified to create the two-tiered Hatcher-Bourque classification, and an Excel-based tool called the Bourquifier was developed to assist in the slippery task of classifying individual binominals. Both the Hatcher-Bourque classification and the Bourquifier are offered to the research community in order to promote more collaboration in the field of semantic relations than has hitherto been the case.

The analysis of the data provides insights into the ways in which humans tend to conceptualize the world. It suggests, contra Bauer, that MERONOMY and PURPOSE are far more widespread, and thus more important, than the LOCATION relation. Of the two types of MERONOMY – Basic (MER) and Reversed (MER-R) – the former is far more frequent than the latter, which indicates that the conceptualization of a complex meaning is much more likely to involve modification by the whole (or, more generally, the container) than modification by the parts (or, more generally, the contents). The Basic MERONOMY relation (MER) occurs most frequently with body parts and in the semantic field of agriculture and vegetation. It can express about one third of the 100 meanings used in this survey; it is found in all 106 languages of the sample; and it can be expressed using any one of the nine nominal modification strategies.
Bauer’s suggestion that the next most frequent type is where the head is made from the material in the modifier is also not supported out by the data: both PURPOSE and COORDINATION are much more common than COMPOSITION. PURPOSE is most often encountered in the semantic field Modern world to denote advanced technological concepts; it only occurs in 89 of the 106 languages, no doubt because some of the languages in the sample do not have words for concepts of that kind; significantly, the only morphosyntactic strategy that does not occur with this relation is the classifier strategy, *cls*, but this also the most sparsely populated of all strategies. COORDINATION is used primarily to denote animates of a certain age, gender or both; it is therefore not surprising that it occurs mostly in the domains of kinship, animals, agriculture and vegetation. Cases such as these account for over 90% of binominals that exhibit this relation, and once again, every morphosyntactic strategy is attested in the data.

The Basic LOCATION relation is the fourth most frequent type overall and occurs three times as often as the Reversed relation; in other words, it is more usual to conceptualize an object in terms of where it is located than what is located at, near or in it. It is found in almost all of the languages of the sample (97 out of 106) and can be expressed by any of the nine strategies. It is most often encountered in the fields of the natural world and basic technologies and concepts.

The other fairly frequent relations are Basic POSSESSION and Reversed COMPOSITION (the reversed form of POSSESSION is uncommon, and the Basic form of COMPOSITION does not occur in the data at all). The range of meanings they can express is limited: only 12% in each case; all the same, they can be expressed by any strategy.

Apart from COMP, every one of the 25 relations was found in the data, but some were very rare, in particular CAUS, SRC-R, TEMP-R, TAX-R. While these are fairly peripheral in noun-noun compounds (and their functional equivalents), they may be more common in other types of compounds, for example those in which the head or the modifier is an action-morph rather than a thing-morph (see the Introduction for the precise definition of binominal used in the present study).

The data for the low-level relations suggests that there is no overall correlation between morphosyntactic strategy and semantic relation: many relations are expressed by every strategy, most are expressed by almost every strategy, and those that are expressed by just a few strategies are those where the data is sparse. This impression is confirmed by the analysis of high-level relations: every one of the five relations of Hatcher2 are attested with every one of the nine morphosyntactic strategies, so we can state quite categorically that there is no such overall correlation. It is thus not the case some strategies are used to express some relations, while other strategies are used for other relations.

However, while this applies cross-linguistically, it does not mean that there are no such correlations within individual languages.

In fact, the opposite is the case: As I showed in Pepper (2010), the Cameroonian language Nizaa (ISO 639: sgi) uses left-headed and right-headed compounds for two distinct sets of relations. Zúñiga (2014) reports something similar for Mapudungun (ISO
639: arn), as does Atoyebi (2010) for Oko (ISO 639: oks). Bourque himself (p. 253) compares N N and N à N binominals in French and shows that the two constructions have very different profiles (for example, PURPOSE and USE account for 47.5% of all French N à N binominals in his database, but only 12.65% of his N N binominals). Some of the contributions in this volume start to address this issue for other languages, but there is much work to be done. That work would be much more productive if researchers were to adopt the same classification system, and that is the purpose of Hatcher-Bourque.

References


