DE GRUYTER MOUTON DOI 10.1515/phras-2021-0005 — YoP 2021; 12: 75–146

Gabrijela Buljan and Lea Maras

"Shall I (compare) compare thee?"

A corpus-based study of a special type of modification of English *as*-similes

Abstract: This paper presents the results of a corpus-based analysis of a special type of modification of the English (*as*) Adj *as* NP similes. The modification involves filling the property slot with a cognate noun-adjective compound, i.e., a compound adjective consisting of the original adjective and the noun representing the original source of comparison, and inserting a new source of comparison into the construction (*red as blood* vs *blood-red as a raw steak*). Our data come from three sources: the Corpus of Contemporary American English (COCA), the iWeb Corpus, and material published on the Google website. Using quantitative methods we first explored whether there is a relationship between various distributional and formal features of the "original" *as*-similes and their likelihood of exhibiting this type of modification behavior. We then performed a quantitative and qualitative analysis of the semantic and, to a lesser extent, discourse-related features of authentic examples of this type of modification. Our results indicate that while these modifications are not abundant, the *as*-similes that have been found to modify in this way are significantly different from the *as*-similes that have been found not to modify in this way, on a number of formal and distributional features. The analysis of the semantic and discourse-related features of the modifications themselves revealed (i) the typical semantic domains of the three nominal entities featured in the modified simile: the original source, the new source, and the target, including the semantic fit among the domains of those entities; (ii) the typical semantic domains of the properties for which the three nominal entities are compared, including the semantic fit among the domains of those properties; and (iii) the typical text varieties accommodating these modifications. The latter results confirm, and to some extent elaborate on, some earlier findings about the semantic and discourse-related profiles of similes at large.

Keywords: *as*-simile, modification, source of comparison, target of comparison, expressivity.

Gabrijela Buljan, University of Osijek, Croatia, gbuljan@ffos.hr
Lea Maras, University of Osijek, Croatia, lmaras@ffos.hr

1 Introduction[1]

In her *Similes Dictionary*, editor, publisher, and theater critic Elyse Sommer notes:

> Its effectiveness for expressing thoughts more clearly and vividly makes the simile one of most widely used figures of speech in written and spoken English. Similes crop up in newspaper and magazine articles, fiction and nonfiction, dramas as well as daily conversations. The ones with the most zip tend to metamorphose into common expressions that are used unchanged or refreshed. In the age of sound bites and tweets they are more than ever timely and, to borrow an ever-popular simile, useful as a Swiss army knife for drawing pithy word sketches that are more robust than a single word and more spontaneous than a formal quote. (Sommer 2013: ix)

The *Similes Dictionary* features examples from sources as diverse as the Bible, Socrates, Shakespeare, movies, music, and TV shows. It boasts an impressive collection of 16 000 entries organized into nearly 1 300 thematic categories and an introduction with tips for creating novel similes. Sommer is aware that a simile dictionary can never be complete. However, for all its attention to novelty, the dictionary has missed one interesting convention-breaking pattern, which we like to think of as the simile equivalent of the Russian Matryoshka doll. Examples like *blood-red as a raw steak* (cf. *red as blood*) or *paper-thin as movie posters* (cf. *thin as paper*) involve a feat of creative modification, whereby the property slot of the simile is filled with a cognate noun-adjective compound (*blood-red*), i.e., a compound adjective consisting of the original adjective and the noun representing the original source of comparison, and a new source NP is inserted into the construction (*a raw steak*).[2] The two sources of comparison thus coexist within the fabric of the modified simile. Since the new source on the simile's outer margin tends to be structurally more complex than the original source merged with the adjective (see Section 4.3.1), the image of the Matryoshka doll – the set of wooden dolls of decreasing size placed one inside another – makes for a nifty, if only

1 This article is a substantially revised and expanded version of the second co-author's bachelor's thesis (Maras 2020). We are grateful to the editors and publishers for their patience and to the anonymous reviewers for their constructive criticism. Any shortcomings that may still exist are entirely our own. A special thanks goes out to the creative minds behind the fantastic examples discussed in this paper.

2 In these modifications, new sources are expressed as full NPs. Old sources are not full NPs since – being the first members of noun-adjective compounds – they lack determiners (cf. *sharp as* ***a razor*** vs ***razor****-sharp as shrapnel* [iWeb]). In fact, in our data, old sources within modified similes are overwhelmingly structurally simple and consist of a single noun (see Tab. 9 in Section 4.3.1). This is why we will continue to speak of adjective-*noun* compounds, the fusion of the original adjective and the original source *noun*, etc.

partial, metaphorical shorthand for this type of modification. Some examples are given in 1–5 below:

(1) *On a map showing levels of solar radiation, with the sunniest areas colored deep red, the kingdom is* ***as blood-red as a raw steak***. (COCA)[3]

(2) *For most Abdeen residents, however, the promises of the good life are* ***as paper-thin as movie posters***. (COCA)

(3) *Sadly, these celebrations also encourage an activity* ***as ice cold as a white Christmas*** *– the holiday breakup.* (iWeb)

(4) *It was before noon in northern Afghanistan, and the country felt* ***as empty and skull-white as a moon***. (COCA)

(5) *My experience with the HTC One is that this phone is* ***as rock solid as Mt. Everest****, in both build and operation*. (iWeb)

The type of modification can be schematized as follows:

I. Base-form simile	X_{NP} ... [(*as*) Adj *as* Y_{NP}] X... *(as) red as blood* X_{NP} = target of comparison Adj = original property designation Y_{NP} = original source of comparison
II. Modified simile	X_{NP} ... [(*as*) Y_n-Adj *as* Y_{NP}'] X ... *(as) blood-red as a raw steak* X_{NP} = target of comparison Y_n-Adj = original property designation fused with the Y_{NP} noun (CNAC) Y_{NP}' = new source of comparison

Such modifications are still rare (see Section 4.1). They are most likely one-off usage events or "occasional variants" (Langlotz 2006: 199) created for specific discursive

3 For space reasons, we shall indicate the source corpus only for the numbered examples separated off from the main text and for the creative similes in running text. See Appendix for references to the examples quoted from Google, and go to COCA (http://corpus.byu.edu/coca/) and iWeb (https://www.english-corpora.org/iWeb/) for detail about the examples quoted from the two corpora. Neither the authors nor the publisher claim any right over any copyrighted content quoted in this article. Every effort has been made to identify and trace copyright holders for material quoted from Google and we gratefully acknowledge the permissions granted.

purposes and are unlikely to become part of the entrenched phraseolexicon (cf. Naciscione 2010: 42). Still, as a delightful manifestation of the phenomenon of renewal of language expressivity,[4] they deserve a closer look. The goal of this paper is to examine this type of modification more closely by focusing on the formal and distributional features of various aspects of the original *as*-similes that may be related to this type of modification behavior, as well as on the semantic and, to a lesser extent, discourse-related features of the modifications themselves.

The plan of the article is as follows. Section 2 explains the background and the theoretical framework for the study. The methodology is detailed in Section 3 and the results are presented and discussed in Section 4. Section 5 lays out some conclusions.

2 Background

Similes have been studied since antiquity, but there is no agreement yet on the exact extension of the term. There is, for one, a distinction between stock similes in a language's phraseological system and similes as figures/tropes (cf. Omazić 2002; Wikberg 2008: 128). According to Omazić (2002: 101), similes as tropes are a feature of online language production, whereas similes as phraseological units (PUs) have become conventionalized, partly desemanticized, and are reproduced as units. Omazić acknowledges that (non)compositionality ("desemanticization") and immutability are gradient notions and that similes as PUs may be fully compositional (also Omazić 2015: 19) and accessible to modification.[5] No categorical distinction can be drawn between similes as figures and phraseological similes on the criterion of conventionality either. Conventionalization is a gradual process that may lead any novel simile to entrenchment through frequent use in a community of speakers. Hao and Veale (2010: 643) call attention to the grey area between creative one-off similes and the stock similes from printed sources. Many similes are common enough to be familiar but not so common as to be considered a part of the phraseolexicon. As a matter of principle, then, we make no categorical distinction between the two types of similes. We acknowledge, though, that

4 Renewal of expressivity involves phenomena like morphosyntactic change (double superlatives in *mostest*), lexical coinages (*fantabulous* < *fantastic* and *fabulous*), word formation (*safety squints* 'improvised eye protection by squinting the eyelids thus lowering the chance of injury to the eye', a compound created by analogy from *safety goggles* [*AVE Dictionary*, Accessed 29 April 2020 https://avedictionary.com/safety-squints/]), semantic change, e.g. *filthy* (in *filthy rich*) bleaching down to a degree intensifier, etc. (cf. Miller 2014: Ch. 8).

5 "The best way of accommodating such a phenomenon [PUs] is a scale. This indicates that we have to deal with different scales of complex features along which the members are placed based on the *intensity* of their specific features [...]" (Omazić 2015: 22–23).

particular expressions sit at various places on the continuum of conventionality – which is why stock similes like *hard as nails* may appear different (more "formulaic") than completely novel similes ("figures" or "tropes" in the sense described above) like *slippery as a kitten in its birth-sac* (COCA).

We are concerned here with renewal of expressivity, and since there can ever only be a need for renewal if there had been a perceived weakening or loss of expressivity, our study will mainly deal with corpus-retrieved similes that do have a sufficiently recognizable, conventional form and may be loosely regarded as instantiations of established units of the English phraseological system (see Methodology, Section 3).

A conceptually more complex issue is whether and how similes differ from similar forms of analogical reasoning, viz. metaphor and literal comparison. This question has been debated for centuries in rhetoric, philosophy, linguistics, and psychology (Aristotle, Quintilian, and Cicero, as cited in Fogelin 2011: 27; Aisenman 1999; Barnden 2012, 2015; Bredin 1998; Carston and Wearing 2011; Chiappe and Kennedy 2000, 2001; Gentner and Bowdle 2001; Glucksberg 2001; Israel et al. 2004; Lakoff and Johnson 1980; Miller 1993 [1979]; Ortony 1993 [1979]; Tirrell 1991; Todd and Clarke 1999; Utsumi 2011). While we agree that the issue is central, this study is not meant to contribute to this debate. We discuss the distinction between simile and literal comparison insofar as necessary to introduce our view of similes and explain our principles of data selection.

2.1 On simile (vs literal comparison)

We share Israel et al.'s (2004: 125) understanding of similes as a category subsuming a range of explicit figurative comparative constructions, which in English include but are not limited to those with the prepositions/conjunctions *like* and *as*.[6] Other properties proposed as criterial for distinguishing similes, especially from literal comparison, include:

a. the source and target entities are sufficiently unlike (cf. Miller 1993 [1979]: 373; Ortony 1993 [1979]);
b. the source entity is a model/representative/paragon of the property, i.e., exhibits the property to a high degree (Barnden 2015; Israel et al. 2004; Moon 2008; Wikberg 2008);

6 Israel et al. (2004: 125) define similes as any construction that invites a comparison of two very different entities, arguing that any construction that can express literal comparison can also be used as a simile, e.g. *The duchess – you've seen her portrait ... sir, it no more approached her than a weed comes up to a rose*, etc.

c. the source NP is non-referring; it is used attributively as a type representing the conceptual background for the description/evaluation of the referential target (Bredin 1998; Wikberg 2008: 133).

To summarize, in similes the target entity is described or evaluated by comparison with another "sufficiently unlike" source entity. The source entity is construed as a model of the property concerned, and the speaker intends the source NP to be interpreted attributively, rather than referentially, i.e. as representing the conceptual background for the description/evaluation of the referential target. None of these properties stand if construed as based on discrete binary oppositions.

A key semantic property distinguishing similes, e.g., *This lawyer is like a shark*, from literal comparisons like *A barracuda is like a shark* (examples from Barnden 2015: 49) is the figurative feel of similes that comes from mentally comparing "the unlike" (cf. Miller 1993 [1979]: 373). However, disparity/similarity are non-binary notions and are subject to construal (Israel et al. 2004: 126; Wikberg 2008: 127). Luckily, this can be accommodated into a framework like cognitive linguistics, which is comfortable with fuzzy categories and encyclopedic semantics (we return to this shortly).

Further, whereas in literal comparison source–target associations are not meant to go beyond what is stated, e.g. *He is as tall as his mother* (it is not implied that the mother is particularly tall), in simile, targets are associated with sources with intense features (Israel et al. 2004: 126–127). The sources exhibit the property to an extreme degree, allowing the speaker to convey her "superlative evaluation of the target" (Israel et al. 2004: 128). Sources in similes involve entities construable as a model, a representative (Hao and Veale 2010: 642; cf. also Ortony 1993 [1979]), or a paragon for the property in question[7] (Israel et al. 2004: 127), e.g. *he's thin as a rake* or even *She's as thin as Twiggy*. In Barnden (2015), deploying sources with intense features into similes (and metaphors) is called "source-based exaggeration". This feature needs to complement the "unlikeness" criterion if a construction of comparison is to be interpreted as prototypical. A non-exaggerative comparison like (of plutonium) *shaped into a ball more or less*

7 This is not defeated by counterexamples like ironic *clear as mud*, where *mud* is an anti-paragon for clarity. Israel et al. (2004: 127) claim that, because similes by definition feature entities acting as paragons, poetic considerations like alliteration or rhyme may force semantically unmotivated sources; e.g. *cool as a cucumber*, *fine as wine*. It is tempting to recast this in Construction Grammar terms and say that integrating such similes with licensing constructions involves coercing the incompatible source concept into the construction and the construction filling in the missing component of exaggeration (Goldberg 1995; Michaelis 2003, 2004). For some construction-based approaches to the study of PUs see Ivorra Ordines (2022) and Corpas Pastor (2021).

as big as a grapefruit (COCA) does compare two unlike entities, but with no superlative evaluation of the target, it is a step closer to literal comparison.

Finally, similes may also differ from literal comparisons interpretively. From his philosophical angle, Bredin (1998: 74–75) draws a distinction between similes and what he calls "ordinary" (i.e. literal) comparisons by claiming that similes are *predicative comparisons* in which the predicate describes the subject, while literal comparisons are *symmetrical comparisons* with referentially independent subjects and predicates. Paraphrasing Bredin, in a predicative comparison, by asserting or denying the target's likeness to something else we are enriching our knowledge of the target. In a symmetrical comparison, the target and source refer independently to two different things, whose likeness is being asserted or denied. This would seem like a sound basis for the distinction. However, in his plea for a more explicit examination in psychological processing studies of how similes (*My lawyer is like a shark/ like a shark I saw at the zoo yesterday*) and be-form metaphors (*My lawyer is a shark*) link up with referential vs predicative interpretation, Barnden claims that "in the be-form metaphor case, *but not in the simile case*, a very strong default is to take a simple indefinite source term such as a shark 'predicatively', and not to even consider the referential possibility other than in highly exceptional circumstances" (2012: 274 [emphasis ours]). Theoretically, this leaves room for similes with referentially specific sources and further blurs the boundary between simile and literal comparison.

Furthermore, referentiality itself is a tortured linguistic term. With Givón (2001), we assume that referentiality is not a logical relationship between referring expressions (NPs) and entities existing in the Real World. It is a matter of mapping from linguistic expression to entities in the Universe of Discourse. The Universe of Discourse is a realm "established by a particular speaker who then intends entities in it to either refer or not refer" (Givón 2001: 439). Thus, *a horse* and *it* are meant to refer to an actual horse in the Universe of Discourse in *She's looking for a horse; it escaped last Friday* and are only meant as the non-referential type in *She's looking for a horse; it had better be white* (Givón 2001: 439). Referentiality is also not a binary notion. There is a continuum between clearly referring and clearly non-referring interpretations based on how strong the speaker's intention is to refer to a specific individual (Givón 2001: 449). In *He bought books* Past Tense reference (realis) suggests a referential interpretation of *books*, but Plurality imparts the irrelevance of the individual reference of *books*. How does this inform our view of similes? While it may well be that most similes involve non-referring uses of source NPs, non-referentiality cannot be assumed or even read off from the grammatical coding of sources; rather, we may speak of degrees of referential strength, which is a combination of the grammatical devices employed on source NPs and speaker's intent. To give one example, grammatically definite NPs tend to be interpreted referentially (Givón 2001: 441) – and we may reasonably assume the same applies to proper nouns, like *Twiggy* in *She's as thin as Twiggy*. But, although the speaker

arguably intends to refer to the individual by that name, Twiggy is actually invited into the Universe of Discourse to stand as the paragon of thinness, to represent the type "extremely thin people", and is not quite referential. By recruiting this knowledge, i.e., by asserting the target's likeness to Twiggy, we are enriching our knowledge of the target, leaving the source entity in the conceptual background.

Summing up, with Barnden (2015: 49) and others (Carston and Wearing 2011: 297–298; Israel et al. 2004: 126), and in line with our cognitive-linguistic orientation, we assume a non-discrete distinction between similes and literal comparison.[8] The simile is a fuzzy category (Lakoff 1987; Rosch and Mervis 1975; Rosch 1977), whose prototype is represented by any construction of comparison which is/where:

1. figurative in the sense that the source and the target are construed as conceptually distinct entities (cf. human target vs environment source in *He is deaf as a stone*); they are even more figurative if the adjective itself is understood figuratively (*hard* for 'unfeeling' in *hard as stone*);
2. the source concept represents the model or paragon for the property (a status defined contextually, e.g. when a person is described as *subtle as a cockroach crawling across a white rug* [COCA]);
3. the source concept represents a non-referential type, intended to be interpreted as conceptual background relative to which the referential target entity is understood or evaluated;
4. connected to 3, the act of comparison is asymmetrical because the target and source of comparison differ in salience. In Langacker (1987: Ch. 3.1.2), comparison is understood broadly as a fundamental, complex cognitive event occurring in a variety of domains. It allows novel experience to be interpreted with reference to old experience, and therefore implies an inherent asymmetry between the old experience as the standard, and the new experience as the target of comparison. According to Langacker (2008: 58), "[t]he categorizing structure lies in the background, taken for granted as a preestablished basis for assessment, while the target is in the foreground of awareness as the structure being observed and assessed".

To these, we add two properties to capture the difference between similes perceived as familiar/formulaic and thus as PUs, and those perceived as novel, on-line occurrences:

5. similes become conventionalized in a community of speakers and perceived as part of the phraseolexicon;
6. similes are most likely processed as units in their canonical, unmodified form.[9]

8 Barnden (2015) uses the term "likeness statement" to avoid choosing between the two terms.

9 The problem of processing similes is too complex to explore here. For some detail see Barnden (2015).

A construction of comparison meeting all these criteria is the prototype of the simile category ([*man*] *hard as a stone*). Some similes are less prototypical either because a property is present to a lesser degree (e.g. conceptual distinctness) or because it is absent (e.g. non-referentiality of the source). Criteria 5 and 6 will not be discussed further. More comments on conceptual distinctness (criterion 1) are in order because it connects to criteria 2 (paragon-status of sources), 3 (non-referentiality of sources), and 4 (difference in salience in the source-target pair).

Conceptual distinctness is a fuzzy notion and has less to do with the ontological status of entities designated by the lexical items (e.g. human vs environment) than with the distinctness of the conceptual domains relative to which those entities are construed in context.[10] For instance, two human entities may be construed as sufficiently distinct if one is a non-entity from the conceptual domain of mundane, ordinary life and the other a cultural icon construed against the razzle-dazzle of the world of the entertainment industry (*my sister* and *Twiggy*).[11] In cognitive linguistics, the meaning of linguistic expressions is characterized relative to conceptual domains and there is a whole matrix of domains from which any can be recruited for the contextually fitting understanding of the linguistic expression. According to Langacker (1987: 158–161), a domain is central for the interpretation of a linguistic expression if it meets four criteria, of which the first three are pertinent here. A central domain is (a) *characteristic* for the designation of the linguistic unit, i.e. unique to the class of designated entities and thus sufficient to identify a class member; (b) *generic*, i.e. the knowledge is not restricted to specific exemplars but applies to whole classes of entities, (c) *conventional*, i.e. it represents knowledge that is shared within a community of speakers[12] and (d) *intrinsic*, i.e. it makes no essential reference to external entities. When it comes to source entities, understanding similes like *hard as stone* or *busy as a bee* involves accessing conventionally shared knowledge about classes of entities like natural objects and animals.

10 Cf. Israel et al. (2004: 126): "[l]iteral comparison involves entities which evoke similar domain matrices, but which may differ in their specifications within one or more domains. Figurative comparison, on the other hand, [...] involves the alignment of concepts with very different domain matrices".

11 In Hao and Veale's (2010: 642) corpus study of 20 299 ironic *about as*-similes, 12% of similes included proper-named source NPs like *Paris Hilton*, *Michael Moore*, etc. This means that for them comparisons with sources quite like the targets were still counted as similes as long as the sources were highly representative of the property concerned.

12 Conventionality as an aspect of centrality of conceptual domains in the domain matrix of a linguistic expression must be distinguished from the conventionality of a simile as a feature of the simile prototype (cf. criterion 5 above: similes become conventionalized in a community of speakers and perceived as part of the phraseolexicon).

Understanding a simile like *thin as Twiggy* in turn, is based on the conceptual domain representing conventional knowledge about a specific exemplar, and only by extension about the generic category the exemplar represents. Finally, for semantically motivated similes, the knowledge recruited from the domains evoked by source NPs will be more or less characteristic (even if not unique to them), i.e. it will allow the construal of sources as typical hosts for the "extreme" properties involved in comparison. Hardness is not a unique property of stones nor is diligence a property unique to bees, but this knowledge is sufficiently characteristic of members of those classes. The bottom line is, the paragon nature of the source entity (prototypicality criterion 2) is part of what is characteristic knowledge about the source entity. The typically non-referential nature of sources (prototypicality criterion 3) resonates with genericness as the criterion for domain centrality; generic knowledge about sources is knowledge that pertains to the whole class/type and this general knowledge, accessed through source NPs, is deployed to characterize rather than to refer to specific entities known to the utterer. Finally, the difference in saliency (prototypicality criterion 4) between target and source entities stems from the function of source entities and their domains in acts of comparison. They provide background knowledge that allows for a target's more vivid description or evaluation.

In (6a–e) below we list some modifications of the similes *thin as (a) stick/tissue/paper*, proposing a loosely linear progression from the most prototypical *as*-simile in 6a to the completely literal comparison in 6e. We ignore here the prototypicality criteria 5 and 6, since none of these creative examples are likely to be perceived as conventional or to be processed holistically. For easier reference, all examples include subscripts for the new source (S'), the target (T), and the property (P):

(6) a. *For most Abdeen residents, however,* ***the promises of the good life$_T$*** *are* ***as paper thin$_P$ as movie posters$_{S'}$*** (COCA)

S'-T very distinct, P figurative, S' non-referential, generic knowledge, S' paragon, S'<<T salience.

b. ***The storyline$_T$*** *may be* ***as stick thin$_P$ as Twiggy$_{S'}$*** [...] (Google)

S'-T very distinct, P figurative, S' somewhat referential, specific (unique) knowledge, S' paragon, S'<T salience

c. *More than anything, I wanted to have such a life, to have my mailbox jammed with* ***letters$_T$ as blue and tissue-thin$_P$ as my grandmother's vein-colored hands$_{S'}$*** (Google)

S'-T very distinct, P literal, S' referential, specific knowledge, S' paragon, S'<T salience

d. *At 6 feet 9 and 185 pounds,* ***Phaler*** $_{T}$ *is* ***as stick-thin*** $_{P}$ ***as a runway model*** $_{S}$, (Google)

S'-T somewhat distinct, P literal, S' non-referential, generic, S' paragon, S'<<T salience

e. *It wasn't that* $\boldsymbol{I}_{T}$ *wanted to be* ***as stick-thin*** $_{P}$ ***as Twiggy*** $_{S}$, (Google)

S'-T somewhat distinct, P literal, S' somewhat referential, specific (unique) knowledge, S' paragon, S'<T salience

f. ***She*** $_{T}$*'s* ***as stick-thin*** $_{P}$ ***as her mom*** $_{S}$, (invented example)

S'-T insufficiently distinct, P literal, S' referential, specific (unique) knowledge, S' not a paragon, S'≈T salience

2.2 PU and simile modifications

PU modifications can be defined as deliberate, creative, and idiosyncratic ad hoc changes of the canonical PU structure and/or meaning which produce semantic, stylistic, affective, or pragmatic effects in discourse (Omazić 2015: 35). According to Omazić, PU modification rests on finding harmony between a set of contradictions involving some of the definitional properties of PUs: familiarity vs novelty; entrenchment vs creativity; stability vs variability; figurative vs literal meaning; automated vs creative production (2015: 36). There are now a number of corpus, discourse, and cognitive studies proving that PUs are variable and quite abundant in corpora (Geeraert 2016; Langlotz 2006; Naciscione 2010; Omazić 2002, 2003, 2007, 2015; Vo 2011). It has also been found that modifications occur with any type of PU, even semantically opaque ones (Langlotz 2006; Naciscione 2010; Omazić 2003).

Interestingly, according to Moon (1998), similes in particular tend to be lexicalized. Wikberg (2008: 129) attributes the lack of creativity to the relatively low frequency of the *as*-simile in the BNC corpus he explored.[13] Nevertheless, Wikberg does acknowledge some innovations: "However, *white as snow* appears once as *white as newly fallen snow*, which might be a more realistic

13 According to Moon (1998), similes (including proverbs and most idioms) generally occur very rarely, with frequencies of less than one token per million words. However, this study did not consider the distribution of fixed expressions across different text types (Omazić 2015: 31–32). Other studies also report a relative scarcity of all classes of PUs in any register, but they note uneven distributions of PUs across registers (e.g. Biber et al. 1999). For instance, Moon (2001) found that in the Bank of English corpus, print journalism featured most idioms, tabloid newspapers included more than broadsheets and idioms were avoided in expository non-fiction and academic writing. For more detail see Omazić (2015: Ch. 3).

version today" (Wikberg 2008: 135). Contrary to Moon (1998), Omazić (2015) found in her BNC-based study of PU modification that "idioms of comparison" are among the most frequently modified PUs. In Omazić's work, "idioms of comparison" involve constructions of comparison involving *as* (i.e. *as*-similes) and *like*. Hao and Veale (2010) examined a corpus of authentic ironic similes like *as welcome as a root canal without anesthesia*, where sources express properties ironically opposed to those coded in the adjectives. Veale finds the syntactic form of *as*-similes to be a scaffolding for creativity that can be exploited "with remarkable freedom":

> There is something appealingly democratic and unpretentious about similes. Not only are they pervasive in language, they are at home in any register of speech and any genre of text, from tabloid newspapers to romantic poetry (Fishelov 1992). Conveniently, most languages provide a wealth of pre-fabricated similes that are as well-known to native speakers as the adjectival features they serve to exemplify. [...] But just as importantly, languages like English make it easy for speakers to mint their own similes on the fly, by imposing low barriers to creation. (Veale 2012: 329)

There are, therefore, already a number of corpus-based studies exploring variation/modification in English *as*-similes. But to the best of our knowledge, no study has explored the specific type of modification examined here. We will rely on aspects of a model of PU modification proposed by Omazić. In her extensive work on the topic (2002, 2003, 2007, 2015; Omazić and Čačija 2020), Omazić sets up a dynamic model that accounts for principles and constraints on PU modification (cf. also Langlotz 2006). A closer examination is beyond the scope of this paper, but we lay out below the four elements of the model and briefly explain those that will be pertinent here:

(1) Personal constraints: individual propensity for PU modification based on personality type, flexibility with language, etc.
(2) Modification principles and constraints (inspired by Blending theory [Fauconnier and Turner 2002]), which include:
    (a) constitutive principles of PU modification (essentially a roadmap for how to arrive from the canonical to the modified PU, involving notions like inputs, projections, emergent structure, etc.);
    (b) modification principles (semantic, grammatical, and lexical constraints on what is possible; they set limits to how far one can go in modifying PUs);
    (c) vital relations (modifications can be successful if they rest on salient, vital relations between elements of the base-form and the new context, like change, time, space, cause-effect, etc.).

In brief, the principles and constraints under (2) ensure that the modified PU is recognizable as being based on the canonical form, that it is grammatically coherent, that it is interpretable in the new context, and that users can grasp its relevance.

(3) Contextual constraints: type of medium, type of discourse and genre, etc.

(4) Functional constraints: a PU modification needs to fulfil a purpose to be effective since purposeful modifications create new meaning, provoke, impress, entertain, etc.

This study is about a type of simile modification that is clearly purposeful, possibly highly author-style/personality driven, and discourse sensitive. It is paradigmatic (featuring lexical change), and involves syntagmatic structural realignment, if merging the original source and property lexemes can be seen as part of the modification process itself. If indeed it can, we could argue that our modifications weaken Moon's (2008) claim that, because of the basic structure of *as*-similes, there is little room for syntagmatic variation. The type of modification examined here is also easily recognizable and interpretable since the "old" source of comparison is not discarded after the new one is introduced. Assuming that the innovator must have had some familiarity with the original simile to bend it, we may consider the original simile as part of the established phraseolexion, regardless of how close or distant from the *as*-simile prototype it may be. That being the case, the general principles and constraints applicable to modifying PUs are likely to play a role here too. The modification constraints that may be specific to this type of modification, like the syllabic size of adjectives and original source nouns, will be addressed in Section 4.2. We will not explore the functional or author/personal constraints since these mainly involve processing issues concerning the speaker's motivation/production and the hearer's reception of modifications. Neither will we work with the global constitutive principles or Blending theory-specific constructs like "vital relations". We will use quantitative and qualitative methods to examine:

(A) the relationship between some distributional and formal features of what we will define below as base-form similes and the Matryoshka-type modification/Matryoshka-type similes (from here on M-modification or M-similes);[14] and

(B) the semantic and discourse-related features of the M-similes themselves.

14 Where appropriate, we will use derivatives thereof, like the verb "to M-modify", etc.

A. Distributional and formal features of base-form similes

The following distributional and formal features of base-form similes will be explored for their association with M-modification:

1. The overall token frequency of what we call "base-form similes" (B-similes). Base-form similes are here defined as all our corpus-retrieved similes that fit the [X_{NP} … (*as*) Adj *as* Y_{NP}] schema and act as a potential scaffolding for M-modification.[15] Note that B-similes are defined in purely structural terms, which means they include canonical similes and those that already count as modifications because of lexical substitutions in the source or property slots e.g. *white as snow* vs *white as the moon* (COCA) vs *white as the frost balls on a Christmas tree* (COCA); *silent as a tomb* vs *quiet as a tomb*.[16] This allows us to avoid making arbitrary decisions as to which B-similes are part of the canon. Instead, we merely register the objective measure of frequency to establish if there is a difference in token frequency between the two subsets of B-similes, viz. the B-similes that have been found to M-modify and the B-similes that have been found not to M-modify. From here on, we will be referring to these as "two subsets of B-similes". The reasonable assumption is that B-similes with higher token frequencies may have lost their luster and may be generally more prone to M-modification than B-similes of lower token frequencies. Some support for this comes from claims that frequency of exposure breeds familiarity and diminishes information value: "Core use is not always best suited for the purpose of verbalizing a person's feelings, thoughts, and experiences, nor does it adequately convey the meaning which the discourse situation requires. *As a frequently recurrent item*, core use sounds more common and conveys less information in comparison with the infrequent unusual stylistic instantiations" (Naciscione 2010: 39, [emphasis ours]).[17]

---

15 Our "base-form similes" do not correspond to Naciscione's "base form" of a PU, defined as a decontextualized archetypical form of a PU stored in long-term memory (2010: 8). Base-form similes are here defined in purely structural terms, but all are corpus-retrieved instances of actual use. Some may be considered instances of core use – constituting the "perfect" example of a PU in context (Naciscione 2010: 35) like *white as snow*, but many fit Naciscione's "instantial stylistic uses" – unique stylistic applications of a PU in discourse (Naciscione 2010: 43), e.g. (*an idea*) *fresh as disco music and mullets* (COCA). For more discussion of the notion of *base form* in studies of PU variation, see Petrova (2011: Ch. 4.4).

16 Moon (1998) found that many PUs do not have one standard form but multiple equally institutionalized variants.

17 Clearly, familiarity cannot be equated with frequency of exposure. Frequency is an objective text-based measure; familiarity is a knowledge-based and subject-dependent construct (cf. Grzybek and Chlosta 2009; Mieder 1989).

2. Token frequency of cognate noun-adjective compounds (CNACs) (e.g. *blood-red*). The reasonable assumption is that preexisting CNACs may have precipitated M-modification by strengthening the entrenchment, and weakening the expressivity of B-similes, i.e. of the original source – property pairs. This assumption could be supported by Norrick's (1987) claim about striking semantic similarities between the meaning relations expressed in similes and CNACs (see also Moon 2008: 7; for a different view see Novoselec and Parizoska 2012); and by that of Moon (2008: 32) who, arguing a different point, submits that highly frequent CNACs may make the simile sound familiar even when the simile is not very frequent itself.
3. The family size of alternatives. This is defined as lexical variation of sources, i.e. the sum of B-similes with similar or identical meanings where the same property is matched with a number of different source NPs, e.g. *white as* X = *white as a sheet/snow/a lily/the moon/bone*. We will refer to such semi-schematic similes with open source slots as "B-simile templates", and sometimes, for simplicity, only as "templates".[18] The reasonable assumption is that B-simile templates with bigger families of sources already show a tendency to modify and may also be prone to M-modification.[19]
4. The frequency of source NPs in general (i.e. in the corpora used), and the frequency of their use as sources of comparison in the B-simile databases. Whether or not the token frequency of each complete B-simile, i.e. its "overuse", has any bearing on the existence of its M-modification, higher token frequencies of B-simile components, in particular source NPs, may have rendered those sources too worn out to be useful. The reasonable assumption is that B-similes with source NPs that are generally more frequent in corpora and/or more frequent in our B-simile databases are more likely to M-modify. This assumption is partly justified by Mancuso and Laudana (2019), who report a positive correlation between the frequency of manipulated idioms and the frequency of lemmas: idioms which more often occur in corpora in a manipulated form involve frequent words. It remains to be seen whether the same applies to M-modification.

18 B-simile templates are different from B-similes, since different B-similes may fit the same B-simile template; e.g. *as white as* X can be matched with specific B-similes like *white as a bone*, *white as chalk*, etc. This distinction must be borne in mind when interpreting quantitative results in Section 4.2.

19 Moon (2008) claims that many similes fall into sets or clusters where members of the set have similar or identical meanings and some of the lexical elements in common but other lexical elements that vary. She illustrates this with similes like *thin as a beanpole*, *thin as a lath*, *thin as a rail*, *thin as a rake*, etc.

5. The syllabic size of adjectives in the two subsets of B-similes. Since the adjective needs to become the second CNAC member in the M-simile, there may be restrictions on its size. The assumption is that B-similes with shorter adjectives are more likely to be associated with M-modification than those with longer adjectives. Admittedly, Moon (2008) established that the majority of adjectives in her corpus of similes were monosyllabic anyway. It remains to be checked whether this is the case in our databases, and whether there might still be a difference between the syllabic sizes of adjectives in the two subsets of B-similes.
6. The syllabic size of source nouns within syntactically simple source NPs in the two subsets of B-similes. Since the source noun needs to become the first CNAC member in the M-simile, the assumption is that M-modification is more likely with short source nouns.

Since we could not afford a semantic and discourse analysis of the entire B-simile databases, only M-simile tokens will be analyzed for their lexico-semantic and discourse-related features.

B. Lexico-semantic features of M-similes

B.1 Lexico-semantic features of old sources, new sources, and targets

7. The semantic domains of the old source entity, the new source entity, and the target entity (e.g. Environment, Tools, Animals, etc.). We also consider here: (a) whether M-similes show uneven affinities to particular lexemes from the same semantic domain for their old sources. For instance, although the Human Body domain is exploited more than some other domains, not all of its lexemes are (*blood* is common in M-modification: *blood-red as*... but *lymph* is not used at all); (b) the syntactic complexity of new and old sources (cf. the simple old source and the complex new source in *as bone-dry as a bale of wastepaper compacted by feelings of helplessness* [COCA]).
8. The match between the semantic domains of old sources and targets, new sources and targets, old sources and new sources, and among all three nominal entities at once (e.g. no match in: *Anxiety*$_i$ *as flour*$_j$*-fine as sand from Aram*$_k$ [Google]; where $_i$ = Emotion/State/Disposition, $_j$ = Food and Drinks, $_k$ = Environment).

B.2 Lexico-semantic features of properties implicit in the three nominal entities

9. The semantic domain of the properties implicit in old sources, new sources, and targets (e.g. Temperature, Color, Textural Smoothness, etc.). We also consider here whether M-similes show a preference for particular adjectives from the same semantic domain. For instance, although the Color domain is heavily exploited, not all of its lexemes are (equally frequently) used (*white*, *black* and *red* are common in M-similes, e.g. *lily-white as Norway* [Google], *jet-black as Zoro's costume* [iWeb], *blood-red as a vampire* [COCA], *beige* is not found at all).
10. The match between the properties of old and new sources, old sources and targets, new sources and targets, and among the properties of all three entities at once (e.g. *a stranger*$_{i}$ ... *steel*$_{j}$ *cold as ice*$_{j}$ [Google]; where $_{i}$ = Emotion, $_{j}$ = Temperature)

Our semantic categories, i.e. domains, are inspired by Dixon (2005), Givón (2001), and Moon (2008). However, they should not be considered discrete since they are also a matter of construal. For instance, *lobster* could be categorized under both Animal, and Food and Drinks, but it will be tagged as Food in *lobster-red as a sunburn* (Google), since lobsters only turn red after cooking.

C. Discourse-related features of M-similes

11. Which text varieties are typical hosts for M-modifications?
    Because of the complex and ever-changing landscape of register/genre/text types, especially those inherently Internet-based (e.g. blogs, cf. Giltrow and Stein 2009), the categories we will use in addressing this context-based feature of M-modification should be seen as our tentative proposals pending a more dedicated discourse-oriented analysis.

3 Methodology

To collect as many examples of M-similes as possible we used three sources. We started with the Corpus of Contemporary American English (COCA) since this is the corpus where we first came across some M-similes by a lucky happenstance, while working on a different project. The decision to go beyond COCA and include the iWeb Corpus and Google as additional sources was motivated by

our assumption that the two web-based resources, especially the virtually unconstrained content indexed by Google, would be likely to deliver more instances of creative M-similes than COCA (see further in text).[20] For simplicity, we will refer to all three sources as "corpora". Since M-similes develop on the backs of B-similes, so to speak, our first step was to tap the three corpora for B-similes, and the B-similes were later checked for the presence of M-modification. Each set of B-similes obtained from the three corpora will be referred to as a "database", giving us three databases in total.

COCA returned the most results through automatic queries targeting the schematic structure of B-similes <_j *as* _nn> ('any [Adj *as* noun] sequence') and <_j *as* _a _nn> ('any [Adj *as* determiner noun] sequence'). Specifically, COCA delivered 13 000 results, of which 2 710 passed muster as B-similes after manual cleaning (see below). For unknown reasons, the automatic search of iWeb yielded only 255 results, which seriously misrepresents the presence of B-similes in that corpus. When probed manually, iWeb was found to have many more examples of B-similes, which did not figure among the 255 results. Google, in turn, does not have an automatic search feature comparable to that of COCA and iWeb, which made it impossible to mine B-similes straight from Google. For these reasons, COCA was used as the primary source of B-similes, and iWeb and Google were then searched manually for the same B-similes originally retrieved from COCA, yielding a total of 1 285 B-similes in iWeb and 2 579 B-similes in Google (see Tab. 2). There are obvious disadvantages to this derivative approach to collecting B-similes from iWeb and Google, but the method would have constituted a more serious problem if we had actually set out to compare data across the three corpora. While this would have been an endeavor worthy of effort, especially for assessing the reliability of corpus evidence in studying PUs,[21] it was not our goal. In this study, our

20 For comparison, COCA includes about 1 billion words from around 500 000 texts almost evenly divided between spoken, fiction, magazines, newspapers, academic journals, blogs, other web pages, and TV/Movie subtitles. iWeb contains around 14 billion words in approximately 22 million webpages from around 95 000 systematically chosen websites (see https://www.english-corpora.org/iWeb). Google's search engine, in turn, indexes billions of webpages alone (see https://en.wikipedia.org/wiki/Google).

21 Moon (2008: 26–27) claims that "corpora provide no real evidence for individual as-similes being particularly frequent [...] that a simile lexicon identified from one corpus is unlikely to map perfectly onto one identified from a different corpus, particularly with respect to the lowest frequency items". Moon also claims that "[r]eliance on corpus data in the investigation of idioms, proverbs, similes and the like has been criticized on the grounds that corpora do not provide an accurate representation of current language use, because they contain the wrong kind of data. Thus, for example, as-similes may not be as rare in English as corpus evidence suggests" (Moon 2008: 21).

primary objective was to check as many B-similes as possible for the existence of their M-modification in order to gain as much insight as possible into the nature of M-similes. On balance, both iWeb and especially Google proved to feature more tokens (Google also yielded more types) of M-similes than COCA, which is why we decided to include iWeb and Google data in our analyses. Moreover, it is hardly surprising that Google yielded the most M-similes since, unlike the balanced and systematically constructed iWeb, and especially COCA, Google's content is virtually unconstrained and is more likely to feature a considerable proportion of highly informal types of discourse (like tabloid newspapers, advertising, social networks, review/opinion articles, blogs) where similes and other PUs are more at home (Moon 2008: 22).

Returning now to COCA and our first step in the selection of B-similes. The total number of hits from the automatic search of COCA was 13 000, with 2 710 B-similes (types not tokens) remaining after manual cleaning.[22] We first summed up singular and plural variants, and minor spelling variants (*grey/gray*), then weeded out non-target structures, which included:

- literal comparisons (*Here, [...] you eat so you get strong as Mama*);[23]
- comparisons of proportions (*He was as American as he was Jewish*);
- comparisons involving more targets and properties (*He was as granite-willed as his forehead was monumental*);
- adjectives followed by temporal *as* prepositional phrases (*I was so happy as a child*);
- adjectives as parts of concession and temporal clauses introduced by *as* (*beautiful as she was, she was already twenty*; *it quickly changes to gleaming snow-white as the tree matures*);
- similes with reversed semantic components (*there's nothing as strong as the bond between a mother and a child*, where the target *the bond between a mother and a child* is found in the source slot and represents an instance of the highly generic category expressed in the target slot by indefinites like *nothing*);
- Adj *as hell* similes and those ending in *as shit, as fuck, as heck*. Moon (2008: 5) excludes such similes from her analysis on the grounds that the source NP

22 The actual number of B-similes in COCA is 2 707. The three "missing" cases refer to the M-similes that were picked up alongside B-similes by our initial automatic query. They include: *pissgrey as the underpants of the six graders in the dressing rooms of the swimming pool*; *rubber-red as a pencil eraser's*, and *skull-white as a moon*. Incidentally, they were not attested in COCA in their unmodified form.

23 We also ran a manual quality check of all tokens of B-similes obtained from COCA and iWeb to verify that they do indeed qualify as similes in the sense defined in this paper. We did not do the same for Google for reasons that will be explained later in the text.

is more of an "emphatic particle" (except in semantically motivated *hot as hell*).[24] We agree with Moon's assessment, but we do not lose sight of the fact that "emphasis" underlies all similes, including many others lacking semantic motivation. We discarded those examples for yet another reason. Surprisingly, the iWeb showed no results of *as hell* similes. Also, these similes virtually never featured any CNACs or M-modifications. Interestingly though, *as fuck*, *as hell*, and *as heck* were each attested once on Google as new sources. Here is just one example: *C'mon nigga that shit looks overcooked* [...] *The only time breast fluffs up like that when it's absolute* ***sand dry as fuck*** *in center*.

The B-similes obtained from each corpus then served as a baseline for identifying corresponding M-similes. This was done using a lexically specific query involving the hyphenated and non-hyphenated CNAC matching each B-simile followed by *as*, e.g. *blood-red as/blood red as*.

All tokens of M-similes obtained from the three corpora were combined for the analysis of their semantic and discourse-related features, except duplicates (i.e. the same linguistic example of an M-simile appearing in a different corpus). An analysis of tokens is necessary since different M-modifications of a single B-simile may have different lexical, semantic, and discourse-related features. However, identifying possibly distinctive semantic and discourse-related features of M-modification lies beyond the scope of this paper since this would require a painstaking token-based comparison of those features in the very large number of B-similes in our three databases that have, and those that have not been found to M-modify.

As for the formal and distributional features of M-similes, since they are common to all tokens of a single type, they were subject to a type-based quantitative analysis. This made it possible to address potentially distinctive formal and distributional features of those B-similes that have been found to M-modify, by comparing them to the B-similes not found to M-modify. Since distributions vary between COCA and iWeb, the distributional features were analyzed in each corpus separately, including, where possible, in data from Google. For this quantitative aspect of the analysis, we did not discard duplicate tokens since token counts measure the presence of a structure in a particular corpus, i.e. distributions are corpus-specific. At this point we should add an important cautionary note. Quantitative data from Google were only partly usable. Some searches returned extremely high counts with too many non-target structures to clean

24 Adams (1973: 97) claims the same for an even wider range of items as parts of noun-adjective compounds: "[i]t is easy to see how a first element in this kind of compound may become simply an intensifier, as in the following examples which are no longer understood as involving an explicit comparison [...] *dog-tired, stock still, stone deaf*".

manually; therefore, these analyses had to be abandoned (analyses of source NP or CNAC frequencies, see below). Finding specific B-similes (or M-similes for that matter) in Google was not an issue since enclosing the lexically specific search strings in double quotes should not allow non-target structures. Still, the token frequencies of many individual B-similes ran into the thousands, sometimes millions (and were only given as approximations), which made it impossible to run a manual quality check of all those results. Also, since many B-similes occurred with very high token frequencies, this resulted in a staggering token frequency total (see Tab. 2). The reader is invited to keep this in mind while reading the following sections. We should note though, that despite this drawback, our distributional analyses performed on the Google data gave results that were, in the main, consistent with those obtained from iWeb and COCA. And, of course, we should not forget Google's one redeeming quality – it remains unchallenged even by iWeb, but especially by COCA – in delivering many fantastic M-similes for the analysis of the lexico-semantic and discourse-related features of M-modification.

Finally, more rigorous statistical modelling, with variables related to form, distribution, meaning, and use being combined and assessed for their potential to predict M-modification, must be left for the future mainly because the target data are still comparatively too scarce.

4 Results and discussion

4.1 M-similes: general descriptives

Since M-similes are the main focus of this study, we first present in Table 1 general descriptives for M-similes found in the three corpora.

Type frequency of M-similes is the total number of different M-similes corresponding to the B-similes in our databases. *Token frequency of M-similes* indicates the sum of all tokens of M-similes corresponding to the B-similes in our databases. Thus, *blood-red as a raw steak* (COCA) and *lobster-red as a sunburn*

Tab. 1: M-similes: type frequency, token frequency, and type/token ratio

M-similes	Type frequency (TypF)	Token frequency (TokF)	Type/token ratio (T/T ratio)
iWeb	13	42	0.3
COCA	14	15	0.9
Google	139	280	0.5

(Google) count as two M-simile types, going back to *red as blood* and *red as lobster*, respectively. The two instances of M-modification of *red as blood* attested in COCA – *blood-red as a raw steak* and *blood-red as a vampire* – count as two tokens of one M-simile type. Note that we treat as tokens of a single M-simile type all the M-similes that share the old source and property (above *blood* and *red*), regardless of whether they also share the new source. In the *blood-red* M-similes, the two tokens do not share new sources (*a raw steak* vs *a vampire*), but in *pitch dark as night* (COCA) and *pitch-dark as the night* (COCA) they do. *The type/token ratio* is the averaged measure of dispersion of M-simile tokens across M-simile types. Values closer to 0 imply that a smaller number of M-simile types accounts for all attested M-simile tokens; values closer to 1 suggest greater variety, viz. the tokens of M-similes are spread out across more M-simile types. Note that this ratio needs to be interpreted relative to the overall number of types and tokens. Theoretically, where there is only one type and one instance of a target structure in a corpus, the T/T ratio would be the ideal 1, but in reality, there would be no variability. Where, however, there is at least some (albeit small) number of types and tokens (as is the case here in iWeb and COCA), then even within such small pools of examples, we may speak of lower or higher T/T ratios, relatively speaking. In our databases, iWeb has a low T/T ratio because many M-simile tokens (N = 19) belong to a single M-simile type, i.e. that based on the B-simile *clear as crystal*. For such instances we might tentatively propose that low T/T ratios suggest that new tokens of M-similes may (also) be built by analogy to familiar exemplars rather than from productive schemas like [(*as*) CNAC *as* NP].[25] All in all, with the exception of Google, the type and token frequencies of M-similes are quite low.

4.2 Distributional and formal features of B-similes & M-modification

4.2.1 Token frequencies of B-similes

Table 2 shows quantitative data necessary for the description and comparison of the two subsets of B-similes. Note first the very low proportion of the B-similes which were found to M-modify in all three databases: 0.5% (COCA), 1.0% (iWeb) and 5.4% (Google).

25 For more about schema-based and analogy-based accounts of building novel expressions see e.g. Langacker (1987: 445–447), also Bybee (2010) and Taylor (2012).

Tab. 2: Descriptive statistics for two subsets of B-similes: those that M-modify and those that do not M-modify (iWeb, COCA, Google)

		TypF	%	TokF	%	Mode	Min	Max	Mean Rank	Sum of Ranks
iWeb	**+M**	13	1.01	5 745	9.19	1*	1	1 408	1063.65	13 827.50
Bsim	**-M**	1 272	98.99	56 794	90.81	1	1	2 853	638.70	812 427.50
	Total	1285	100.0	62 539	100.0					
COCA	**+M**	14	0.52	177	2.22	0	0	29	1902.50	26 635.00
Bsim	**-M**	2 696	99.48	7 808	97.78	1	1	116	1352.66	3 646 770.00
	Total	2 710	100.0	7 985	100.0					
Google	**+M**	139	5.39	106 355 397	9.09	129 000	5	7 230 000	1819.04	252 846.00
Bsim	**-M**	2 440	94.61	1 063 642 806	90.91	1	1	28 800 000	1259.86	3 074 064.00
	Total	2 579	100.0	1 169 998 203	100.0					

* Multiple modes exist, the smallest value is shown; Bsim +M: B-similes with M-modification; Bsim -M: B-similes without M-modification

In iWeb, the most common token frequency (*mode*) of B-similes with or without M-modification is 1. In COCA, the mode of B-similes without M-modification is also 1, but is equal to 0 in B-similes with M-modification. The latter is due to the three "orphan" M-similes automatically extracted from COCA, which had no tokens of the "parent" B-simile (see footnote 22). Looking at iWeb and COCA data, we also see that the minimal token frequency of B-similes of 1 or even 0 is not a barrier to M-modification. Google stands out since the minimal token frequency of B-similes with M-modification is slightly higher (5), and the mode of B-similes with M-modification is considerably higher (129 000).[26] But notably, in all three corpora, B-similes without M-modification have much higher maximum token-frequency values than B-similes with M-modification, which suggests that M-modification is not an exclusive privilege of the most token-frequent B-similes. We conclude that the lowest frequency of B-similes is not a barrier to, nor is the very highest frequency a necessary condition for, M-modification. Now, whether this also implies that the occurrence of M-similes has statistically nothing to do with high token frequencies of B-similes, is a question that will be addressed shortly.

First, let us comment on some of the curious features of the data. It made no sense to report the mean token frequencies of either subset of B-similes since the standard deviation, kurtosis and skewness values (indicators of the normality of data distribution) showed considerable to extreme dispersion (and in some cases, accumulation) of data away from this central value. Looking at TypF data in Table 2, it stands out that there are vastly many more B-similes without M-modification that those with M-modification. But, on closer inspection of data, we confirmed our initial subjective impression that most B-similes in the subset of B-similes without M-modification are hapaxes or have low token frequency. In iWeb, non-modifying B-similes with tokens not exceeding 10 account for 802/1 272 cases (264 of which are hapaxes), that is 63.1% of the database. In COCA, they account for 2 582/2 696 cases (1 564 of which are hapaxes), which is 95.8% of the database, and in Google for 589/2 440 cases (152 of which are hapaxes), which represents 24.1% of the database. This explains why the subsets of B-similes with M-modification could show consistently higher central values, specifically the mean rank (as well as the unreported mean and median), than B-similes without M-modification, despite being generally less frequent than the latter. Whether this difference reaches statistical significance was tested with the Mann Whitney

26 In Google, all B-similes with M-modification had different frequencies, except for two B-similes, which registered the same token frequency of 129 000 (recall that Google frequencies were rounded and given as approximations).

U statistic.[27] In all three databases, we found a statistically significant difference in token frequencies between the two subsets of B-similes (iWeb: U = 2799.500, $p < .001$; COCA: U = 11214.000, $p = .003$; Google: U = 96044.000, $p < .001$). Thus, M-modification was proven to be more likely with B-similes of higher token frequencies, suggesting that "overuse" might be a factor contributing to M-modification. Still, we should not lose sight of the fact that the most common token frequency of the not-so-many B-similes with M-modification is still 1 or even zero, which means that high token frequency cannot be an absolute criterion for M-modification.

4.2.2 Cognate noun-adjective compounds (CNACs)

A related distributional feature hypothesized to be associated with M-modification is the co-existence of (frequent) cognate noun-adjective compounds (CNACs). The reasonable assumption was that CNACs may have precipitated M-modification by weakening the expressivity of original source – property pairs. This analysis is based on the token frequencies of CNACs in iWeb and COCA only (Tab. 3), since Google delivered very many results with many non-target items, such as object nouns followed by adjectives as object complements, e.g. X *washed the sheet white*.

In iWeb, the minimum frequency of CNACs among B-similes with M-modification is 124, which suggests that M-modification was only found with B-similes that had matching CNACs. Put differently, in iWeb, B-similes which had no CNACs did not have M-modification either. Upon closer inspection of data, we found that B-similes with no CNACs actually accounted for as many as 661 of 1 272 (52%) of all B-similes without M-modification (this can also be deduced from comparing the two columns representing the type frequency of B-similes and the type frequency of CNACs). Other B-similes without M-modification had CNACs of varying frequencies, but again, the bulk, i.e. 79.3% had CNACs with low token frequencies, not exceeding 10. On the other hand, B-similes with M-modification always had CNACs, with the minimum CNAC frequency being 124 and the maximum 46 616. In COCA, the situation was different. The minimum frequency of CNACs was 0 in both subsets of B-similes, and cases of B-similes with M-modification but without CNACs were actually the most common (3 out of 14 cases). However, it is noteworthy that the three cases of unattested CNACs for

27 Our databases consistently failed to show normal or equally-shaped distributions, as confimed by the Kolmogorov-Smirnov normality test, skewness and kurtosis values, and the visual inspection of histograms.

Tab. 3: Descriptive statistics for CNACs associated with two subsets of B-similes (iWeb and COCA)

		TypF (Bsim)	TypF (CNAC)	TokF (CNAC)	Mode (CNAC)	Min (CNAC)	Max (CNAC)	Mean Rank (CNAC)	Sum of Ranks (CNAC)
iWeb	**Bsim +M**	13	13	146 411	-*	124	46 616	1 266.46	16 464.00
	Bsim -M	1 272	611	108 265	0	0	10 351	636.63	809 791.00
	Total	1 285	624	254 676					
COCA	**Bsim +M**	14	11	2 354	0	0	433	2 360.00	33 040.00
	Bsim -M	2 696	416	6 024	0	0	476	1 350.28	3 640 365.00
	Total	2 710	427	8 378					

* All frequencies were registered only once

B-similes with M-modification were the CNACs of the three "orphan" M-similes from COCA, viz. *piss-gray*, *rubber-red*, and *skull-white*. In all other cases CNACs were present with frequencies ranging from 16 to the maximum 433. Concerning B-similes without M-modification, CNACs were absent in as many as 2 280 of 2 696 B-similes (85%). The difference between the two subsets of B-similes in terms of the token frequencies of their CNACs proved significant in both COCA ($U = 4809.000$, $p < .001$) and iWeb ($U = 163.000$, $p < .001$). In other words, the token frequency of CNACs corresponding to B-similes with M-modification is significantly higher than that of CNACs corresponding to B-similes without M-modification. This does not mean that the presence of frequent CNACs is the cause of M-modification. Instead, we could say that, while the preexistence of CNACs is not a necessary condition for M-modification and the CNACs may (perhaps exceptionally) only arise as part of the M-modification process, M-modification is still significantly more likely to be found with B-similes with fairly frequent CNACs.

4.2.3 Family size of B-simile templates

Concerning the lexical variation of B-similes in the source slot, our assumption that B-simile templates ("templates" for short) with a bigger family of source alternatives may be more prone to M-modification than those with a smaller family of alternatives was justified by our data. We found a statistically significant difference between the two subsets of templates in all three databases (iWeb: $U = 111.000$, $p < .001$; COCA: $U = 384.000$, $p < .001$; Google: $U = 2577.000$, $p < .001$).

Tab. 4: Descriptive statistics for the family size of B-simile templates (iWeb, COCA, Google)

		TypF	TokF	Mode	Min	Max	Mean Rank	Sum of Ranks
iWeb	**Bsim +M**	10	221	21	9	43	289.40	2 894.00
BsimTemp	**Bsim -M**	295	1 064	1	1	45	148.38	43 771.00
	Total	305						
COCA	**Bsim +M**	10	365	19	2	102	574.10	5 741.00
BsimTemp	**Bsim -M**	607	2 350	1	1	68	304.63	184 912.00
	Total	617						
Google	**Bsim +M**	55	1 147	1*	1	94	510.15	28 058.00
BsimTemp	**Bsim -M**	529	1 434	1	1	52	269.87	142 762.00
	Total	584	221					

* Multiple modes exist. The smallest value is shown; BsimTemp: B-simile template; Bsim+M: B-simile templates that include B-similes with M-modification; Bsim-M: B-simile templates that include B-similes without M-modification

According to Table 4, in iWeb and COCA, the templates for B-similes with M-modification have a bigger minimum family size than the templates covering B-similes without M-modification (in Google, their minimum family sizes are equal). In COCA and Google, the templates for B-similes with M-modification also have a bigger maximum family size than the templates comprising B-similes without M-modification (in iWeb the situation with the maximum family size is only slightly reversed). Overall, these data suggest that M-modification tends to be found within families of B-similes where at least some variety of sources already exists, and preferably within bigger families of B-similes. Concerning the templates in iWeb and COCA that include B-similes without M-modification, it is noteworthy that their minimal family size of 1 is also the most commonly occurring (*mode*) family size in those subsets. This means that no M-modification was found for the many B-similes that were the sole members of their respective templates. The situation was slightly different for Google, where the most frequent (*mode*) family size of templates comprising B-similes with and B-similes without M-modification was equally 1. Still, a close inspection of the Google data revealed that whereas in the case of templates for B-similes without M-modification this accounts for 301 of 529 cases or 56.9%, in the templates for the B-similes with M-modification, those with a family size 1 occur only three times, and that accounts for a very small fraction of all cases, viz. 5.5%.

4.2.4 Frequencies of source NPs

The frequency of source NPs in iWeb and COCA (Tab. 5),[28] and their frequency as source NPs in our B-simile databases from all three corpora (Tab. 6) were also tested as potential factors distinguishing the two subsets of B-similes. The two subsets of B-similes were found to differ significantly, in that those which do M-modify have source NPs that are more token-frequent in general (iWeb: U = 1854.000, p = .002; COCA: U = 3452.000, p < .001) and that are also more frequently deployed as sources in B-similes (iWeb: U = 1344.000, p < .001; COCA: U = 4214.500, p < .001; Google: U = 28332.000, p < .001) than the source NPs of the B-similes without M-modification. The first finding is consistent with Mancuso and Laudana (2019), who report that idioms which more often occur in corpora in a manipulated form are made up of frequent words. Note also the big discrepancy in Table 5 in the minimum corpus token frequencies of NPs acting as sources in B-similes without and those with M-modification. The former occur with a minimum frequency of 1, the minimum frequency of the latter runs into (a) thousand(s). B-similes with M-modification house significantly more token-frequent NPs as sources than B-similes without M-modification.

A closer inspection of the results behind those reported in Table 6 revealed that for all three databases, in B-similes not participating in M-modification, the bulk of source NPs occurs only once. In COCA, these singly occurring NPs account

Tab. 5: Descriptive statistics for the source NPs in the two corpora: iWeb and COCA

		TypF	TokF	Mode	Min	Max	Mean Rank	Sum of Ranks
iWeb	**Bsim +M**	12	7 982 999	-*	24 434	1 932 334	482.00	5 784.00
Source NP	**Bsim -M**	630	236 655 914	1**	1	24 565 941	318.44	200 619.00
	Total	642						
COCA	**Bsim +M**	12	365 201	-*	1 875	136 532	1 276.83	15 322.00
Source NP	**Bsim -M**	1 558	20 439 176	1	1	1 669 055	781.72	1 217 913.00
	Total	1570						

* All frequencies were registered only once; ** Multiple modes exist, the smallest value is shown; Bsim +M = source NPs found in B-similes with M-modification; Bsim -M = source NPs found in B-similes without M-modification

28 The general token frequencies of source NPs were not checked for the Google database since, unlike with iWeb and COCA, we could not control for parts of speech, nor search Google by lemmas rather than word-forms.

Tab. 6: Descriptive statistics for the source NPs in the three B-simile databases: iWeb, COCA, Google

		TypF	TokF	Mode	Min	Max	Mean Rank	Sum of Ranks
iWeb	**Bsim +M**	12	62	2	1	12	524.50	6 294.00
Source NP	**Bsim -M**	630	1 221	1	1	21	317.63	200 109.00
	Total	642						
COCA	**Bsim +M**	12	54	1*	1	17	1 213.29	14 559.50
Source NP	**Bsim -M**	1 558	2 629	1	1	25	782.21	1 218 675.50
	Total	1570						
Google	**Bsim +M**	105	489	1*	1	26	1 123.17	117 933.00
Source NP	**Bsim -M**	1 340	2 079	1	1	19	691.64	926 802.00
	Total	1445						

* Multiple modes exist. The smallest value is shown. Bsim +M = source NPs found in B-similes without M-modification; Bsim -M = source NPs found in B-similes without M-modification

for 1 168/1 558 (75%) of all source NPs; in iWeb, for 400/630 (63.5%) of all source NPs; and in Google, for 1 024/1 340 (76.4%) of all source NPs. This does not mean that the subset of B-similes with M-modification does not include source NPs occurring in only one B-simile, but such NPs are comparatively infrequent – they account for 1/12 (8.3%) of source NPs in the iWeb database, 3/12 (25%) of source NPs in the COCA database, and 24/105 (22.9%) of source NPs in the Google database. Concerning source NPs in B-similes with M-modification, given the comparatively small number of such NPs in all three databases (iWeb: N = 12, COCA: N = 12, Google: N = 105), it is unsurprising that the sum of their token frequencies in the database is not large (62, 54, and 489, respectively). Still, their mean rank is always higher than the mean rank of NPs in B-similes without M-modification. This is undoubtedly due to the fact that between 63.5% and 76.4% of NPs in the latter group occur only once. All of this points to the conclusion that, while M-modification is not impossible with B-similes featuring infrequently exploited source NPs, M-modification still mainly occurs with B-similes whose NP sources more often play the part.

4.2.5 Adjective syllables

We next explored the syllabic size of adjectives to verify the assumption that M-similes are more likely with shorter adjectives due to limited space in the CNAC. Admittedly, in Moon's (2008) corpus of similes most adjectives were found to be monosyllabic anyway. According to the raw frequencies in Table 7, the majority

Tab. 7: Frequencies of adjectives of different syllabic sizes in B-similes with and without M-modification (iWeb, COCA, Google)

			Polysyllabic						Total
		1-syllabic	2-syllabic	3-syllabic	4-syllabic	5-syllabic	6-syllabic	>1 Total	
iWeb	**Bsim +M**	12	1	0	0	0	0	1	13
adj syllables	**Bsim -M**	872	289	87	20	4	0	400	1 272
	Total	884	290	87	20	4	0	401	1 285
COCA	**Bsim +M**	13	1	0	0	0	0	1	14
adj syllables	**Bsim -M**	1559	821	230	64	20	2	1137	2 696
	Total	1572	822	230	64	20	2	1138	2 710
Google	**Bsim +M**	123	15	1	0	0	0	16	139
adj syllables	**Bsim -M**	1404	752	204	60	18	2	1036	2 440
	Total	1527	767	205	60	18	2	1052	2 579

of adjectives in B-similes in our three databases are also monosyllabic, regardless of whether the B-similes M-modify or not.[29] However, B-similes without M-modification also feature a number of polysyllabic adjectives, which is only exceptionally or rarely the case with B-similes that do M-modify. Thus, although adjectives proved to be almost exclusively or most frequently monosyllabic in B-similes with and without M-modification, respectively, the proportion of monosyllabic to polysyllabic adjectives is still higher in B-similes with M-modification than in B-similes without M-modification in all three databases. The difference was significant in COCA (χ^2 (1, N = 2 710) = 5.652, p = 0.017) and Google (χ^2 (1, N = 2 579) = 50.879, p < .001), but was non-significant in iWeb (χ^2 (1, N = 1 285) = 2.366, p = .124).[30] We conclude that in B-similes with M-modification, monosyllabic adjectives remain virtually unrivalled by the comparatively few polysyllabic adjectives, while B-similes without M-modification have a sizable proportion of adjectives of more than one syllable.

4.2.6 Source noun syllables

Finally, we checked whether, like adjectives, nouns from the syntactically simple NP sources are also characteristically short in B-similes participating in M-modification. After all, the source noun needs to become the first member in the spatially constricted CNAC. We restricted this analysis to those B-similes whose sources are expressed as fairly simple NPs (head noun with/without a closed-class determiner), so that only the determiner can be assumed to be left out as the noun gets absorbed into the CNAC. Table 8 details, for each set, the number of B-similes with nouns consisting of one or more syllables. Clearly, monosyllabic source nouns are the most frequent in both subsets of B-similes, but the proportions of short and longer nouns are quite close in both (albeit closer in B-similes without M-modification). When tested for significance, B-similes with M-modification and those with no M-modification proved not to be different in terms of the proportions of short and long source nouns in any of the databases: iWeb (χ^2 (1, N = 1 141) = .000, p = 1.000), COCA (χ^2 (1, N = 1 958) = 1.197, p = .274), Google (χ^2 (1, N = 1 935) = 1.482, p = .223).

29 Two-syllabic adjectives are found in M-similes *as rock-**solid** as* NP (iWeb) and *as rock **steady** as* NP (COCA). Google delivers more: *three-dollar-bill-**phoney** as the Nevada neon strip itself*, *butter-**yellow** as a plumeria blossom, as rock **solid** as a balloon made to look like a rock,* etc., even one tri-syllabic adjective in *stone **serious** as nuclear war*.

30 We calculated the difference between monosyllabic and cumulatively all polysyllabic adjectives.

Tab. 8: Frequencies of source nouns of different syllabic sizes in B-similes with and without M-modification (iWeb, COCA, Google)

			Polysyllabic					Total
		1-syllabic	2-syllabic	3-syllabic	4-syllabic	5-syllabic	>1 Total	
iWeb	**Bsim +M**	8	5	0	0	0	5	13
noun syllables	**Bsim -M**	660	383	72	12	1	468	1 128
	Total	668	388	72	12	1	473	1 141
COCA	**Bsim +M**	10	4	0	0	0	4	14
noun syllables	**Bsim -M**	1 034	692	181	34	3	910	1 944
	Total	1 044	696	181	34	3	914	1 958
Google	**Bsim +M**	66	59	11	1	0	71	137
noun syllables	**Bsim -M**	970	629	163	33	3	828	1 798
	Total	1 036	688	174	34	3	899	1 935

4.3 Semantic features

The four semantic aspects of M-similes announced in Section 2.2 will be discussed in two subsections:

1. semantic domains of the three nominal entities: old source, new source, and target, and the existence of (mis)matches between/among them (Section 4.3.1)
2. semantic domains of the properties for which the old source, new source, and target are compared, and the existence of (mis)matches between/among them (Section 4.3.2)

Recall that this analysis was done at token level and only on non-duplicate M-similes from all three corpora (N = 289).[31] Also, unlike in Section 4.2, no comparison was attempted between the semantic features of B-similes that have, and those that have not been found to M-modify.

4.3.1 Semantic domains of old sources, new sources, and targets

Probably the most exciting thing about M-similes is the replacement of the old source entity in its role as *comparatum* by a new source. Looking into what kinds of new entities come into play and how they combine with the old ones might reveal something about how expressivity is restored in M-modification. Before that, it is worth pointing out that the newly recruited sources are predominantly expressed as structurally more complex NPs (Tabs. 9 and 10). The proportion of complex source expressions in the overall structure of sources is significantly higher in new sources than in old ones (χ^2 (1, N = 578) = 288.429, p < .001). The two examples of structurally complex old sources in our database include *as **three-dollar-bill** phoney as the Nevada neon strip itself* (Google) and *as **night and day** different as passive and active* (Google). Examples of various kinds of structural complexity in the new sources are given in Table 10. As for structural simplicity, in old sources this means the presence of the source noun only (no determiner), in new sources this means an NP featuring a closed-class determiner and the head noun.[32]

31 Note that in Table 1, the unexpressed sum total of all tokens of M-similes in the three databases (where duplicates were not yet discarded) would be higher (N = 337).

32 Since new sources are expressed as full NPs but old sources are not, we avoid the label "NP" in Tables 9 and 10.

Tab. 9: Structural complexity of old and new sources in M-similes

	Structurally simple	%	Structurally complex	%	Total	%
Old sources	287	99.31	2	0.69	289	100.00
New sources	92	31.83	197	68.17	289	100.00

Tab. 10: Types, frequencies, and examples of structurally complex new sources in M-similes

	Type of structural complexity in new sources	N	Example
1	NP with a single simple premodifier/ genitive NP as a modifier or determiner	98	*Rail thin and with skin as nut-brown as **a Greek Islander**, he gives the appearance of Tolkien's Tree Beard* (Google)
2	NPs with various (combinations of) pre- and postmodifiers	57	*American cinema of the 1980s is a suitable monument to the Gipper: 10 years of what Harry Lime might refer to as cuckoo-clock art …, cast with actresses and actors as whistle-clean as **dirndl-clad milkmaids and farm boys of the Alps**.* (Google)
3	NP with a single prepositional phrase postmodifier	30	*…watercolors underscore the elegance of the words … with brushstrokes as whisper-soft as **a layer of sheer Egyptian gauze**.* (Google)
4	NP with a single clausal postmodifier	11	*About as rock solid as **a balloon made to look like a rock**, myself and others poked plenty of holes in it …* (Google)
5	Coordinated nouns	1	*Loving peace and making peace are as night and day different as **passive and active**.* (Google)

Entities from various semantic domains appear as old and new sources and targets, but with varying frequencies (Tab. 11). We discuss in detail the top-three-ranked old source domains to see how they combine with the domains of new sources. The remaining domains are presented in a summary fashion.

Let us first comment on the frequency rank of the domains of old sources from Table 11 for how this matches the results reported in Norrick (1987). Norrick (1987) analyzed 366 dictionary-collected stock similes and found that most source NPs (38%) belong to the category "animals", followed by "natural products" (19%) and "artefacts" (14%). We cannot tell if the same distribution would be found if we analyzed semantically the full set of B-similes in our three databases. However, Norrick's findings are not entirely consistent with our findings

concerning the old sources featured in M-similes. Table 11 shows that M-similes prefer, by a wide margin, old sources from the domain Environment, while Animals as old sources rank very low. Admittedly, the inventory and extension of the semantic domains in Norrick's study differs from ours, as do the sources of our data, but it is noteworthy that our most robust category Environment is not the most frequent source in Norrick's (1987) stock similes despite the fact that his category of "natural products" is broader and includes sources like *honey* (here classed under Food), *bone* (here: Human Body)[33] or *flowers* (here: Flora). Importantly, though, Norrick's two top-ranking semantic categories in the CNACs corresponding to the stock similes *do* match ours: "natural classes make up the largest class of vehicles by far, while artifacts run a weak second" (1987: 148). Norrick's finding concerning CNACs is very interesting since CNACs are part of the fabric of M-similes. Whatever explains the shift in frequency-rank of Animal and Environment sources in stock similes vs their CNACs in Norrick's study, the fact that the same two categories were found to be the most prolific in independent CNACs in Norrick's study and in old sources as parts of CNACs in M-similes, is relevant. It may be construed as yet another argument for a possible link between the existence of CNACs and the tendency to M-modify existing B-similes.

We could study the data in Table 11 from two perspectives: either starting from target domains to explore how each is matched with the domains of old and new sources, or starting from old source domains to see how each is matched with the domains of new sources and targets. Either way, we cannot afford complete analyses. To save space, we first provide general statistics about pairwise and cumulative semantic matches between/among old sources, new sources, and targets (Tabs. 12a, b), and then focus on the three most frequent old source domains to see how they match those of new sources (Tabs. 13–15). We are, after all, more interested in what may be gained by changing up the sources of comparison than with the question of which targets are most commonly found in M-similes.

At this level of granularity, we compared the nominals for the semantic domains from Table 11. This means that e.g. the old and the new source were treated as a match in ***crystal*** *clear as* ***spring water*** (iWeb) since both fit the category Environment, despite obvious differences. In contrast, the target and the new source were treated as non-matching in ***teeth*** *bone-white as* ***dentures*** (COCA), despite obvious similarities, because they occupy different domains, viz. Human Body vs Tools. The comparison of any pair of nominals shows that most commonly they involve entities from different

33 We decided to class as Human Body any instantiations of body parts that could be either human or animal in origin, like *blood, bone*. The category Animal Body was reserved for the lexical expression of those body parts that can only belong to an animal, like *feather*.

Tab. 11: Frequency-ranked semantic domains feeding old and new source NPs and target NPs*

No	Sem. domain: old sources	TokF	No	Sem. domain: new sources	TokF	No	Sem. domain: targets	
1	Environment (*desert, mud, rock*)	110	1	Environment (*the cracked earth on which he stands*)	39	1	Human Body	71
2	Tools (*razor, paper*)	64	2	Tools (*the hunting dagger strapped to my belt*)	31	2	Humans	47
3	Food and Drinks (*nut, milk*)	34	3	Weather, Natural Cycles and Processes (*a North Shore sunset*)	24	3	General Abstract	25
4	Flora (*lily, petals*)	22	4	Food and Drinks (*a Monty Python after-dinner mint*)	22	4	Tools	23
5	Human Body (*blood*)	17	5	Humans (*runway models; Elvis; a Missouri meth-head; a Greek islander; Popeye's girlfriend*)	21	5	Cultural Artefacts	18
6	Bodily Processes and Functions (*death, piss*)	10	6	Human Body (*a corpse's hand; Craig David's beard; the fingers of a surgeon; Snow White's hair; a young Aboriginal's limbs; a prom queen's thighs*)	20	6	Food and Drinks	13
7	Weather, Natural Cycles and Processes (*snow, dawn*)	9	7	Cultural Artefacts (*those giant human babies sculpted by Mueck*)	20	7	Environment	13
8	Furniture and Furnishings (*rail, pillow*)	7	8	Clothes and Accessories (*a freshly pressed tuxedo*)	19	8	Clothes and Accessories	9
9	Animal Body *(feather*)	5	9	Flora (*lotuses of the pond new opened*)	16	9	Animals	9
10	Animals (*otter*)	3	10	Animal (*a roping pony at roundup*)	12	10	Result (*brushstrokes*)	8
11	Humans (*baby*)	3	11	Architectural Artefacts (*the pyramids, the palace of Versailles*)	9	11	Flora	7
12	Supernatural Entities (*ghost*)	3	12	Animal Body (*the down of a new-born chick*)	8	12	Location	7

Tab. 11: (Continued)

No	Sem. domain: old sources	TokF	No	Sem. domain: new sources	TokF	No	Sem. domain: targets	
13	Extraterrestrial Entities (*moon*)	2	13	Location (*the towers they work in; the Nevada neon strip itself; Norway*)	7	13	Activity, Process, Event	7
			14	Furniture and Furnishings (*a billiard table; fine china*)	7	14	Weather, Natural Cycles and Processes	5
			15	Activity, Process, Event (*nuclear war; a plant blooming*)	6	15	Architectural Artefacts	5
			16	State, Disposition, Emotion (*a Las Vegas hangover; a tranquil mind*)	5	16	Character and Personality	5
			17	Bodily Processes and Functions (*a heartbeat*)	5	17	Furniture and Furnishings	5
			18	Intensifier (*heck, fuck, hell*)	5	18	State, Disposition, Emotion	4
			19	Extraterrestrial Entities (*the sun*)	4	19	Animal Body	4
			20	Supernatural Entities (*a vampire*)	3	20	Supernatural Entities	1
			21	Body of Supernatural Entity (*an angel's wings*)	2	21	Organization	1
			22	General Abstract (*passive and active*)	2	22	Unclear	2
			23	Character and Personality (*Mr. Frederiksen's humor*)	1			
			24	Organization (*the Mississippi Highway Patrol*)	1			

*Some domain names have been shortened for simplicity in our tables (sometimes in text too). An example (or more) is given for each old source and new source domain, and only for those target domains not already exemplified among the domains of the two sources.

Tab. 12a: Old sources, new sources and targets: semantic domain match between each pair of nominals

	No match	Match	Missing	Total
Old source + Target	263 (91.00%)	24 (8.30%)	2 (0.70%)	289 (100.00%)
New source + Target	237 (82.00%)	50 (17.30%)	2 (0.70%)	289 (100.00%)
Old source + New source	230 (79.60%)	59 (20.40%)	-	289 (100.00%)

*Missing: the target was unclear due to GDPR restrictions that precluded access to the complete example or the context was otherwise insufficient.

Tab. 12b: Old sources, new sources and targets: semantic domain match among all three nominals

Old source + target & new source + target & old source + new source	N	%
Match: no–no–no	168	58.13
Match: no–no–yes	50	17.30
Match: no–yes–no	46	15.92
Match: yes–no–no	18	6.22
Match: yes–yes–yes	5	1.73
Missing*	2	0.70
Total	289	100.00

*Missing: the target was unclear due to GDPR restrictions that precluded access to the complete example or the context was otherwise insufficient.

domains. This is so with 91% of old source–target pairs, 82% of new source–target pairs and 79% of old source–new source pairs (for some examples of the latter see Tabs. 13–15). When all three entities are considered together, the commonest case (58.13%) is M-similes with all nominals from different domains (example 7), the least common (1.73%) is where all three come from the same domain (example 8). The second most common scenario (17.30%) is M-similes where neither source matches the target, but the two sources are drawn from the same domain (example 9). Close in frequency (15.92%) are M-similes where the target-inconsistent old source is replaced with a new source matching the target (example 10). The second least common case is M-similes whose new source comes from a different domain than that of the matching old source and target (6.22%), as in example (11). This suggests that, when selecting new sources, speakers tend to make the conceptual leap to a domain different from that of the old source, and usually neither matches the domain of the target.

(7) *I can assure you will spy cattle in every shade of brown, black and grey, some patterned with spots and patches, and the* ***calves***$_i$ *all* ***milk***$_j$*-pale as* ***moonlight***$_k$ (Google)

(8) ***The air***$_i$ *…is as* ***crystal***$_i$ *clear as* ***spring water***$_i$ (iWeb)

(9) *My experience with the HTC One is that this* ***phone***$_i$ *is as* ***rock***$_j$ *solid as* ***Mt. Everest***$_j$. (iWeb)

(10) *He's wearing* ***a chef's smock***$_i$ *as* ***jet***$_j$*-black as* ***a shocktrooper's uniform***$_i$ … (COCA)

(11) *Meanwhile, cook remaining* ***butter***$_i$ *in a saucepan over medium-high heat until* ***nut***$_i$ *brown as* ***an Hermes bag***$_j$ (Google)

Tab. 13: Old source Environment: frequency-ranked semantic domains of new sources matched with old source Environment

	New source domains	TokF	%	Example
1	Environment	30	27.27	***rock*** *hard as* ***Mt. Everest*** (iWeb)
2	Weather, Natural Cycles & Processes	11	10.00	***pitch****-black as* ***midnight*** (iWeb)
3	Tools	9	8.18	***rock*** *hard as* ***nails*** (iWeb)
4	Architectural Artefacts	8	7.27	***sky****-high as* ***the buildings*** (iWeb)
5	Cultural Artefacts	8	7.27	***jet****-black as* ***Yojo, the ebony idol of Queequeg*** (iWeb)
6	Human Body	7	6.36	***rock*** *solid as* ***your abs*** (iWeb)
7	Food and Drinks	7	6.36	***lava****-hot as a* ***hot pocket*** (Google)
8	Humans	6	5.45	***stick****-thin as* ***Twiggy*** (Google)
9	Clothes & Accessories	4	3.64	***jet****-black as* ***a shocktrooper's uniform*** (COCA)
10	Animals	3	2.73	***stick****-thin as* ***a mantis*** (Google)
11	Activity, Process, Event	3	2.73	***stone****-serious as* ***nuclear war*** (Google)
12	Furniture & Furnishings	2	1.82	***rock*** *solid as* ***Optimus Prime's couch*** (Google)
13	Extraterrestrial Entities	2	1.82	***boulder****-heavy as* ***the moon*** (Google)
14	Location	2	1.82	***crystal*** *clear as* ***the beaches and pools he frequents*** (iWeb)
15	Animal Body	2	1.82	***soot*** *black as* ***a raccoon's (eyes)*** (Google)
16	*Other	1 × 6	0.91 × 6	***crystal*** *clear as a* ***vegetarian girl's urine*** (iWeb)
Total		**110**	**100.00**	

*Other (6 new source domains, each occurring once): Bodily Processes & Functions; Flora; State, Disposition, Emotion; Intensifier; General Abstract; Character & Personality

We next analyze in detail the three most frequent domains of old sources (Environment, Tools, and Food and Drinks, see Table 11), and specify for each the most frequent domains of their new source partners.

1. Environment, Natural Materials and Products (N = 110). This category includes aspects of the environment – defined as the air, water, and land on which humans, animals, and plants live, including natural objects and materials. It is the most robust category of old sources and the most versatile one (Tab. 13). The membership is large and usable enough to allow these old sources to be most frequently matched both with members of their own domain and with new sources from

Tab. 14: Old source Tools: Frequency-ranked semantic domains of new sources matched with old source Tools

	New source domains	TokF	%	Example
1	Human Body	9	14.06	***satin**-smooth as **a prom queens thighs*** (Google)
2	Tools	8	12.50	***razor**-sharp as **the axes that had beheaded the hundred convicted wizards of the Temple team*** (COCA)
3	Cultural Artefact	7	10.93	***whip**-quick as **cinema in the 50s*** (Google)
4	Clothes & Accessories	6	9.38	***porcelain**-pale as **her kimono*** (Google)
5	Humans	6	9.38	***whistle**-clean as **dirndl clad milkmaids and farm boys of the Alps*** (Google)
6	Flora	5	7.81	***needle**-sharp as **wet seaweed*** (Google)
7	Environment	5	7.81	***steel** cold as **ice*** (Google)
8	Food & Drinks	3	4.69	***porcelain**-smooth as **a cool Irish stout on a cool summer's evening*** (Google)
9	Animal Body	3	4.69	***sandpaper**-rough as **a cat's (tongue)*** (Google)
10	Location	3	4.69	***three-dollar-bill**-phoney as **the Nevada neon strip itself*** (Google)
11	Weather, Natural Cycles & Processes	2	3.13	***porcelain**-white as **freshly-fallen snow*** (Google)
12	Animals	2	3.13	***ink**-black as **witches' cats*** (Google)
13	Intensifier	2	3.13	***whip**-smart as **hell*** (Google)
14	*Other	1 × 3	1.56 × 3	***drum**-tight as **a Chinese cabinet*** (COCA)
Total		**64**	**100.00**	

*Other (3 new source domains, each occurring once): Bodily Processes & Functions; State, Disposition, Emotion; Furniture & Furnishings

20 more domains, albeit most often only once or rarely. Still, taken together, new sources from non-cognate domains (i.e. those different from the domain of the old source) outnumber new sources from the cognate domain.

2. Man-Made Tools and Materials (N = 64). This domain includes man-made tools, practical utensils, and materials that are instrumental in performing actions or are used for specific purposes. As the second most frequent domain of old sources, it also pairs with a fair number of new source domains (Tab. 14). Like Environment, it seems to be slightly conservative as it very often partners with members of its own domain. Still, the most prolific domain is Human Body, even if it only wins by a narrow margin.

3. Food and Drinks (N = 34). This domain includes all animal products, plant-based products (consumable flora), and other substances consumed to provide nutritional support or satisfy thirst. There is also a considerable dispersion of cases across new source domains (Tab. 15). Only three domains of new sources occur with a frequency of 5 or more, nine of the remaining 12 occur once, and three occur twice. This domain, too, most readily pairs up with new sources from its own domain, the next most frequent being Tools, and Clothes and Accessories.

Tab. 15: Old source Food and Drinks: Frequency-ranked semantic domains of new sources matched with old source Food and Drinks

	New source domains	TokF	%	Example
1	Food & Drinks	9	26.47	***marshmallow****-smooth as* ***butter*** (Google)
2	Tools	5	14.71	***wafer****-thin as* ***a contact lens*** (Google)
3	Clothes & Accessories	5	14.71	***nut*** *brown as* ***an Hermes bag*** (Google)
4	Environment	2	5.88	***flour****-fine as* ***sand from Aram*** (Google)
5	Activity, Process, Event	2	5.88	***honey****-slow as* ***a plant blooming*** (Google)
6	State, Disposition, Emotion	2	5.88	***lobster****-red as* ***a sunburn*** (Google)
7	*Other	1 × 9	2.94 × 9	***butter****-yellow as* ***a plumeria blossom*** (Google)
Total		**34**	**100.00**	

*Other (9 new source domains, each occurring once): Bodily Processes & Functions; Humans; Animals; Flora; Architectural Artefacts; Weather, Natural Cycles & Processes; Cultural Artefacts; Furniture & Furnishings; Intensifier

Tab. 16: Summary of remaining old source – new source pairings

	Old source domains (TokF)	Number of new source domains	TokF of examples per new source domain
4	Flora (22)	12	Flora (4), Weather (3), Humans (3), Environment (2), Clothes and Accessories (2), Body of Supernatural Entity (2), Tools (1), Human Body (1), Food and Drinks (1), Location (1), Activity, Process, Event (1), Organization (1)
5	Human Body (17)	9	Tools (4), Weather (3), Flora (2), Furniture and Furnishings (2), Cultural Artefacts (2), Food and Drinks (1), Supernatural Entities (1), Extraterrestrial Entities (1), Animal Body (1)
6	Bodily Processes & Functions (10)	6	Tools (3), Bodily Processes (2), Weather (2), Flora (1), Human Body (1), Clothes and Accessories (1)
7	Weather, Natural Cycles & Processes (9)	9	Weather (1), Flora (1), Animals (1), Supernatural Entities (1), Extraterrestrial Entities (1), Cultural Artefacts (1), Animal Body (1), State, Disposition, Emotion (1), General Abstract (1)
8	Furniture & Furnishings (7)	4	Humans (4), Weather (1), Animals (1), Supernatural Entities (1)
9	Animal Body (5)	5	Animal Body (1), Animals (1), Flora (1), Furniture and Furnishings (1), Location (1)
10	Animals (3)	2	Animals (2), Environment (1)
11	Humans (3)	3	Tools (1), Food and Drinks (1), Cultural Artefacts (1)
12	Supernatural Entities (3)	3	Weather (1), Humans (1), Intensifier (1)
13	Extraterrestrial Entities (2)	2	Human Body (1), Animals (1)

The remaining domains of old sources are presented in summary fashion in Table 16.

The fact that some domains of old sources are more often involved in M-similes does not entail that the full lexical potential of those domains has been exploited. In Table 17 we provide the type/token ratio of lexemes for each domain of old sources. The idea is to see whether relatively higher token counts of M-similes per domain involve heavier concentrations of M-simile tokens around particular lexemes. This would mean that a semantic domain is only partly exploited since some of the source lexemes are keener to participate in M-similes than others.

Lexical type frequency indicates how many different lexemes from the same semantic domain of old sources were found to participate in M-modification. *Lexical token frequency* is the sum of all occurrences of all old source lexeme types from the same domain. The *lexical T/T ratio* measures the dispersion of lexical tokens

Tab. 17: Lexical type and token frequency per semantic domain of old sources and their lexical T/T ratio

	Sem. domain of old sources	Lexical TypF	Lexical TokF	Lexical type/ token ratio	TokF of particular lexemes
1	Environment	32	110	0.3	*crystal* (22), *stick* (10), *rock* (9), *jet* (8), *pitch* (8), *stone* (7), *soot* (5), *coal* (4), *ruby* (4), *marble* (4), *ice* (3), *sky* (3), *iron* (2), *dirt* (2), *dust* (2), *emerald* (1), *gold* (1), *mud* (1), *glacier* (1), *cave* (1), *desert* (1), *sand* (1), *tinder* (1), *ash* (1), *diamond* (1), *flint* (1), *granite* (1), *boulder* (1), *lead* (1), *lava* (1), *alabaster* (1), *sea* (1)
2	Tools	27	64	0.4	*razor* (8), *satin* (6), *porcelain* (5), *paper* (5), *tissue* (3), *parchment* (3), *pencil* (3), *needle* (3), *steel* (3), *drum* (2), *dagger* (2), *neon* (2), *sandpaper* (2), *velvet* (2), *whip* (2), *glass* (2), *whistle* (1), *three-dollar bill* (1), *silk* (1), *rubber* (1), *poker* (1), *plastic* (1), *pin* (1), *knife* (1), *ink* (1), *chalk* (1), *bowstring* (1)
3	Food & Drinks	15	34	0.4	*wafer* (10), *milk* (5), *nut* (4), *butter* (3), *beet* (2), *beetroot* (1)*, *honey* (1), *flour* (1), *banana* (1), *gourd* (1), *marshmallow* (1), *vanilla* (1), *syrup* (1), *pancake* (1), *lobster* (1)
4	Flora	4	22	0.2	*lily* (9), *petal* (6), *ebony* (5), *daisy* (1), *reed* (1)
5	Human Body	4	17	0.2	*bone* (8), *blood* (7), *skull* (1), *flesh* (1)
6	Bodily Processes & Functions	3	10	0.3	*death* (1), *piss* (1), *whisper* (8)
7	Weather, Natural Cycles & Processes	7	9	0.8	*snow* (2), *fire* (2), *dawn* (1), *night and day* (1), *lightning* (1), *cloud* (1), *summer* (1)
8	Furniture & Furnishings	3	7	0.4	*rail* (4), *pillow* (2), *sheet* (1)
9	Animal Body	2	5	0.4	*feather* (3), *ivory* (2)
10	Animals	3	3	1.0	*buck* (1), *cat* (1), *otter* (1)
11	Humans	1	3	0.3	*baby* (3)
12	Supernatural Entities	1	3	0.3	*ghost* (3)
13	Extraterrestrial Entities	1	2	0.5	*moon* (2)

* *Beet* and *beetroot* were counted as two separate lexemes despite being the same type of vegetable.

across lexical types. This information is different from the type and token frequencies of M-similes and their T/T ratios in Table 1, since a particular source lexeme may be involved in different M-simile types, e.g. *stone-steady as the pyramids* (Google) vs *stone-still as a sarcophagus* (Google). Here, the lexical type frequency is 1, but the type frequency of M-simile is 2. Note that, like with the T/T ratio of M-similes, the lexical T/T ratio also needs to be interpreted relative to the overall number of lexical types and tokens. As Table 17 shows, several old source domains have fairly low T/T ratios (0.2 and 0.3). Each involves one or two overextended lexemes. In the category Environment (T/T = 0.3), this is *crystal*, which in all instances is involved in the M-simile type *crystal-clear as X*. *Stick* (only in *stick-thin as X*), *rock*, *jet*, *pitch* and, perhaps also *stone* also have a notable presence. With Flora, it is *lily*, *petal* and perhaps also *ebony*, with Human Body – *bone* and *blood*, etc.

So far, our main concern was with source entities but not to leave targets completely on the sidelines, here are a few general comments about the semantics of targets. As can be gleaned from the target column in Table 18, M-similes reflect a strong human concern with self. The most frequent targets instantiate the domains of Human Body and Humans. These domains are considerably less exploited as old sources, but, interestingly, make a stronger appearance among new sources, which is where we typically find references to fictional characters (*Huck Finn*), famous real individuals (*Lance Armstrong*), classes of humans defined by profession (*Calvin Klein models on the runway*), ethnicity (*a Brit*) or other social groupings (*a Missouri meth-head*). The third most frequent target domain is General Abstract entities. This is unsurprising given the natural human tendency to reach for the concrete to better understand the abstract – as widely documented by research in conceptual metaphor theory inaugurated by Lakoff and Johnson (1980). When analyzing the three most frequent targets: Human Body, Human, General Abstract, we found that all are most often matched with old sources from the same two categories: Environment (Human Body: N = 23/71 or 32.4%; Human: N = 15/47 or 31.9%; General Abstract: N = 12/25 or 48%) and Tools (Human Body: N = 16/71 or 22.5%; Human: N = 11/47 or 23.4%, General Abstract: N = 8 or 32%). However, all three are matched with a wider range of new sources than old sources: the target Human Body became more strongly associated with new sources from Flora (N = 11/71 or 15.5%), followed by Human Body (N = 9/71 or 12.7%). The target Human linked most often with the new source domain Humans (N = 15/47 or 31.9%), and Tools (N = 5/47 or 10.6%). General Abstract targets appeared more conservative as they continued to prefer new sources from the domains Environment (N = 8/25 or 32%), but Tools (N = 4/25 or 16%) and Food and Drinks (N = 4/25 or 16%) shared 2nd rank. This alone suggests that speakers feel little need to stick to the same kinds of sources as found in B-similes when describing and evaluating their targets afresh by comparison with new sources.

Tab. 18: Summary presentation of target-old source and target-new source domain pairings

	Target domains (TokF)	Number of old source domains	TokF of examples per old source domain	Number of new source domains	TokF of examples per new source domain
1	Human Body (71)	12	Environment (23), Tools (16), Flora (8), Human Body (8), Food (4), Animal Body (4), Weather (2), Humans (2), Bodily Processes (1), Furniture (1), Supernatural (1), Extraterrestrial (1)	17	Flora (11), Human Body (9), Environment (8), Tools (8), Weather (5), Humans (5), Food (4), Animal Body (4), Animals (3), Clothes (3), Furniture (2), Cultural Artefacts (2), Location (2), State, Disposition, Emotion (2), Supernatural (1), Extraterrestrial (1), Architectural Artefacts (1)
2	Humans (47)	9	Environment (15), Tools (11), Flora (5), Food (5), Furniture (4), Animals (3), Human Body (2), Supernatural (1), Extraterrestrial (1)	15	Humans (15), Tools (5), Environment (3), Food (3), Animals (3), Clothes (3), Archit. Artefacts (3), Intensifier (3), Human Body (2), Furniture (2), Weather (1), Supernatural (1), Cultural Art. (1), Body of Supernat. Ent. (1), Gen. Abstract (1)
3	General Abstract (25)	5	Environment (12), Tools (8), Food (3), Flora (1), Bodily Processes (1)	10	Environment (8), Tools (4), Food (4), Human Body (2), Architectural Artef. (2), Weather (1), Bodily Processes (1), Animals (1), Cultural Artefacts (1), Clothes (1)
4	Tools (23)	8	Tools (7), Environment (5), Food (4), Flora (2), Bodily Processes (2), Weather (1), Human (1), Supernatural (1)	9	Cultural Artef. (5), Environment (3), Tools (3), Weather (3), Clothes (3), Human Body (2), Food (2), Bodily Processes (1), Body of Supernat. Entity (1)
5	Cultural Artefacts (18)	3	Environment (10), Tools (7), Bodily Processes (1)	9	Cultural Artef. (5), Environment (3), Weather (3), Location (2), Tools (1), Humans (1), Food (1), Clothes (1), Animal Body (1)

Tab. 18: (Continued)

	Target domains (TokF)	Number of old source domains	TokF of examples per old source domain	Number of new source domains	TokF of examples per new source domain
6	Food (13)	5	Environment (5), Food (4), Tools (2), Flora (1), Furniture (1)	9	Food (3), Weather (2), Location (2), Environment (1), Tools (1), Clothes (1), Activity, Process, Event (1), Intensifier (1), Character (1)
7	Environment (13)	5	Environment (6), Tools (3), Food (2), Human Body (1), Animals (1)	8	Environment (4), Furniture (2), State, Disposition, Emotion (2), Tools (1), Flora (1), Animals (1), Cultural Artef. (1), Animal Body (1)
8	Clothes (9)	5	Environment (4), Tools (2), Food (1), Bodily Processes (1), Animal Body (1)	8	Clothes (2), Environment (1), Flora (1), Weather (1), Humany Body (1), Animals (1), Cultural Artef. (1), Intensifier (1)
9	Animals (9)	5	Environment (4), Food (2), Flora (1), Weather (1), Furniture (1)	7	Weather (3), Environment (1), Food (1), Animals (1), Supernatural (1), Clothes (1), Activity, Process, Event (1)
10	Result (8)	5	Environment (3), Tools (2), Weather (1), Food (1), Bodily Processes (1)	6	Tools (2), Human Body (2), Environment (1), Weather (1), Cultural Artef. (1), Animal Body (1)
11	Flora (7)	4	Weather (3), Human Body (2), Environment (1), Food (1)	7	Tools (1), Flora (1), Weather (1), Extraterrestrial (1), Cultural Artefacts (1), Clothes (1), State, Disposition, Emotion (1)
12	Location (7)	3	Environment (4), Human Body (2), Food (1)	6	Environment (2), Human Body (1), Food (1), Animals (1), Extraterrestrial (1), Architectural Artefacts (1)
13	Activity, Process, Event (7)	5	Environment (2), Flora (2), Weather (1), Food (1), Bodily Processes (1)	5	Activity, Process, Event (3), Tools (1), Weather (1), Location (1), General Abstract (1)
14	Weather (5)	3	Environment (2), Human Body (2), Food (1)	5	Tools (1), Weather (1), Food (1), Bodily Processes (1), Furniture (1)

Tab. 18: (Continued)

	Target domains (TokF)	Number of old source domains	TokF of examples per old source domain	Number of new source domains	TokF of examples per new source domain
15	Architectural Artefacts (5)	2	Environment (3), Tools (2)	4	Architectural Artefact (2), Environment (1), Flora (1), Weather (1)
16	Character and Personality (5)	3	Tools (3), Environment (1), Flora (1)	4	Tools (2), Flora (1), Cultural Artefacts (1), Clothes (1)
17	Furniture (5)	3	Environment (2), Bodily Processes (2), Tools (1)	5	Weather (1), Food (1), Bodily Processes (1), Extraterrestrial (1), Animal Body (1)
18	State, Disposition, Emotion (4)	2	Environment (2), Food (2)	3	Environment (2), Bodily Processes (1), Activity, Process, Event (1)
19	Animal Body (4)	3	Tools (2), Environment (1), Food and Drinks (1)	4	Environment (1), Tools (1), Cultural Artefacts (1), Clothes (1)
20	Supernatural Entities (1)	1	Environment (1)	1	Animals (1)
21	Organization (1)	1	Flora (1)	1	Organization (1)
22	Unclear (2)	1	Environment (2)	2	Environment (1), Food (1)

4.3.2 Semantic domains of properties

Nominal entities can be compared if they have some shared ground. In B- and M-similes, the property coded in the adjective acts as the *tertium comparationis*. B-similes involve two entities and their properties, but M-similes are all about matching the properties of three nominals. The "shared ground" usually means that all three entities share exactly the same property, or more accurately, the same value (exaggeratedly, cf. Section 2.1) of a typically scalar property (example 12). However, the idea of a shared ground is challenged by ironic (M-)similes, where some of the entities exhibit opposite property values (example 13). Finally, a more abstract shared ground must be assumed in M-similes with polysemous adjectives, where entities may be compared for different, but still related properties (example 14).

Before examining how the properties of the three nominals in M-similes match, we first analyze the properties of each of the three nominals separately and categorize them into semantic domains (Tab. 19). Whether or not properties qualify as the same is a matter of construal. Even in the seemingly easily-understood, experientially grounded domains of physical properties, things are not exactly straightforward. Take two examples. The properties coded by *still* in *patient as stone-still as a sarcophagus* and by *steady* in *defense as stone-steady as the pyramids* both imply immovability. Yet in *stone-still*, *still* is saliently construed as the absence of (desirable) ability to move, while *steady* in *defense as stone-steady as the pyramids* is saliently construed as the ability to withstand (undesirable) movement. Similarly, within the same M-simile, viz. (*a tale*) *crystal clear as the skies above Paradise Islands* (iWeb), clarity of crystal and clarity of skies both mean absence of "blemishes" viz. chemical or physical impurities and clouds respectively. But in each entity a different aspect of this property is salient: in the case of crystal – its transparency; in skies – the (temporary) state of not being overcast.

To keep the analysis doable and avoid excessively inflating the number of domains for properties, the domains have been postulated at the level of scales (most properties are scalar). Each domain (scale) is meant to bring together adjectives coding various scalar values, e.g. the domain labeled Tactile Resistance would include both the adjectives *hard* and *soft*. The downside is that this level of analysis does not allow the exploration of ironic effects that come from assessing source and target entities as equal while featuring opposite scalar values. But it does not hide polysemy since properties of nominal entities linked by polysemy will fall into different categories, e.g. Effectiveness and Geometric Sharpness, when one's leadership is described as *needle sharp as wet seaweed*. Sadly, space prevents a thorough analysis of the kinds of polysemy involved, although it should be noted that most adjectives have concrete basic senses, and metaphorically,

Tab. 19: Frequency-ranked semantic domains of properties implicit in old sources, new sources, and targets

	Sem. domain of properties of old sources	TokF		Sem. domain of properties of new sources	TokF		Sem. domain of properties of targets	TokF
1	Color (hue: *white*, saturation: *pale*, lightness: of blood: *dark*)	102	1	Color (hue, saturation, lightness)	89	1	Color (hue, saturation, lightness)	87
2	Thickness (*thin, skinny*)	41	2	Thickness	39	2	Thickness	33
3	Transparency (*clear*)	22	3	Tactile Resistance	20	3	Textural Smoothness	19
4	Textural Smoothness (*smooth, rough*)	19	4	Transparency	18	4	Tactile Resistance	18
5	Tactile Resistance (*soft, hard*)	19	5	Textural Smoothness	18	5	Cognitive Clarity	14
6	Geometric Sharpness (*sharp*)	11	6	Geometric Sharpness	10	6	Value and Quality	13
7	Temperature (*cold*)	9	7	Luminance	9	7	Intensity of Action	8
8	Humidity (*dry*)	9	8	Material Strength	9	8	Emotion, Disposition, State	8
9	Material Strength (of iron: *strong*)	9	9	Humidity	7	9	Luminance	6
10	Intensity of Action (of whisper: *thin, soft*)	8	10	Temperature	6	10	Racial Profile	6
11	Luminance (of dawn: *bright*)	4	11	Unmotivated	6	11	Auditory Pleasantness	6
12	Speed and Agility (*slow*)	4	12	Intensity of Action	6	12	Humidity	5
13	Tension (*taut*)	3	13	Emotion, Disposition, State (of surgeon: *cold*)	4	13	Physical Purity	5
14	Immovability (*still, solid*)	3	14	Speed and Agility	4	14	Quantity (of fandom: *thin*)	5
15	Height (*high*)	3	15	Immovability	3	15	Cleverness (of novel: *sharp*)	4
16	Unmotivated (of *stone*: *deaf, serious*)	3	16	Racial Profile (*white*)	3	16	Temperature	4
17	Stability (*steady*)	2	17	Height	3	17	Other Abstract Property	4

Tab. 19: (Continued)

	Sem. domain of properties of old sources	TokF		Sem. domain of properties of new sources	TokF		Sem. domain of properties of targets	TokF
18	Weight (*heavy*)	2	18	Weather Condition (of sky: *clear, heavy*)	3	18	Geometric Sharpness	4
19	External Bodily State (*naked*)	2	19	External Bodily State	3	19	Abstract Strength (of propensity: *strong*)	4
20	Value and Quality (of dirt: *cheap*)	1	20	Other Abstract Property	3	20	Speed and Agility	4
21	Taste (*salty*)	1	21	Physical Purity	2	21	Transparency	3
22	Viscosity (of syrup: *thick*)	1	22	Pleasantness	2	22	Immovability	3
23	Shape (*flat*)	1	23	Stability	2	23	Auditory Clarity (of voice: *clear*)	3
24	Stiffness (*stiff*)	1	24	Taste	2	24	External Bodily State	2
25	Authenticity (*phoney*)	1	25	Value and Quality	2	25	Moral Purity	2
26	Similarity (*different*)	1	26	Appeal	2	26	Appeal	2
27	Appeal (of characters: *plain*)	1	27	Auditory Pleasantness (of song: *hard*)	2	27	Stability	2
28	Naturalness (of daisy: *fresh*)	1	28	Auditory Pitch (of bus engine roar: *thick*)	1	28	Effectiveness (of phone: *smooth*)	1
29	Other Abstract Property (of death: *black*)	1	29	Authenticity	1	29	Auditory Impairment	1
30	Pleasantness (of summer: *sweet*)	1	30	Cognitive Clarity (of a tranquil mind: *clear*)	1	30	Authenticity	1
31	Physical Deficiency (of dirt: *poor*)	1	31	Event Qualification (of war: *serious*)	1	31	Event Qualification	1
32	Physical Purity (of crystal: *pure*)	1	32	Naturalness	1	32	Material Strength	1

Tab. 19: (Continued)

	Sem. domain of properties of old sources	TokF		Sem. domain of properties of new sources	TokF		Sem. domain of properties of targets	TokF
33	Unclear (of whistle: *clean*)	1	33	Abstract Strength (*a Greek destiny*)	1	33	Experience (of players: *green*)	1
			34	Shape	1	34	Naturalness	1
			35	Similarity	1	35	Pleasantness	1
			36	Social Condition (of Huck Finn: *poor*)	1	36	Shape	1
			37	Stiffness	1	37	Similarity	1
			38	Tension	1	38	Social Condition	1
			39	Weight	1	39	Stiffness	1
						40	Strictness (of rules: *strict*)	1
						41	Taste	1
						42	Weight	1

metonymically or metaphtonymically derived abstract senses wittily exploited in (M-)similes.[34] In this section we only identify and rank for frequency the semantic domains of properties for all three nominals (Tab. 19) and discuss whether or not these match, similarly to how we proceeded with nominals in Section 4.3.1.

(12) *Lips as red and **velvet-soft as petals** and a tongue as sharp as thorns.* (Google)
(softness of lips, velvet, and petals)

(13) *About **as rock solid as a balloon made to look like a rock**, myself and others poked plenty of holes in it and he ignored them completely* (Google)
(hardness of rock, but softness of balloon)

(14) *Enhance your days on the water with audio **as crystal-clear as the water around you*** (iWeb)
(visual transparency of water vs auditory clarity of sound)

Color, with its three aspects – hue, saturation, and lightness – tops the list as the most frequent property domain in all three nominal entities, viz. the old sources, the new sources, and the targets. The three nominals all have Physical Thickness as the second most frequent domain. If we focus on the remaining four domains with a frequency of at least 10 (those ranked 3rd to 6th in Table 19), it is noteworthy that old and new sources share the same four domains, all representing physical properties of matter: Transparency, Textural Smoothness, Tactile Resistance, and Geometric Sharpness. This is partly consistent with Norrick (1987), who found that similes cluster around certain *tertia*, specifically color and other "directly perceived" properties like *sharp*, *soft*, *cold*, etc. Targets have only Textual Smoothness and Tactile Resistance among the remaining four domains with at least two-digit tokens; the other two domains are more abstract, viz. Cognitive Clarity, and Value and Quality. This finding is not surprising since general abstract entities were among the most frequent target nominals too – and it is only natural that abstract entities would be described/evaluated for their abstract properties by

34 Goossens (1990) proposes *metaphtonymy* as a cover term for various interaction patterns between metaphor and metonymy. The details are beyond the scope of this paper, but we illustrate here one example of metaphor-metonymy interaction in our data. Underlying the example (of a lover) *steel **cold** as ice* (Google) is the conceptual metaphor EMOTION IS TEMPERATURE (Kövecses 2005: 156), specifically its entailment EMOTIONAL INDIFFERENCE IS LOW TEMPERATURE. However, this metaphor builds on the metonymic cause-effect relationship between emotions and their physiological effects, like changes in body temperature. The conceptual metaphor EMOTION IS TEMPERATURE arises by generalizing the notion of body heat to heat. This generalized source domain of temperature can then structure the target domain of emotions even when the particular physiological effect is absent.

comparison with concrete properties of their concrete source counterparts (see examples 21–23):

(15) *The following morning your mother is sitting at the kitchen table, sobbing and trembling, with a lot of paper tissues against her face, moist, yellow tissues,* ***as pissgrey as the underpants of the sixth graders in the dressing rooms of the swimming pool*** *…* (COCA)
(Color)

(16) *… out of it has emerged not a butterfly, but a predatory horror,* ***stick-thin as a mantis*** *…* (Google)
(Thickness)

(17) *Teeth* ***as razor sharp as stake knives*** (Google)
(Geometric Sharpness)

(18) *As pure, fresh and* ***crystal clear as a tide pool****, the Men's Care Shower Gel is the irrefutable wake-up call of your cleansing ritual* (Google)
(Transparency)

(19) *To my shock, neither muscle nor bone are evident on Donny's hairy shoulders and back, altogether bluff with mounds of fat; the ass is small and* ***rock-hard as a couple of gourds*** (COCA)
(Tactile Resistance)

(20) *Bubble me up until I am no longer coarse but* ***otter-sleek as a blood slicked stone*** (Google)
(Textural Smoothness)

(21) *After this journey, the answer is* ***as crystal clear as the freshly melted water in Iceland's lakes*** (iWeb)
(Cognitive Clarity)

(22) *… make sure you've got a business plan* ***as rock solid as your abs*** (iWeb)
(Value and Quality)

(23) *Capriccio: Experience was* ***as wafer thin as the pizza*** (Google)
(Value and Quality)

An analysis of the fit between/among properties inherent to the three nominals is shown in Tabs. 20a, b.

Unlike nominal entities, in all pairwise and cumulative comparisons of all three properties, the strongest scenario is one with both/all properties matching (for details see Tabs. 20a, b). This is not because of any undue

Tab. 20a: Properties of old sources, new sources and targets: semantic domain match between each pair of properties*

Property	No match	Match	Total
Old source + Target	95 (32.87%)	194 (67.13%)	289 (100.00%)
New source + Target	100 (34.60%)	189 (65.40%)	289 (100.00%)
Old source + New source	41 (14.19%)	248 (85.81%)	289 (100.00%)

*Due to GDPR restrictions and insufficient context, the target entity itself was unclear in two cases, but we could ascertain the nature of the properties from the rest of the simile, e.g. *brown* of some physical *residue* is clearly a color property, even if it is unclear what the target entity is.

Tab. 20b: Properties of old sources, new sources and targets: semantic domain match among all properties at once*

Properties: old source + target & new source + target & old source + new source	N	%
Match: no–no–no	7	2.42
Match: no–no–yes	15	5.19
Match: no–yes–no	19	6.57
Match: yes–no–no	71	24.57
Match: yes–yes–yes	177	61.25
Total	289	100.00

*Due to GDPR restrictions and insufficient context, the target entity itself was unclear in two cases, but we could ascertain the nature of the properties from the rest of the simile, e.g. *brown* of some physical *residue* is clearly a color property, even if it is unclear what the target entity is.

generalization that would lead to more elegant description since we were quite liberal in postulating as many separate domains as the data warranted. It is simply because comparisons can only work if there is a *tertium comparationis*, so a fair degree of matching was expected. A closer look into non-matching properties would in fact be far more interesting, as it would reveal various patterns of polysemy and how (much) polysemy contributes to the "creative flavor" of M-similes, possibly in comparison to other forms of simile modification, like simple lexical substitutions of source NPs. Sadly, this must be sidelined for space reasons. We submit here some illustrative examples of non-matching properties between old and new sources. Color as the old source property is readily matched with any of the following metonymic, metaphorical, or metaphtonymic properties in new sources: Racial Profile (e.g. of political rallies: *as lily-white as Norway* [Google]), Abstract Property (of beer: *pitch-black as*

Mr. Frederiksen's humor [iWeb]), Luminance (of horse: *pitch dark as the winter's night* [Google]), Emotion, Disposition, State (of a walking cane: *as ebony-dark as the soul of Satan* [Google]), Intensity (of mouth: *as ruby red as the desire of the Sanc Graal* [Google]). Temperature, unsurprisingly, is matched with Emotions, Dispositions and States (e.g. of person: *steel-cold as a surgeon* [Google]). Textural Smoothness pairs up with Taste (of person: *porcelain-smooth as a cool Irish stout on a cool summer's evening* [Google]) or General Pleasantness (of finger: *petal-soft as the rays of the early sun* [Google]). Material Strength matches Abstract Strength in *iron-strong as a Greek destiny* (Google) and Humidity pairs with Aesthetic Appeal in (of writing) *dust-dry as an organic chemistry textbook* (Google).

We close this section by examining, for the six most frequent property domains of old sources, the lexical variation in their respective M-similes. Table 21 presents their lexical type- and token frequency and the lexical T/T ratio, showing this time even more clearly that the lexical potential of those domains has remained largely untapped. The T/T ratios are very low, not going beyond 0.2. Each domain has one or two lexemes that account for almost all M-simile tokens.

4.4 Discourse-related features

In observing the kinds of text where M-similes can be found, we do not commit to any definitive distinction between individual register or genre types, or even

Tab. 21: Lexical type- and token frequency for the first six domains of properties in old sources, and lexical T/T ratio

	Sem. domain of old sources	Lexical TypF	Lexical TokF	Lexical type/ token ratio	TokF of particular lexemes
1	Color	12	102	0.1	*white* (28), *black* (23), *red* (13), *dark* (12), *pale* (11), *brown* (5), *yellow* (3), *bright* (2), *grey* (2), *green* (1), *orange* (1), *pink* (1)
2	Thickness	2	41	0.04	*thin* (37), *skinny* (4)
3	Transparency	1	22	0.04	*clear* (22)
4	Textural Smoothness	5	19	0.2	*smooth* (12), *soft* (3), *rough* (2), *sleek* (1), *fine* (1)
5	Tactile Resistance	2	19	0.1	*soft* (13), *hard* (6)
6	Geometric Sharpness	1	11	0.09	*sharp* (11)

Tab. 22: The frequency-rank of text varieties where M-similes were found

	Text variety	N	%
1	Creative writing and non-fiction prose	115	39.79
2	Online journalism	66	22.84
3	Company/commercial websites	43	14.88
4	Non-commercial websites/blogs of online communities	20	6.92
5	Personal blogs	20	6.92
6	Online forum/discussion sites	14	4.84
7	Social networks (Twitter, Facebook, etc.)	6	2.08
8	Other (Quora: question and answer website, n/a)	5	1.73
Total		**289**	**100.00**

generally between registers and genres. We adopt in principle the explanation of registers and genres as different perspectives on text varieties (Biber and Conrad 2009: 2). Both are defined functionally, by considering the communicative purposes of different varieties of language used in different situations: the register perspective focuses on a careful analysis of linguistic patterns associated with the situation of use or specifically, with particular purposes, topics, degrees of interactiveness, and mode whereas the genre perspective focuses on "the conventional structures used to construct a complete text within the variety" such as conventional ways to open or close a letter (Biber and Conrad 2009: 2).

Our discourse categories (Tab. 22) are a mix of what could be considered registers and genres. For instance, fiction (under creative writing) could best be analyzed as a register significantly associated with linguistic features typical of narrative production (Biber et al. 2000: 148). Blogs, on the other hand, qualify as a genre, at least in their original role as personal online diaries. However, none of the categories are discrete, and (sub)categories that carry different names may show increasing overlaps. A recent volume dedicated to Internet genres emphasizes the "volatility and chameleon-like properties of Internet genres", adding that "[t]here is a constant and fast proliferation of genres–and of forms of communication that are candidates for being a genre. [...] Existing genres quickly differentiate into sub-species" (Giltrow and Stein 2009: 9). Blogs, for instance, started out as personal online diaries, but have now evolved into professionally-edited platforms for company advertising, delivering news, and instructional resources (so-called edublogs). As for their role in journalism, according to Britannica, "To meet increasing consumer demand for up-to-the-minute and highly detailed reporting, media outlets developed alternative channels of dissemination, such as online distribution, electronic mailings, and direct interaction with

the public via forums, blogs, user-generated content, and social media sites such as Facebook and Twitter".[35] Essentially, any content created and shared (also publicly commented) via various social media or other services on a fairly regular basis can now be called a blog.[36] In sum, the categories proposed below represent text varieties loosely based around communicative purposes like describing, commenting, evaluating, advertising, etc., purposes which make them suitable hosts for M-similes. The texts varieties cover largely similar topics (home, cooking, travel, food, celebrities, sports, culture, business, relationships, nature, music, etc.) and most have a notable degree of interactiveness (especially those in 4 through 8).

The M-similes found in the most prolific category of creative writing and non-fiction prose are ideally suited to the purposes of the variety, viz. to develop characters, vividly describe entities, including abstract feelings and the setting (see 24–25). Most examples were found in fiction (N = 85), then poetry (N = 18), creative non-fiction (N = 5), non-fiction prose (N = 4), and song lyrics (N = 3). According to Moon (2008: 30), "the strong association in BoE[37] between as-similes and fiction suggests a convention of story-telling, with similes belonging to its sub-lexicon as a kind of descriptive commonplace [...]". The next most robust category is interpretive material featured in newspapers and magazines (see example 26), i.e. opinion and commentary articles found on the websites of media outlets. Note that, consistently with earlier claims (Moon 2008: 21), M-similes were not found in dry, informative, matter-of-fact news reports.[38] Creative M-similes also made a notable appearance on company/commercial websites (see example 27). A web presence is a now a must in running a business. Via their websites, companies relay important information to the readers. Amongst other things, they describe and advertise their products and services, sometimes relying on nifty (M-)similes to catch the eye of potential customers. The remaining categories represent

35 https://www.britannica.com/topic/journalism. Accessed 17 December 2020.

36 Another example is the informal discussion running online on whether question-and-answer websites like Quora fall into the category of social networking sites together with Twitter or Facebook. For detail, see https://www.quora.com/Is-Quora-a-social-media-site. Accessed 17 December 2020.

37 BoE stands for the Bank of English corpus.

38 According to Britannica, in its attempts to hold audience, present-day journalism includes an increasing number of "articles on the background of the news, personality sketches, and columns of timely comment by writers skilled in presenting opinion in readable form. By the mid-1960s most newspapers, particularly evening and Sunday editions, were relying heavily on magazine techniques, except for their content of 'hard news', where the traditional rule of objectivity still applied" https://www.britannica.com/topic/journalism. Accessed 17 December 2020.

some of the most interactive and most informal text varieties, which may involve emotionally tense situations or discourses of conflict (Naciscione 2010: 42); it is here that users mint M-similes to provide fresh and exciting perspectives, characterizations and evaluations, exchange wit, opinions, insults, questions and answers, share stories and experiences.[39] There is virtually no limit to the topics discussed on non-commercial websites/blogs of online communities (communities of people with shared interests), personal blogs (understood as online personal diaries), forums/discussion sites (websites, or sections of a website[40] which allow people to post messages and create open group conversation in the form of posted messages), social networks like Twitter and Facebook, and question and answer sites (Quora): they cover celebrities, food, home furnishing, personal accessories, nature and environment, relationships and emotions, spirituality, theater, travel, gaming and technology, etc. (see examples 28–31):

(24) *The sky was **pitch-black as ink**.* (iWeb: fiction)

(25) *The color red is related to the undead. Decomposing corpses often acquire a ruddy color, and this was generally taken for evidence of vampirism. Thus, the folkloric vampire is never pale, as one would expect of a corpse; his face is commonly described as florid or of a healthy color or dark, and this may be attributed to his habit of drinking blood. (The Serbians, referring to a redfaced, hard-drinking man, assert that he is "**blood red as a vampire**."* (COCA: nonfiction prose)

(26) *That is necessary to keep profits on track in an industry where margins can often seem **as wafer-thin as a slice of supermarket ham*** (Google: newspaper)

(27) *Old Growth Imperial Stout is **ebony dark as the night skies around winter solstice*** (Google: commercial website)

(28) *I'm sorry but it was too late. To Sheree its **crystal clear as the Maui waters** that Bob hasn't changed* (iWeb: RealityTea blog for Reality TV shows)

(29) *Prince you are beautifully masculine, composed and you are **as porcelain smooth as an Irish stout on a cool summer's evening**.* (Google: blog)

(30) *My experience with the HTC One is that this phone is **as rock solid as Mt. Everest**, in both build and operation* (iWeb: Brighthand forum)

39 For a useful discussion of chats and forums as suitable places for studying, amongst other things, the pragmatics and variability of PUs, see Kleinberger Günther (2006). For some more discussion on the usefulness of chats and forums in studying PUs from the aspect of "conceptual orality", see Mellado Blanco (2012).

40 https://techterms.com/definition/web_forum. Accessed 17 December 2020.

(31) *He is **pitch dark as the winter's night**. Tomorrow at 10:24 approved stallion Feel Good will start in the finals for 7 year olds* (Google: Facebook)

5 Conclusions

Taken at face value, *as*-similes appear quite unexciting. According to Moon (2008: 7), they are semantically simple since they correspond to the meaning of the adjective; pragmatically they are simple since they emphasize the degree of the property – with the supposition that the property is in reality or by convention "the canonical feature" of the source nominal. And yet, in her study of conventional *as*-similes Moon found that "what started as a deliberately simple and limited study [...] produced problems which not only prevented easy answers but seem to have implications for corpus-based phraseological studies in general" (2008: 3–4). Clearly, their apparent simplicity does not mean similes are not worthy of attention; among others, Hao and Veale's (2010) and Veale's (2012) studies are loud testimonies against this assumption. In this study, we chose to focus on the more creative twists and bends of *as*-similes that result in what we referred to as M-similes (Matryoshka-type similes). This is a special type of modification that involves a fusion of the old source noun and the adjective into a cognate noun-adjective compound (CNAC) – or replacement of the original adjective with the pre-existing CNAC – and insertion of a fresh source NP in its place. The outcomes of M-modification, like *blood-red as a raw steak* or *paper-thin as posters*, are best construed as occasionalisms (Langlotz 2006: 199) and not as future members of the established phraseolexicon. And, while we agree with cognitive psychologists' claim that "creativity is hard to measure or access, and is by definition unpredictable" (from Naciscione 2010: 43), we still attempted to "see beyond the originality" (Naciscione 2010: 65) and capture the essence of M-modification.

In the foregoing sections we presented a detailed analysis of the formal and distributional features of B-similes for their association with M-modification. We also examined the semantic, and to a lesser extent, discourse-related aspects of authentic examples of M-similes. The goal was to uncover modification patterns and constraints (Omazić 2015; Omazić and Čačija 2020), some of which may be typical of M-modification. All data were collected from three sources: the two tagged and parsed electronic corpora COCA and iWeb, and Google. We worked with three corpora to retrieve as many examples of B-similes and M-similes as possible but did not compare corpus data since technical limitations prevented sourcing B-similes directly from iWeb and Google. A set of 2 710 similes corresponding to the schema [*as* Adj *as* NP], which we called base-form similes

(B-similes), was mined from COCA and each was checked for the existence of M-modification. The iWeb and Google were then probed for the same B-similes and their possible M-modifications. There was not a complete overlap in the sizes of these three databases, since some of the B-similes mined from COCA were absent in iWeb and Google. Also, some analyses were only done on COCA and iWeb databases due to unreliability of the quantitative data from Google.

In the first part of our analysis, we compared the formal and distributional features of those B-similes that were found to M-modify and those that were not found to M-modify. Since these two subsets of B-similes were very different in size (and distribution) in all three databases, i.e. since M-modifying B-similes accounted for a small fraction of the complete databases, we used nonparametric statistics to assess the significance of the formal and distributional differences between various aspects of the two subsets of B-similes. Our assumptions that M-modification is associated with token-frequent B-simile, with B-similes that have token-frequent CNACs, with B-simile templates that already have stronger families of source NP alternatives, with B-similes whose source NPs are either more frequent generally or more frequently deployed as sources in B-similes, and with B-similes featuring monosyllabic adjectives were all confirmed at statistically significant levels (the latter finding was nonsignificant in only one database). The syllabic size of source nouns did not differ significantly between the two subsets of B-similes. Our results are consistent with claims that frequently used PUs may have lost their expressivity (Naciscione 2010: 39), that frequently occurring CNACs may also contribute to B-simile familiarity (Moon 2008: 32), and that idioms that are more frequent in corpora in modified form tend to consist of frequent words (Mancuso and Laudana 2019).

Semantic and discourse-related features had to be analyzed at token level due to their context-dependency. To make the analysis doable, we only analyzed tokens of M-similes. In other words, we did not attempt a comparison of the semantic and discourse-related features of B-similes that have, and those that have not been found to M-modify – which would be comparable to how we proceeded with features of form and distribution. Since M-similes characteristically involve a switch to a new source, it was interesting to first check whether the two sources differ in syntactic complexity. Whereas old sources were almost without exception single nouns, they were predominantly matched with more complex new source NPs, which sometimes featured elaborate structural modification. This undoubtedly contributes to their information load.

We then sorted old source, new source, and target entities into semantic domains and ranked those domains by frequency. Old and new sources were found to most often fall into two domains: Environment, Natural Materials and Products, and Man-Made Tools and Materials. Target entities most commonly

instantiated the domains Human Body, Human, and General Abstract entities. This reflects human concern with self and human tendency to construe Abstract Entities by comparison to more concrete ones. Interestingly, the most frequent domains of old sources differ from those in Norrick (1987), who found that regular *as*-similes most often involve animals as source entities, followed by natural products (corresponding partly to our Environment) and artefacts (corresponding partly to our Tools). But the two most frequent domains of sources that Norrick observed in independent CNACs, viz. Environment and Tools, did match the domains of our old sources in M-similes. This lends further credence to the proposed association between CNACs and M-modification.

Next, we examined the semantic domain match between any pair of nominal entities and all three entities at once. The comparison of pairs of nominals showed that most commonly they involve entities from different domains. This suggests that, when selecting new sources, speakers prefer to make the conceptual leap to a domain different from that of the old sources, and most commonly neither matches the domain of the target. Regardless of how frequent a domain was, our results also indicate that the lexical resources of each old source domain are not fully exploited. Most domains of old sources, especially the most frequent ones, involve one or possibly several overused lexemes. This may have contributed to the blunting of such frequently occurring old source NPs and eventually to their structural and conceptual "demotion" in the course of M-modification. A similar procedure was applied to the properties of the three nominals. Since comparisons can only work if the compared entities share common ground, it was unsurprising that in all pairwise comparisons and in the cumulative comparison of properties of all three nominals, the strongest case was one with both/all properties matching. The most frequent properties in all three nominals involve experientially-grounded, physical properties of matter, viz. Color, Thickness, Transparency, Tactile Resistance, and Textural Smoothness. More abstract properties, like Cognitive Clarity, Value and Quality occur close to the top of the frequency list with target entities, which is again unsurprising given that target entities themselves are very often abstract entities. When the properties of some of the three entities were found not to match, this was mainly due to the inherently polysemous nature of the adjective. This was not studied in detail but some interesting patterns of polysemy were detected, like the metonymic link between Color and Race in e.g. (of an organization – itself standing metonymically for its members) *as lily-white as the Mississippi Highway patrol* (iWeb) or the metaphtonymic link between Temperature and Emotion in e.g. (of a person) *as steel cold as a surgeon* (Google). As was the case for nominal entities, the lexical potential of the property domains was also underexploited, since one or two lexemes per property domain accounted for most M-simile tokens.

Finally, our analysis of text varieties showed that M-similes sit most comfortably in creative writing, typically fiction, interpretive news and magazine material, rather than matter-of-fact news reports (consistently with Moon 2008), product description and advertising on company/commercial websites, and in highly personal, interactive, often confrontational discourses typical of online community blogs, personal blogs, online discussion sites (cf. Kleinberger Günther 2006; Mellado Blanco 2012) and social networks.

Although we attempted to provide a broad and deep analysis of the many facets of M-similes, we have only scratched the surface. Future studies should test our findings against new data (new corpora), examine in finer detail the semantics of M-similes, including various patterns of polysemy and their motivating mechanisms. Perhaps most urgently, they should consider how the semantics of M-similes differ from the semantics of B-similes (or only B-similes from the established phraseolexicon), and how M-similes differ from other types of creative *as*-simile modifications.

Appendix

Bibliography of similes quoted from Google by order of appearance

(entries are prefaced by their respective similes, or full quotes if so required under copyright permission)

as stick thin as Twiggy

Devlin, Vivien. "Shout! The Mod Musical, Momentum Grand @ St. Stephens, Review." EdinburghGuide. Date published 9 August 2015. https://edinburghguide.com/festival/2015/edinburghfringe/shoutthemodmusicalmomentumgrandststephensreview-15873.

as blue and tissue-thin as my grandmother's vein-colored hands

Simon, Andrea. "Bashert: A Granddaughters Holocaust Quest", p. 44. Accessed 23 September 2021. https://books.google.hr/books?id=JHTCxBOn9eUC.

as stick-thin as a runway model

Bolch, Ben. "Phaler Looks to Bulk Up." Latimes. Date published 23 December 2002. https://www.latimes.com/archives/la-xpm-2002-dec-23-sp-hsbbkreport23-story.html.

as stick-thin as Twiggy

SWARA85. "Flab to Fit: My Journey on the Scale." Chai and CCD. Living, Loving, and Learning in India (blog). Date posted 3 May 2013. https://chaiandccd.wordpress.com/2013/05/03/flab-to-fit-my-journey-on-the-scale/.

as flour-fine as sand from Aram

Caylor, Duane K. "The day the rain began." Firstthings. Date published March 2012. https://www.firstthings.com/article/2012/03/the-day-the-rain-began.

lily-white as Norway
Berezow, Alex. "U.S. on Verge of Multi-Party System?" (originally posted on RealClearPolitics). Alexberezow. Date published 24 December 2014. https://www.alexberezow.com/u-s-on-verge-of-multi-party-system/.

steel cold as ice (Courtesy of author, Mary Gauthier)
Gauthier, Mary. "False from true." Mojim. Accessed 29 September 2021. https://mojim.com/usy141605x9x2.htm.

lobster-red as a sunburn
Rogers, Patrick. "A Bird's-Eye Perspective Can Find Beauty in the Planet's Dirtiest Places." NRDC. Date published 3 July 2018. https://www.nrdc.org/onearth/birds-eye-perspective-can-find-beauty-planets-dirtiest-places.

sand dry as fuck
Anonymous. "/ck/ - Food & Cooking". Warosu. Date posted 4 April 2019. https://warosu.org/ck/thread/12125750#p12126241.

three-dollar-bill-phoney as the Nevada neon strip itself (courtesy of Ray Robertson)
Robertson, Ray. 2003. "Mental Hygiene: Essays on Writers and Writing", p. 73. Accessed 25 September 2021. https://books.google.hr/books?id=GR6dQqxaE14C.

butter-yellow as a plumeria blossom
Steingold, Alison Clare. 2008. "The Uber Tuber." HanaHou: The Magazine of Hawaiian Airline 11(4). Hanahou. Date published September 2008. https://hanahou.com/11.4/the-uber-tuber.

as rock solid as a balloon made to look like a rock
SilvaDreams. "About as rock solid as a balloon made to look like a rock, myself and others poked plenty of holes in it and he ignored them completely … ." Warframe forum. Date posted 23 September 2016. https://forums.warframe.com/topic/687186-who-is-the-stalkerslight-spoilers/page/3/.

stone serious as nuclear war
Zee4321. "There's a lot of jokes about 'jeez guys it's never aliens calm down', but detecting an alien civilization should be as stone serious as nuclear war." Reddit. Date posted 15 August 2019. https://www.reddit.com/r/Futurology/comments/cqiepx/astronomers_have_detected_a_whopping_8_new/?utm_source=amp&utm_medium=&utm_content=comments_view_all.

as night and day different as passive and active
Brownworth, Russel. "The voice of God." Rocky Road Devotions. A few Minutes of Help for Today's Journey (blog). Date posted 25 May 2018. http://russellbrownworth.blogspot.com/2018/05/the-voice-of-god.html.

as nut-brown as a Greek Islander
Forster, Marcus. "Bodhisattva #3 – The Exile!". Beat Like Kerouac – now with child! Date posted 17 September 2011. https://beatlikekerouac.com/2011/09/17/bodhisattva-3-the-exile/.

as whistle-clean as dirndl-clad milkmaids and farm boys of the Alps
Von Busack, Richard. "Loserpalooza." Metroactive. Date published 3-9 November 2004. http://www.metroactive.com/papers/metro/11.03.04/loserpalooza-0445.html.

as whisper-soft as a layer of sheer Egyptian gauze
Review of Bunting, Eve. "I Am the Mummy Heb-Nefert." Publishersweekly. Accessed 23 September 2021. https://www.publishersweekly.com/978-0-15-200479-8.

milk-pale as moonlight
"A Kyranian myth: the hunter and the red king". Santharia. Accessed 28 September 2021. http://www.santharia.com/lore/hunter_and_the_red_king.htm.

nut brown as an Hermes bag
FiFi. "Pumpkin pilaf with ricotta & pepitas #GourmetTraveller." FiFi. Date posted 22 May 2018. https://fifi.com.au/2018/05/pumpkin-pilaf-with-ricotta-pepitas-gourmettraveller/.

lava-hot as a hot pocket
Dawn, Shannon. "Labor Day staycation ideas." Dawnsuzanneshannon. Accessed 28 September 2021. https://dawnsuzanneshannon.wixsite.com/dawn-says/single-post/2017/09/01/labor-day-staycation-ideas.

stick-thin as a mantis
Hicks-Jenkins, Clive. "Clive Hicks-Jenkins' Artlog:," Out of the Woods: part 2 (blog). Date posted 30 July 2017. https://clivehicksjenkins.wordpress.com/2015/07/30/.

rock solid as Optimus Prime's couch
McFee, Edwin. "Album Review: Tough Love, The Jimmy Cake." Hotpress. Date published 25 July 2017. https://www.hotpress.com/music/album-review-itough-lovei-the-jimmy-cake-20390595.

boulder-heavy as the moon
Lancaster, John. "The Wedding Speech". Fireriverpoets. Accessed 28 September 2021. https://fireriverpoets.org.uk/competitions/2016-results/the-wedding-speech/.

soot black as a raccoon's
Deirdre. "[Inside_Dierdre: Unrequited_Love,_Part_5]". Wearingthesechains. Date posted 10 March 2006. https://www.wearingthesechains.com/unrequited_love/.

satin-smooth as a prom queens thighs
Surfoverb. "glossy poly is sticky compared to matte or satin poly which is very slick. My fastest neck is poly satin-smooth as a prom queens thighs". How to sand an esquire neck so its like a Road Worn neck. Forum. Date posted 16 January 2014. https://www.tdpri.com/threads/how-to-sand-an-esquire-neck-so-its-like-a-road-worn-neck.456702/.

whip-quick as cinema in the 50s
Caldwell, Brendan. "The Indie Devs You Won't Hear About At E3." Rockpapershotgun. Date published 14 January 2016. https://www.rockpapershotgun.com/the-indie-devs-you-wont-hear-about-at-e3.

porcelain-pale as her kimono
"Viva la ichiruki fuck you." Hashtagartistlife. Date posted 26 December 2016. https://hitenbankai.tumblr.com/post/155046595968/hashtagartistlife-tw-abortion-they.

needle-sharp as wet seaweed
Tinker, Ian. Reader comment on 'The issue is Viking' by Kevin Learmonth. Shetlandtimes. Date posted 28 April 2011. https://www.shetlandtimes.co.uk/2011/04/27/the-issue-is-viking-kevin-learmonth.

porcelain smooth as an Irish stout on a cool summer's evening
Lace, Chantilli. "Chantilly Lace Retro Designs and Fashion," You've Got the Look! (blog). Date posted 3 July 2014. https://chantillylacevintage.wordpress.com/2014/07/03/youve-got-the-look/.

sandpaper-rough as a cat's (tongue)
Mulvany, Catherine. 2007. "Something wicked.", p. 318. Pocket Star Books. Accessed 29 September 2021. https://books.google.hr/books?id=f4_zx4xI0MwC.

porcelain-white as freshly-fallen snow
"Fanfic: Where's the Feather?, Ever After High". Fanfiction. Date posted 16 January 2016. https://www.fanfiction.net/s/11737306/1/Where-s-the-Feather.

Ink-black as witches' cats
Gregory, Charlie. "Tenerife." Poetrysoup. Year posted 2017. https://www.poetrysoup.com/poem/tenerife_880873.

whip smart as hell
Writegrrrl. "Rachel Leibrock," Back to school and other lessons in change (blog). Date posted 20 January 2017. http://www.rachel-leibrock.com/blog/2017/01/20/back-school-and-other-lessons-change.

marshmallow-smooth as butter
"Best of the Nexus till date." Customer review. Amazon. Accessed 29 September 2021. https://www.amazon.in/hz/reviews-render/lighthouse/B0179SM1L6?filterByKeyword=stock+android&pageNumber=1.

wafer-thin as a contact lens
"Veneers and inlays" ku64. Accessed 2 October 2021. https://ku64.de/en/aesthetic-dentistry-dental-aesthetics/.

honey-slow as a plant blooming (Courtesy of Mix Tabor)
Mix, Tabor. 2020. "Moonlight.", p. 117. Tabor Mix. Accessed 26.9.2021. https://books.google.hr/books?id=ChX9DwAAQBAJ&.

as stone-steady as the pyramids
Miller, Bryce. "Ryan Agnew, Aztecs offense show up at crucial time against Fresno State." sandiegouniontribune. Date published 16 November 2019. https://www.sandiegouniontribune.com/sports/sports-columnists/story/2019-11-16/aztecs-san-diego-state-football-fresno-state-offense-defense-mountain-west-ryan-agnew.

stone-still as a sarcophagus
Kaiser, Meghann. "The gathering throng parts like the Red Sea, and the patient, stone still as a sarcophagus, slides under the spotlights of the trauma bay." What didn't you realize would be an unintended consequence of being a doctor? Quora. Date posted 1 October 2019. https://www.quora.com/What-didnt-you-realize-would-be-an-unintended-consequence-of-being-a-doctor.

velvet-soft as petals
Kiddel, Ruby. 2010. "#notjustroses Lips as red and velvet soft as petals and a tongue as sharp as thorns. A rose by every other name." Twitter. Date posted 9 July 2010. https://twitter.com/eroticnotebook/status/18111737180.

razor sharp as stake knifes
Emma. "Similes." Highgateprimaryacademy. Date posted 29 September 2017. https://highgateprimaryacademy.net/y5y6ak2017/2017/09/29/similes.

crystal clear as a tide pool
Ligne St Barth. "Shower Gel HOMME As pure, fresh and crystal clear as a tide pool, the Men's Care Shower Gel is the irrefutable wake-up call of your cleansing ritual." Pinterest. Accessed 29 September 2021. https://www.pinterest.fr/pin/423056958739072839/.

otter-sleek as a blood slicked Stone
SammiSearle. "Sammi Searle's Blog," Old Girl Oblivion (blog). Date posted 25 August 2020. https://searleartblog.wordpress.com/2020/08/25/old-girl-oblivion/.

as wafer thin as the pizza
Jbcfc. "Experience was as wafer thin as the pizza." Tripadvisor. Date posted 26 September 2016. https://www.tripadvisor.co.uk/ShowUserReviews-g227947-d2197439-r425870782-Capriccio-Vilamoura_Quarteira_Faro_District_Algarve.html.

pitch dark as the winter's night
Hannell Dressage Stable. "He is pitch dark as the winter's night. Tomorrow at 10:24 approved stallion Feel Good will start in the finals for 7 year olds. Will you cheer for us?" Facebook. Date posted 12 July 2019. https://www.facebook.com/hannelldressagestable/photos/he-is-pitch-dark-as-the-winters-night-tomorrow-at-1024-approved-stallion-feel-go/3009750962399222/.

as ebony-dark as the soul of Satan
Supercat. "Fighting Spirit." Seakingsfemfight. Accessed 29 September 2021. https://www.seakingsfemfight.com/stories2011/storysupercat11.html.

ruby red as the desire of the Sanc Graal
Cawein, Madison. "The dream of sir Galahad." Sacred-texts. Accessed 28 September 2021. https://www.sacred-texts.com/neu/arthur/art049.htm.

steel-cold as a surgeon
Stringer, Arthur. 1920. "The Prairie Mother." First published by Bobbs-Merrill. Made available by Freeditorial. Accessed 29 September 2021. file:///C:/Users/Korisnik/Downloads/the_prairie_mother_by_arthur_stringer.pdf

petal-soft as the rays of the early sun
Nair, Edasseri Govindan. "Wedding Gift." Poetrynook. Accessed 29 September 2021. https://www.poetrynook.com/poem/wedding-gift%20.

iron-strong as a Greek destiny
Bulwer Lytton Lytton, Edward. "*Lucretia, Or the Children of Night*", p. 174. Accessed 28 September 2021. https://books.google.hr/books?id=4A1sBLrpEBsC.

dust-dry as an organic chemistry textbook
Review of "Inside Rupert's Brain" by Paul R. La Monica. Publishersweekly. Accessed 28 September 2021. https://www.publishersweekly.com/978-1-59184-243-9.

That is necessary to keep profits on track in an industry where margins can often seem as wafer-thin as a slice of supermarket ham (From the Financial Times. 6 September 2009. "A heftier

toll at the till." Felsted, Andrea. © The Financial Times Limited 2013. All Rights Reserved.) https://www.ft.com/content/72320f58-9b0f-11de-a3a1-00144feabdc0.

ebony dark as the night skies around winter solstice
"Caldera Old Groth Imperial Stout 22fl oz". morewines. Accessed 28 September 2021. https://morewines.com/caldera-old-groth-imperial-stout-22fl-oz.html.

References

Adams, Valerie. 1973. *An introduction to modern English word-formation.* London: Longman. https://doi.org/10.4324/9781315504254 (accessed 10 March 2021). Published online 21 June 2016.

Aisenman, Ravid A. 1999. Structure mapping and the simile-metaphor preference. *Metaphor and Symbol* 14(1). 45–51. https://doi.org/10.1207/s15327868ms1401_5 (accessed 26 March 2021). Published online 17 November 2009.

Barnden, John. 2012. Metaphor and simile: Fallacies concerning comparison, ellipsis, and inter-paraphrase. *Metaphor and Symbol* 27(4). 265–282. https://doi.org/10.1080/10926488.2012.716272 (accessed 26 March 2021). Published online 20 September 2012.

Barnden, John. 2015. Metaphor, simile, and the exaggeration of likeness. *Metaphor and Symbol* 30(1). 31–62. https://doi.org/10.1080/10926488.2015.980692 (accessed 26 March 2021). Published online 20 December 2014.

Biber, Douglas, Stieg Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English.* Harlow. Longman.

Biber, Douglas, Susan Conrad & Randi Reppen. 2000. *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press. Published online June 2012.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511814358. Published online June 2012.

Bredin, Hugh. 1998. Comparisons and similes. *Lingua* 105(1–2). 67–78. https://reader.elsevier.com/reader/sd/pii/S0024384197000302?token=8F3DBA67168E38B0E20B2AC9D832C2DACF8F54488F8DBDD2BED020163289EBC6DC5F4AE964F0FF5401D6D5458FB73102 (accessed 21 March 2021). Published online 17 August 1998.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511750526. Published online June 2012.

Carston, Robyn & Catherine Wearing. 2011. Metaphor, hyperbole and simile: A pragmatic approach. *Language and Cognition* 3(2). 283–312. https://doi.org/10.1515/langcog.2011.010 (accessed 21 March 2021). Published online 11 March 2014.

Chiappe, Dan L. & John M. Kennedy. 2000. Are metaphors elliptical similes? *Journal of Psycholinguistic Research* 29(4). 371–398. https://link.springer.com/content/pdf/10.1023/A:1005103211670.pdf (accessed 21 March 2021).

Chiappe, Dan L. & John M. Kennedy. 2001. Literal bases for metaphor and simile. *Metaphor and Symbol* 16(3–4). 249–276. https://doi.org/10.1080/10926488.2001.9678897 (accessed 22 March 2021). Published online 22 June 2011.

Corpas Pastor, Gloria. 2021. Constructional idioms of 'insanity' in English and Spanish: A corpus-based study. *Lingua 254 103013*. file:///C:/Users/Korisnik/Downloads/

Constructional_idioms_of_insanity_in_Eng.pdf (accessed 21 March 2021). Published online 10 February 2021.
Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA)*. Available online at http://corpus.byu.edu/coca/ (accessed January 2021).
Davies, Mark. 2018. *The iWeb Corpus*. Available online at https://www.english-corpora.org/iWeb/ (accessed January 2021).
Dixon, R. M. W. 2005. *A semantic approach to English grammar*, 2nd edn. Oxford: Oxford University Press.
Fauconnier, Gilles & Mark Turner. 2002. *The way we think: Conceptual blending and the mind's hidden complexities*. New York: Basic Books.
Fogelin, Robert J. 2011. *Figuratively speaking* (rev. edn). New York, NY & Oxford, UK: Oxford University Press. DOI: 10.1093/acprof:oso/9780199739998.001.0001. Published online May 2011.
Geeraert, Kristina. 2016. *Climbing on the bandwagon of idiomatic variation: A multi-methodological approach*. Edmonton, Alberta: University of Alberta dissertation. file:///C:/Users/Korisnik/Downloads/5251c2af-c210-4f10-b177-dba6db408387.pdf (accessed 20 February 2020).
Gentner, Dedre & Brian F. Bowdle. 2001. Convention, form and figurative language processing. *Metaphor and Symbol* 16(3–4). 223–247. https://doi.org/10.1080/10926488.2001.9678896 (accessed 21 March 2021). Published online 22 June 2011.
Giltrow, Janet & Dieter Stein (eds.). 2009. *Genres in the Internet. Issues in the theory of genre* (Pragmatics & Beyond New Series 188). Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/pbns.188
Givón, Talmy. 2001. *Syntax: An introduction, volume 1*. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/z.syn1
Glucksberg, Samuel. 2001. *Understanding figurative language: from metaphors to idioms*. Oxford: Oxford University Press. 10.1093/acprof:oso/9780195111095.001.0001. Published online January 2008.
Goldberg, Adele. 1995. *Constructions. A construction grammar approach to argument structure*. Chicago: The University of Chicago Press.
Goossens, Louis. 1990. Metaphtonymy: the interaction of metaphor and metonymy in expressions for linguistic action. *Cognitive Linguistics* 1(3). 323–340. https://www.degruyter.com/document/doi/10.1515/cogl.1990.1.3.323/html. Published online 8 October 2009.
Grzybek, Peter & Christoph Chlosta. 2009. Some essentials on the popularity of (American) proverbs. In Kevin J. McKenna (ed.), *The proverbial "Pied Piper". A festschrift volume of essays in honor of Wolfgang Mieder on the occasion of his sixty-fifth birthday*, 95–110. New York: Peter Lang. http://www.peter-grzybek.eu/science/publications/2009/grzybek_cc_2009_essentials-proverb-popularity.pdf (accessed 21 March 2021).
Hao, Yanfen & Tony Veale. 2010. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds & Machines* 20(4) [Special issue on computational creativity]. 635–650. https://doi.org/10.1007/s11023-010-9211-1 (accessed 21 March 2021).
Israel, Michael, Jennifer Riddle Harding & Vera Tobin. 2004. On simile. In Michel Achard & Suzanne Kemmer (eds.), *Language, culture and mind*, 123–135. Chicago: CSLI Publications. https://www.academia.edu/3995416/On_Simile (accessed 21 March 2021).

Ivorra Ordines, Pedro. 2022. Comparative constructional idioms. A corpus-based approach of the [*más feo que* X] construction. In Carmen Mellado Blanco (ed.), *Productive patterns in phraseology and Construction Grammar. A multilingual approach*, 29–52. Berlin: de Gruyter.

Kleinberger Günther, Ulla. 2006. Phraseologie und Sprichwörter in der digitalen Öffentlichkeit – am Beispiel von Chats. In Annelies Häcki Buhofer & Harald Burger (eds.), *Phraseology in Motion I. Methoden und Kritik*, 229–243. Baltmannsweiler: Schneider Verlag Hohengehren.

Kövecses, Zoltán. 2005. *Metaphor in culture. Universality and variation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511614408. Published online June 2012.

Lakoff, George. 1987. *Women, fire and dangerous things.* Chicago & London: The University of Chicago Press.

Lakoff, George & Mark Johnson. 1980. *Metaphors we live by*. Chicago: The University of Chicago Press.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Volume 1: Theoretical prerequisites.* Stanford, CA: Stanford University Press.

Langacker, Ronald W. 2008. *Cognitive Grammar: A basic introduction*. Oxford: Oxford University Press. 10.1093/acprof:oso/9780195331967.001.0001. Published online May 2008.

Langlotz, Andreas. 2006. *Idiomatic creativity: A cognitive-linguistic model of idiom representation and idiom variation in English* (Human cognitive processing 17). Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/hcp.17

Mancuso, Azzurra & Alessandro Laudana. 2019. Objective frequency values of canonical and syntactically modified idioms: Preliminary normative data. CLiC-it 2019. http://ceur-ws.org/Vol-2481/paper40.pdf (accessed 6 May 2020).

Maras, Lea. 2020. "*As blood-red as a raw steak*: a study of creativity in English *[as ... as]* similes. Unpublished bachelor's thesis. Osijek: Faculty of Humanities and Social Sciences of Osijek University.

Mellado Blanco, Carmen. 2012. Optimización de los recursos TIC en la fraseografía del par de lenguas alemán-español. In María Isabel González Rey (ed.), *Unidades fraseológicas y TIC. Madrid*: Instituto Cervantes (Biblioteca fraseológica y paremiológica, nº 2), 147–166. https://cvc.cervantes.es/lengua/biblioteca_fraseologica/n2_gonzalez/mellado.htm (accessed 15 March 2021).

Michaelis, Laura. 2003. Headless constructions and coercion by construction. In Elaine Francis & Laura Michaelis (eds.), *Mismatch: form-function incongruity and the architecture of grammar*. Standford: CSLI Publications, 259–310. https://www.researchgate.net/publication/277717445_Mismatch_Form-Function_Incongruity_and_the_Architecture_of_Grammar. Published online 22 May 2019.

Michaelis, Laura. 2004. Type shifting in construction grammar: An integrated approach to aspectual coercion. *Cognitive linguistics* 15(1). 1–67. https://www.degruyter.com/document/doi/10.1515/cogl.2004.001/html (accessed 22 March 2021).

Mieder, Wolfgang. 1989. *American proverbs. A study of texts and contexts* (Sprichwörterforschung 13). Bern & New York: Peter Lang.

Miller, Gary D. 2014. *English lexicogenesis*. Oxford: Oxford University Press. 10.1093/acprof:oso/9780199689880.001.0001. Published online April 2014.

Miller, George A. 1993 [1979]. Images and models, similes and metaphors. In Andrew Ortony (ed.), *Metaphor and thought*, 357–400. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139173865.019. Published online June 2012.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: A corpus-based approach.* Oxford: Clarendon Press.

Moon, Rosamund. 2001. The distribution of idioms in English. *Studi Italiani di Linguistica Teorica e Applicata*. 2001(2). 229–241.

Moon, Rosamund. 2008. Conventionalized *as*-similes in English. A problem case. *International Journal of Corpus Linguistics* 13(1). 3–37. doi 10.1075/ijcl.13.1.03moo (accessed 26 March 2021). Published online 19 January 2009.

Naciscione, Anita. 2010. *Stylistic use of phraseological units in discourse.* Amsterdam & Philadelphia: John Benjamins https://doi.org/10.1075/z.159 (accessed 21 March 2021).

Norrick, Neal R. 1987. Semantic aspects of comparative noun-adjective compounds. In Brigitte Asbach-Schnitker & Johannes Roggenhofer (eds.), *Neuere Forschungen zur Wortbildung und Historiographie der Linguistik. Festgabe für Herbert E. Brekle zum 50. Geburtstag*, 145–155. Tübingen: Gunter Narr Verlag.

Novoselec, Zvonimir & Jelena Parizoska. 2012. A corpus-based study of similes and cognate adjectival forms in English, Swedish and Croatian. In Antonio Pamies Bertrán, José Manuel Pazos Bretaña & Lucía Luque Nadal (eds.), *Phraseology and discourse: Cross linguistic and corpus-based approaches*, 101–110. Baltmannsweiler: Schneider Verlag Hohengehren.

Omazić, Marija. 2002. O poredbenom frazemu u engleskom i hrvatskom jeziku [On idioms of comparison in English and Croatian]. *Jezikoslovlje* 3(1–2). 99–129. https://hrcak.srce.hr/31348 (accessed 20 February 2020).

Omazić, Marija. 2003. The metacommunicative setting of phraseological units and their modifications – evidence from the British National Corpus. In Dawn Archer, Paul Rayson, Andrew Wilson & Tony McEnery (eds.), 599–602. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster: Lancaster University. http://ucrel.lancs.ac.uk/publications/CL2003/papers/omazic.pdf (accessed 21 March 2021).

Omazić, Marija. 2007. Patterns of modifications of phraseological units. In Annelies Häcki Buhofer & Harald Burger (eds.), *Phraseology in motion II. Theorie und Anwendung. Akten der internationalen Tagung zur Phraseologie*, 61–108. Hohengehren: Schneider Verlag.

Omazić, Marija. 2015. *Phraseology through the looking glass*. Osijek: Faculty of Humanities and Social Sciences, University J. J. Strossmayer.

Omazić, Marija & Romana Čačija. 2020. Dynamic model of PU modification. In Marija Omazić & Jelena Parizoska (eds.), *Reproducibility and variation of figurative expressions: Theoretical aspects and applications*, 51–67. Białystok: University of Białystok Publishing House. file:///C:/Users/Korisnik/Downloads/1085928.Omazi_Parizoska_IDP_5.pdf (accessed 21 May 2020).

Ortony, Andrew. 1993 [1979]. The role of similarity in similes and metaphors, 2nd edn. In Andrew Ortony (ed.), *Metaphor and thought*, 342–356. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139173865.018. Published online June 2012.

Petrova, Oksana. 2011. *Of pearls and pigs: A conceptual-semantic tiernet approach to formal representation of structure and variation of phraseological units*. Åbo: Åbo Akademi University Press. https://www.academia.edu/1209057/Of_pearls_and_pigs_a_conceptual_semantic_Tiernet_approach_to_formal_representation_of_structure_and_variation_of_phraseological_units (accessed 12 March 2021).

Rosch, Eleanor & Carolyn B. Mervis. 1975. Family resemblances. Studies in the internal structure of categories. *Cognitive Psychology* 7(4). 573–605. https://www.sciencedirect.com/science/article/pii/0010028575900249 (accessed 22 March 2021).

Rosch, Eleanor. 1977. Human categorization. In Neil Warren (ed.), *Studies in cross-cultural psychology*, Vol. 1, 1–49. London: Academic Press.

Sommer, Elyse. 2013. *Similes dictionary*, 2nd edn. Visible Ink Press.

Taylor, John. 2012. *The mental corpus: How language is represented in the mind*. Oxford: Oxford University Press. 10.1093/acprof:oso/9780199290802.001.0001. Published online September 2012.

Tirrell, Lynne. 1991. Reductive and nonreductive simile theories of metaphor. *The Journal of Philosophy* 88(7). 337–358. https://www.jstor.org/stable/2027089 (accessed 21 March 2021).

Todd, Zazie & David D. Clarke. 1999. When is a dead rainbow not like a dead rainbow? Investigating differences between metaphor and simile. In Lynn Cameron & Graham Low (eds.), *Researching and applying metaphor*, 249–268. New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139524704. Published online October 2012.

Utsumi, Akira. 2011. Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science* 35(2). 251–296. https://onlinelibrary.wiley.com/doi/10.1111/j.1551-6709.2010.01144.x (accessed 22 March 2021).

Veale, Tony. 2012. A computational exploration of creative similes. In Fiona MacArthur, José Luis Oncins-Martínez, Manuel Sánchez-Garcia & Ana María Piquer-Píriz (eds.), *Metaphor in use. Context, culture and communication*, 329–343. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/hcp.38.23vea. Published online 17 October 2012.

Vo, Thuc Anh. 2011. *Idiomatic creativity: A pragmatic model for creative idiomatic uses in authentic English discourse*. Nottingham, UK: University of Nottingham dissertation. http://eprints.nottingham.ac.uk/14388/1/555406.pdf (accessed 19 February 2020).

Wikberg, Kay. 2008. Phrasal similes in the BNC. In Sylvianne Granger & Fanny Meunier (eds.), *Phraseology: An interdisciplinary perspective*, 127–142. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/z.139.14wik. Published online 1 June 2008.