UNIVERZITA KARLOVA

Filozofická fakulta

Ústav anglického jazyka a didaktiky

Filologie – anglický jazyk

Disertační práce

Mgr. Jaroslav Emmer

Anglická přirovnání v korpusových datech

English Idioms of Comparison in Corpus Data

Vedoucí práce: prof. PhDr. Aleš Klégr

2023

Prohlašuji, že jsem disertační práci napsal samostatně s využitím pouze uvedených a řádně citovaných pramenů a literatury a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 22. 9. 2023

<div style="text-align: right">Mgr. Jaroslav Emmer v.r.</div>

Abstrakt

Tato práce se věnuje korpusovému zkoumání adjektivních přirovnání. Jelikož přirovnání jsou často zastíněna metaforou, dosavadní výzkum přirovnání na základě empirických dat není příliš rozsáhlý. Předkládaná práce přináší komplexní rozbor adjektivních přirovnání v korpusových datech, přičemž pracuje s různorodými korpusy. Metodologie získávání adjektivních přirovnání z korpusů je detailně popsána včetně souvisejících úskalí. Vzhledem k tomu, že každý korpus vyžaduje jiný způsob dotazování, jsou CQL dotazy prezentovány pro jednotlivé korpusy zvlášť. Korpusové šetření vyústilo v seznam 309 adjektivních přirovnání, a získaná data jsou podrobena formální a obsahové analýze. Po statistickém přehledu dat v rámci jednotlivých korpusů následuje detailní výklad 60 nejfrekventovanějších přirovnání zahrnující jejich význam a funkci, typická komparanda, související přirovnání a dílčí frekvence. Tato přirovnání jsou poté vyhledána v renomovaných online slovnících. Výsledky ukazují, že slovníky opomíjejí mnoho přirovnání, která jsou v korpusových datech frekventovaná. Formální analýza dokládá, že nejčastější anglická adjektivní přirovnání preferují jednoslabičné komponenty. Z obsahové analýzy vyplývá, že přirovnání lze členit do širších skupin podle tertia comparationis (pět podskupin), nebo komparáta (tři podskupiny), ačkoliv se jedná o skupiny velmi obecné. Nakonec je získaný vzorek anglických přirovnání srovnán s českým, což ukazuje významný lexikální překryv. Adjektivní přirovnání nejsou ve své prototypické formě příliš frekventovaná, přesto však zůstávají ve slovní zásobě pevně ukotvená. Ta nejčastější pak patří do tzv. ‚minima přirovnání‘, tedy souboru přirovnání, který je znám zkušeným mluvčím daného jazyka. Tato práce předkládá komplexní popis adjektivních přirovnání na základě empirických dat, nicméně je potřeba další výzkum, zejména v oblasti jejich transformací (konkrétně kompozitních forem adjektivních přirovnání), jejichž frekvence je mnohdy několikanásobně vyšší.

Klíčová slova: CQL dotazy, frazeologie, frekvence, idiomatičnost, korpusy, přirovnání, slovníky

Abstract

This thesis presents a corpus-based investigation that focuses on adjectival similes. As similes are often overshadowed by metaphors, the existing research on similes using empirical evidence is relatively scarce. The presented work provides a complex account of adjectival similes in corpora using data from multiple heterogeneous sources. The methodology for mining adjectival similes from corpora is thoroughly described, along with the associated pitfalls. As every corpus requires a different approach, the CQL query designs are presented individually for each researched corpus. The corpus-based mining yielded a list of 309 unique adjectival similes. The obtained data are subjected to meticulous scrutiny in the form of both formal and content analyses. Following a statistical overview of the data collected from each corpus, the 60 most frequent similes are presented in detail, including their meanings and functions, typical targets, related similes, and frequencies. These similes are then searched in respected online dictionaries, and the findings suggest that dictionaries struggle to keep up with corpus evidence. The formal analysis shows a strong preference for monosyllabic constituting elements. Regarding their content, adjectival similes can be divided into five general ground-centred groups and three source-centred ones, suggesting that general patterns exist despite sometimes being very broad. Lastly, the English simile sample is compared to a Czech one, revealing a significant lexical overlap. Adjectival similes in their prototypical form are relatively infrequent compared to other idiomatic multi-word units. However, they are firmly established in the English lexicon, and the most frequent examples belong to the 'similes minimum' – similes any proficient speaker of the language would know. While this work represents a complex presentation of adjectival similes based on empirical evidence, further research is necessary in the area of their transformations (namely compound adjectival similes), whose frequencies are often much higher.

Keywords: corpora, CQL queries, dictionaries, frequency, idiomaticity, phraseology, simile

Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
AF	Absolute Frequency
BNC	British Nation Corpus
BoE	Bank of English
C/C	Comparison-Categorization
Cam	Cambridge
CAS	Compound Adjectival Simile
COCA	Corpus of Contemporary American English
Col	Collins
CQL	Contextual Query Language
EI	English Idioms
EST	Elliptical Simile Theory
EW	English Web
FSP	Functional Sentence Perspective
i.p.m.	Instances Per Million
MW	Merriam-Webster
MWU	Multi-Word Unit
ODEP	Oxford Dictionary of English Proverbs
OUP	Oxford University Press
Ox	Oxford
PoS, pos	Part of Speech
RF	Relative Frequency
SiS	Simile Stack
t.p.h.	Tokens Per Hundred
T/T	Token-Type
TFD	The Free Dictionary

I.   INTRODUCTION

1 Aim and scope of this work

The introductory section presents the research background and sketches the type of research used in this work. Furthermore, it provides commentary on the formal presentation of adjectival similes.

1.1 Research background

This work deals with idioms of comparison. Similes are often viewed as figures of speech belonging to the literary domain. However, speakers also use them in everyday communication because they have a range of functions and effects, making them valuable tools for various purposes. They can emphasise, entertain, impress and can be used to praise, assess, criticise, offend and jest. They have standardised forms but also novel variants and other transformations. This perhaps makes them appear less frequent, but they typically stand out when used.

The aim of this work is to collect a representative sample of adjectival similes in present-day English, describe it, look at the representation of adjectival similes in English dictionaries and compare the English sample with a Czech one. At the same time, the study aims to describe the methodology of searching for adjectival similes in corpora and its associated pitfalls. Additionally, several complementary goals are presented in section 4.4, along with elaboration.

For several reasons, purely linguistic studies researching adjectival similes are relatively scarce. Firstly, they are traditionally associated with literature. Secondly, they are sometimes overshadowed by metaphors and dismissed as nothing but explicit variants. Thirdly, adjectival similes are multi-word units and require larger datasets to provide enough instances for an investigation.

Fortunately, the technological advancement in the 2020s is unparalleled, allowing us to use artificial intelligence, the internet and large amounts of language data stored in various corpora. Given the nature of adjectival similes, their investigation in a single corpus usually yields only specific data whose general validity is difficult to establish. Therefore, this work explores multiple corpora to better capture the nature of adjectival similes and their occurrence in heterogeneous environments.

1.2 Research specification

The research conducted in this work uses corpus linguistics as a methodology for data mining. The initial step was determining and specifying the researched items. Once the criteria had been established, the following step was the selection of corpora used for data mining (explained in Section 4.1). Next, a corpus-based extraction of adjectival similes was conducted, followed by a thorough data analysis comprising both a qualitative and a quantitative survey of the findings. The final part reviews the research questions and hypotheses in light of the acquired data.

Initially, researching adjectival similes appears convenient as the formal frame makes it easy to design a basic query for a corpus search. However, the formal frame is not exclusive to adjectival similes. Consequently, it is necessary to formulate criteria delimiting what is considered an adjectival simile. Prototypical examples are clear-cut, but other forms exist that can be considered adjectival similes depending on the criteria. This work takes a strict approach to defining adjectival similes and considers related forms transformations.

In my opinion, transformations (especially CAS, see Section 5.3) belong to the area of simile research as they fundamentally represent the same phenomenon semantically and functionally. Nevertheless, transformations are not the primary focus of this work and are only included in specific considerations related to individual adjectival simile types.

1.3 Formal presentation of adjectival similes and examples

Adjectival similes in the prototypical form are notable for their formalised frame, making them easily recognisable, especially when they occur with the initial *as*. The issue lies in the initial *as* being an optional component in adjectival similes. In this study, the presentation of similes in the text adopts the small uppercase script used for conceptual metaphors, with the initial *as* in the brackets: (AS) FRESH AS A DAISY. In tables and other graphical representations, the initial *as* is omitted to save space and make the data easier to survey.

The uppercase script represents a standard simile as a paradigmatic unit of the lexicon. Language examples are presented in italics to distinguish between paradigmatic ((AS) FRESH AS A DAISY) and syntagmatic instances (*she was fresh as a daisy*).

II. RESEARCH CONTEXT

2 Similes

This section is dedicated to the description of similes and their general purpose. They are contrasted with literal comparisons and metaphors, and their status of figurative multi-word units is reviewed and explained. Furthermore, three general classification dichotomies are reviewed, followed by an outline of simile functions and effects.

2.1 Idioms of Comparison

In our experiencing of the world, we are bound to develop a method of how we understand the world around us, different images, sounds, relations and other things. In order to understand it all, we have to design, or at least inherit, specific strategies that help us with our perception of the vast number of stimuli that we naturally seek to understand. One of the common ways to approach this issue is to see the world as a network of interrelated phenomena. As a result, we seek relations among things and design expressions to represent such relations. From early childhood, we notice that certain entities are similar in shape, colour or smell, among just a few.

> Making comparisons is a very human occupation. We spend our lives comparing one thing to another, and behaving according to the categorisations we make.

(Dienhart, 1999: 98)

However, to argue as Davidson (1978: 39) or Searle (1979: 106) that 'anything is similar to anything in some regard' is probably too broad a generalisation, not helping us whatsoever to determine what makes certain similarities significant enough to merit cognitive attention and, even more so, lexical representation.

Since comparing is vital for understanding the world, it is also reflected in language. As a result, probably all languages have a plethora of grammatical and lexical means used for comparing that their speakers recognise as established and conventional. One traditionally recognised category of set expressions serving this purpose is called idioms of comparison, or even more commonly, *similes*. Similes can generally be described as idiomatic units of varying form and meaning whose purpose is to reinforce the feature ascribed to the *target*, the subject of comparison, by comparing it to a prototypical bearer of such a feature – the *source*. The shared feature is either explicit or implicit and may display a varying degree of prominence. It is explained later in this work what makes

certain features felicitous in given contexts despite sometimes being far from typical of the source.

One of the obvious questions is why we resort to comparison instead of mere description. Norrick (1987: 146) makes an important point when he says that "for purely cognitive reasons, speakers store and recycle similes for properties they find difficult to describe digitally." In this sense, 'digital' stands for the direct description of a feature attributed to the subject instead of analogous, where we choose to describe one entity by comparing it to another, more salient entity by foregrounding their shared feature. As typically difficult to describe digitally, Norrick (1987: 146) mentions concepts related to sensory perception, such as colour ((AS) BLACK AS PITCH), temperature ((AS) COLD AS ICE), or texture ((AS) SOFT AS SILK); but also emotive or taboo concepts, such as inner feelings ((AS) HAPPY AS LARRY) and intoxication ((AS) PISSED AS A FART). The lack of digital description may explain the existence of many similes; however, the importance of this argument should not be overstated. Humour, hyperbole, or sarcasm undeniably motivate quite many (un)conventional similes for which the lack-of-digital-description argument would not suffice, for example, (AS) THICK AS PIG SHIT, (AS) DEAF AS A POST or (AS) SHARP AS A BOWLING BALL. The range of functions of similes is explored in more detail in Section 2.6.

Similes are multi-word units (MWUs) that can (but need not) be idiomatic. However, defining what is to be considered an idiom is quite tricky for traditional and present-day phraseology approaches. Seidl & McMordie (1988: 12-13) provide us with the simplest explanation: "An idiom can be defined as a number of words which, when taken together, have a different meaning from the individual meanings of each word." Similarly, Lipka (1992: 96) refers to the process of idiomatization as "the addition or loss of semantic features." However, such an account is insufficient because it focuses only on the semantic aspect. It would rule out many similes whose component words cannot be said to have abandoned the original meaning within the combination, for instance, (AS) WHITE AS A SHEET or (AS) FREE AS A BIRD.

It is imperative to consider word combinations' semantic and formal aspects before we (dis)qualify them as idiomatic. Palmer (1976: 98-99) elaborates that idioms are a special type of collocation whose meaning is opaque, at least to a certain degree. He also points out that idioms are often restricted grammatically and syntactically, but these restrictions vary significantly. Cruse (2006: 82) lists two main features of prototypical idioms: "they are non-compositional, and they are syntactically frozen." Nevertheless,

these features are scalar, and there are combinations where particular syntactic manipulation is permitted without breaking the idiom; for example, *The shortest straw has been pulled for you*. Makkai (1972: 25) distinguishes between 'encoding' and 'decoding' idioms. The former are generally understood without prior knowledge of the expressions, such as *fast asleep* or *high time*; the latter are viewed as unpredictable or even obscure when heard the first time, such as *cold feet* or *black sheep*. Admitting that providing an all-encompassing definition is nearly impossible, Čermák points out the following:

> Still it is evident that all the features discussed, and others that could be mentioned, cannot be squeezed into an unambiguous specific definition applying only to the phraseme and idiom. The very status of these features, defined negatively as anomalies, which are perceived against the background of the regularities of language, is relative. It depends on the elaboration of the concept of regularity (or regular generability) and its explicit meaning in the linguistic description according to a specific coherent theory of language.

(Čermák, 2007: 83)

Lastly, Gray & Biber (2015: 126, 144), representing the distributional, corpus linguistic approach to phraseology, confirm that the nature of word combinations and their idiomatic status is a relevant aspect for corpus research. Again, the difference is presented as bipolar, with combinations being either idiomatic or non-idiomatic but significantly frequent. They also highlight that pre-defined idioms are typically examined in corpus-based, top-down studies, while non-idiomatic word strings tend to be products of corpus-driven, bottom-up design research.

Idioms can, therefore, be described as lexical expressions that are formally anomalous and possibly also semantically outstanding. Čermák (2007) uses the term 'phraseme' for the formal combination and 'idiom' for the semantic aspect of the combination to distinguish between the two areas. While this may at first seem like an unnecessary distinction, many collocations can be considered phrasemes (i.e. fixed combinations) formally, but the semantic idiomaticity is debatable, such as in *take a seat* or *make a decision*. To conclude the discussion about idiomaticity, it is essential to note that it is a scalar feature most prominent in highly idiomatic combinations often cited as prototypical members of the idiom category. However, many other types of multi-word units exhibit various degrees of idiomaticity and should be equally included in any excursions into the idiom(aticity) domain.

2.2 The simile interpretation of metaphor

Many scholars, starting with Aristotle, argued that metaphors are just elliptical similes and tried to support this view with various arguments. The underlying idea of what may be called Elliptical Simile Theory (henceforth EST), which is sometimes referred to as Naïve Simile Theory (its revised version being Figurative Simile Theory), is that all metaphors are trimmed variants of similes. The trimmed content allows us to dispense with the ground and the comparator, leaving the ground for the hearer to extract. This chapter briefly looks at this concept and critically re-evaluates its validity.

O'Donoghue (2009: 125) observes that even today, "theoretical thinking is sharply divided on one central issue: whether [metaphors and similes] are indistinguishable in meaning and so interchangeable, or altogether different in their effects." While it is obvious what O'Donoghue seeks to emphasise and the phenomenon may be viewed as bipolar, the issue is not strictly black and white. Specific simile instances may be considered explicit variants of metaphors whose comparison is by design implicit, such as *James is (like) a pig*. It would, however, be naïve to assume that *James is a pig* and *James is like a pig* are indistinguishable and invariably interchangeable. One of the crucial things to consider here is that metaphors do not provide us with a ground, and it is up to us to reconstruct the likeness to correctly grasp the metaphor, which may sometimes be very difficult. Beardsley argues the following:

> The metaphor is full and rich, apart from any context; indeed, the function of the context is rather to eliminate possible meanings than to supply them. A metaphor is not an implied comparison.
>
> Beardsley (1958: 138)

The first part is also true of implicit similes (see Section 2.5.2), in other words, similes that do not provide us with an explicit ground, such as *John is like a lorry*. In this sense, implicit similes behave like novel metaphors. They are difficult to understand without context until/unless they become recurrent expressions with a conventionalised sense. Once they become associated with a specific ground, the difficulty of interpretation disappears along with the novelty. Gargani (2014: 3) illustrates the different degree of conventionalisation (novelty) with the metaphors *Achilles is a lion* and *Achilles is a gazelle* and states that while *lion* is a common source employed in metaphors with pre-set salient features and qualities associated with lions, *gazelle* is much less frequent and allows for a somewhat subjective interpretation. It follows from this that conventional

metaphors and conventional implicit similes might be similar (not equal) to explicit similes, whereas non-conventional (nonce) metaphors and non-conventional implicit similes remain open to interpretation and, therefore, cannot be assigned a corresponding explicit simile.

In his chapter *Metaphor*, Black (1962: 36) notes the popularity of EST in the first half of the 20th century. He addresses some of the key questions related to the recognition and understanding of metaphors, one of which is also "What are the relations between metaphor and simile?" (Black, 1962: 25). The likeness is an obvious connection; both figures employ similarity or analogy; however, each figure uses different means of arriving at the desired meaning. Simile exploits a pre-set frame or pattern with an explicit ground (unless implicit), which provides the quality shared by the target and the source. Metaphor, on the other hand, obscures the ground, and the hearer must use other clues (context, background knowledge, embeddedness) to find the similarity or analogy. Such description can easily lead to the conclusion that simile and metaphor differ only in form and thus, have us accept the EST view. However, as was explained earlier, this simplified view only holds in the case of simple, conventional metaphors (sometimes called 'dead' metaphors) as the difficulty of finding a corresponding simile is striking when confronted with complex, novel metaphors.

Davidson (1978) comments on EST and highlights some of the main flaws of this approach. He too recognises that "[t]here is (…) the difficulty of identifying the simile that corresponds to a given metaphor" (Davidson, 1978: 38). It is undoubtedly possible with metaphors such as *You're a dog*, but many literary metaphors are too elaborate. Davidson makes a valid point regarding the comparison of metaphors to similes:

> Just because a simile wears a declaration of similitude on its sleeve, it is (…) far less plausible than in the case of metaphor to maintain that there is a hidden second meaning. In the case of simile, we note what it literally says, that two things resemble one another; we then regard the objects and consider what similarity would, in the context, be to the point.

> (Davidson, 1978: 40)

The main issue with Davidson's argumentation is his rejection of figurative content. He considers only the literal similarity or analogy and, to him, anything beyond that is not propositional. It follows from Davidson's views that in similes such as (AS) DUMB AS BRICKS, we are to seek literal similarity, which will, in turn, trigger a specific association and help us understand the intended meaning. However, there are two main objections to

21

this theory. Firstly, it is difficult to imagine such literal comparisons to be arbitrary. There must be some degree of conventionality or embeddedness in language. Secondly, it is highly improbable that random literal comparisons would successfully capture the actual similarity provided by the ground. In the previously mentioned simile, *people* (usually the target) are compared to *bricks* (source). If we take the literal comparison to be the norm, *dumb* (ground) is the feature that is attributed to the source; therefore, *bricks* are supposed to be an apt bearer of that quality and hence make a good candidate for highlighting the similarity of *people* and *bricks*. From the perspective of *how dumb*, BRICKS is to provide the ideal bearer. It is easy to see how this explanation fails to justify the existence of the conventionalised simile (AS) DUMB AS BRICKS. Anything that cannot speak is canonically *dumb*, but that is not a salient feature of inanimate objects. One possible explanation is offered by Norrick (1987: 149), who describes this phenomenon as 'salience imbalance', in other words, the salient properties typically associated with the target and the source are mutually incompatible.

William Lycan (2001: 180) too mentions the Figurative Simile Theory and notes that "[s]imiles are often, perhaps usually, figures of speech", and adds "that literal similarity is symmetric". In other words, if *A* is similar to *B*, then *B* is similar to *A*. In this theory, similes interpreted figuratively are the bearers of cognitive value. Lycan comments on Fogelin's Figuratively Speaking (1988) and his account of figurative similes to provide further details. Fogelin calls forth salience to help us explain the non-symmetrical relation in similes. A particular feature of the target is brought to attention by using a ground that is a typical (defining or at least salient) property of the source employed in the simile. Figure 1 illustrates the three main entities that constitute adjectival similes:

The car[1] is (as) cheap[2] as chips[3]

1) **Target** (comparandum) = the subject of the proposition, i.e. the entity described.
2) **Ground** (tertium comparationis) = the salient feature(s) of the comparatum that allow(s) it to function as a good source for the given subject.
3) **Source** (comparatum) = the object to which the subject is likened, i.e. the entity used to highlight a particular feature.

Figure 1. The main entities in (adjectival) similes.

Fogelin's model seems viable for explicit similes but fails to provide a convincing argument for why metaphors should be treated as trimmed versions of figurative similes.

Regarding the idea of two separate figures expressing the same content, common sense invites us to apply the principle of contrast[1], which is often discussed in relational semantics, to describe what is referred to as 'absolute' or 'total' synonymy. Discussing absolute synonymy, Cruse (2011) argues that languages do not need multiple expressions with the same meaning. Consequently, the originally identical units diverge in conceptual or associative meaning. This idea, of course, is not limited to individual words only and also applies to multi-word expressions. Nevertheless, O'Donoghue (2009: 127) aptly states that "not only is our thinking about metaphor and simile intertwined, but our thinking in general [is] conditioned by the process of comparing one thing to another." We design our judgments and evaluations, linguistic or other, with the help of established concepts, which serve as a springboard for creating new ideas whose novelty is determined through confrontation with what we already recognise as established. This premise always seems to bring theorists back to the idea that the simile/metaphor issue is bipolar and that a choice between the two options has to be made to develop further argumentation. O'Donoghue (2009: 127) mentions that based on Lakoff's (1980) conceptual metaphor account, it would be logical to speak of conceptual simile as long as we accept the view that metaphor is nothing more than an elliptical simile. Moreover, even if we dismiss the elliptical simile interpretation, the idea of conceptual similes still remains valid.

Another important aspect of comparing metaphors and similes is separating the form from the meaning:

> There seems to me to be an important, and often overlooked, distinction between linguistic explicitness and explicitness of meaning: similes are certainly linguistically flagged in a way that metaphors are not, encouraging the hearer to embark on a process of comparison; but they are no more specific in pointing to intended meaning.

O'Donoghue (2009: 143)

Once again, the vital aspect is conventionalisation. Metaphors and similes may be established expressions that are retrievable and repeatedly used by speakers. Conversely, certain metaphors and similes are designed to impress, puzzle or shock and are not intended to become part of the lexicon. Therefore, we must consider complexity and

---

[1] In Construction Grammar, the same notion is recognised (and frequently cited) as *the Principle of No Synonymy*. The principle represents the assumption that no two distinct forms can be considered semantically and pragmatically identical (Goldberg, 1995: 67; Gilquin, 2010: 97-98).

retrievability in any discussion about the likeness between metaphors and similes. Our intuition and experience, while invaluable in some cases, are unreliable for establishing frequency due to a lack of data and its appropriate interpretation. As a result, it is difficult for an individual to assess which metaphors and similes are retrievable and which are just one-off constructions. Corpus research is an excellent tool for tracking the frequency of multi-word combinations and may help us identify the recurring metaphors and similes that form the stock shared by speakers.

There remains one more issue related to the metaphor that ought to be discussed. It was already mentioned that metaphors could be implicit comparisons inviting the reader or listener to extract the ideal ground. However, the comparison does not always need to be a metaphor's purpose. Glucksberg and Haught (2006) note that there are two principal theories of metaphor in psycholinguistics: the comparison and the categorisation theory. The comparison theory closely resembles the EST. The difference between metaphors and similes is mainly that of explicitness of the linguistic form and possibly a different rhetorical effect. From the categorisation perspective, "metaphors and similes are understood in their own right: the metaphor as a categorisation assertion, the simile as an assertion of similitude" (Glucksberg and Haught, 2006: 361). This does not necessarily mean that a metaphor cannot aim to compare, but, at the same time, the comparison should not be seen as the sole purpose of metaphors. Glucksberg and Haught (2006: 375) illustrate the difference between a metaphor and a simile with the example *I was like a sardine* arguing that *I was a sardine* would not sufficiently communicate the intended concept [to be packed like sardines in a can]. In conclusion to their study, Glucksberg & Haught admit that using both the categorical form (metaphor) and the comparison form (simile) can result in the same interpretation of meaning; however, some metaphors may differ from similes. That is a crucial point for the following argument:

> If metaphors cannot always be paraphrased as similes, then metaphors cannot, in principle, be understood in terms of their corresponding similes, and vice versa. This means that comparison theories of metaphor comprehension, which rest on the assumption that metaphors and smiles are equivalent, are fundamentally flawed.

> (Glucksberg & Haught, 2006: 376)

Barnden (2009: 81) takes it one step further by arguing that "non-paralleled items are often crucial to the metaphorical effect." He demonstrates with examples that not every linguistic metaphor requires mapping to be understood and attributes the lack of attention

paid to non-parallelism in theories of metaphor to vague descriptions of how context is used in the interpretation of metaphors. He provides this example:

'I don't think strings are attached. If there are any they're made of nylon – I can't see them'

Barnden (2009: 79)

In this example, Barnden argues that while the stock metaphor NO STRINGS ATTACHED utilises mapping of *strings* onto the target-domain of *constraints* to be understood, *nylon* neither requires nor receives any mapping and still contributes to overall understanding – the 'difficulty of seeing any constraints' is readily grasped by the listener/hearer. Here is how Barnden explains the lack of mapping of *nylon*:

The crucial carry-over here is that of difficulty. Notice carefully that this does not of itself imply that the *translucency* [of nylon] needs to be mapped to the target scenario. Even less does it imply that being-made-of-nylon itself needs to be mapped (similarly for being thin ...) and, even less again, that being-made-of and nylon need separately to be mapped.

Barnden (2009: 83)

It follows that if no mapping occurs, we cannot possibly find a corresponding simile. Such a line of argumentation alone (with many other objections presented earlier) can be treated as conclusive evidence that metaphor and simile are different phenomena that may overlap in some regard but cannot be considered two manifestations of a single figure.

2.3 Literal comparison vs simile

In this work, similes are considered (idiomatic) phraseological units of varying degrees of conventionalisation, both syntactic and semantic. Consequently, the established part-of-speech sequences allow us to harvest most similes from corpora with a predesigned query.[2] However, these syntactic frames are not exclusive to similes, and the query often yields many irrelevant strings. Therefore, it is imperative to establish a clear definition of similes as opposed to literal comparisons of two (or more) entities.

Let us first explore Čermák's (2007: 187-188) discussion of three significant aspects used for idiom recognition. These may be summarised and applied to similes in the following points:

1. recurrence,

---

[2] Novel similes with complex syntactic structures may be difficult to find as there are neither established lexical items, nor any anticipated syntactic structure.

2. the invariability of the components,

3. figurativeness.

'Recurrence' requires reliable intuition for an impromptu assessment. Therefore, most idioms are readily recognised only by native or highly experienced speakers of the language. Nevertheless, our intuition is by no means infallible. Additionally, there are more reliable methods of examining the recurrence of an expression, namely searching an idiom dictionary[3] or a sufficiently large corpus. It is worth noting that when the speaker fails to recognise the simile, the truth and felicity conditions play the central role in interpreting the string. Provided that the meaning of the expression is judged plausible within the context, it is likely to receive the literal comparison interpretation, such as (AS) TOUGH AS NAILS. Nevertheless, if the literal meaning is infelicitous, the expression is classified as a semantically obscure idiom, such as GRIN LIKE A CHESHIRE CAT.[4]

'The invariability of the components' is directly related to idiomaticity. Established similes permit very little paradigmatic selection in the ground (if available), and the source slots without a change of meaning – this makes similes idiomatic. Conversely, there are no paradigmatic constraints in literal comparisons, only the standard morphosyntactic restrictions that are universal within each language.

'Figurativeness' is sometimes problematic as many original metaphors (or metaphorical meanings) have become established senses of words and, therefore, are no longer recognised as idiosyncratic. This effectively compromises the reliability of the speaker's intuition. As a result, the figurativeness of meaning is optional and need not be present in similes, such as LOOK LIKE A DROWNED RAT or (AS) BLACK AS JET (cf. Niculae and Danescu-Niculescu-Mizil's (2014) distinction between literal and figurative comparisons).

Čermák (2007: 386) describes the difference between 'non-established comparisons', whose purpose is to simply express a perceived similarity of two entities at a specific moment, and 'established comparisons' (similes), which are phraseological units stored in our mental lexicon as single items. It follows from the distinction that non-established comparisons are readily generated by speakers of the language with a free selection of the components, while similes are acquired as whole phraseological units.

---

[3] Unlike their printed predecessors, online dictionaries are no longer limited by the number of pages. This results in the inclusion of many 'less common' expressions that would otherwise have to be omitted in printed dictionaries in order to save space.
[4] This simile is an example of allusion to Lewis Carroll's *Alice's Adventures in the Wonderland*.

Although some similes are literal in meaning, their respective component slots remain limited in terms of lexical variability. Similes are ready-made phraseological units and, as Čermák (2007: 393) points out, attempts at decomposition and analysis of individual components and their role seldom provide any beneficial outcome. Conversely, literal comparisons are only restricted by the syntactic frame and allow for the free paradigmatic selection of the lexical elements and the analysis of their constituents.

Additionally, there is one fundamental cognitive difference between literal comparisons and similes. It was mentioned earlier that the source of a simile is a prototypical bearer of the feature or quality ascribed to the target. It follows from this premise that we store in our mental lexicon a large set of prototypes[5] serving as felicitous sources in various similes, and since there are many conventional similes, this prototype set must be shared among speakers. Thus, similes are categorisations based on analogy. Literal comparisons do not rely on prototypes since there is no target-source structure, and we can consider both compared entities simply subjects of the comparison.[6]

Lastly, the most significant difference between literal comparisons and similes lies in their functions. The primary function of literal comparisons is to compare two entities. Conversely, similes are categorisations whose primary function is determined by comparing two dissimilar entities.

> What links the simile elements is some variously specified set of common features; what separates them is all the rest. Thus *dis*similarity is an intrinsic part of simile.

> Moon (2011: 134)

2.3.1 Dissimilarity and reversibility

The concept of 'dissimilarity' is often treated as a defining attribute of similes. It is frequently discussed in works on similes:

> We adhere (…) to the criterion, prevalent in many definitions of simile, that it consists of a comparison between two *unlike* things. A simile must entail marked semantic distance between the source and target terms or, put differently, target and source need to belong to different taxonomies.

> Tartakovsky & Shen (2019: 209)

---

[5] The number of prototypes reflects the number of categories that we have designed/adopted for our understanding of the world.
[6] The structure could be viewed as target-target.

The first point is problematic, as it requires a clear operational definition of (un)likeness. The second point is more specific and also revealing. If we stipulate that simile requires its target and source to belong to different taxonomies, we now have a clear criterion. However, this criterion is not exclusive to established similes. Consider these examples:

(1) *This film is almost as good as drinking.*
(2) *He smells like blue cheese.*

Both sentences illustrate that there can be a comparison of two entities from different taxonomies, which would not be classified as an established simile. In (1), we compare *film* to *drinking*, with the former being [a form of audio-visual entertainment] and the latter referring to [(a session of) alcohol consumption]. It is nearly impossible to see how these entities could become co-taxonyms. If we wanted to have an "umbrella term" for these two concepts, it might be 'free-time activities', with *film* interpreted as 'film watching'. However, such a line of argumentation appears unnecessary. The difference in taxonomies is even more evident in (2), where a person is compared to blue cheese on the ground of 'unpleasant smell'. Ultimately, unlikeness, or more commonly dissimilarity, of the target and the source ought to be treated as a requirement in similes, but this criterion alone cannot be used to distinguish between literal comparisons and established similes. However, we can postulate that (1) and (2) are examples of nonce similes, a transitional category between literal comparisons and established similes.[7]

The purpose of a simile is to highlight a particular feature, often by using a literally infelicitous comparison.

(3) *Her words were sharp as a razor.*
(4) *Tim sleeps like a log.*

Both (3) and (4) are established similes whose target and source are from different taxonomies. Moon (2011: 149) describes the phenomenon of comparing two dissimilar things as a 'category mismatch' (i.e. abstract *(words)* ≠ concrete *(razor)*, or human *(Tim)* ≠ inanimate object *(log)*). The category mismatch allows us to use a simple 'reversibility test' to determine whether a comparison is literal or a simile.

(5) *A laptop is as good as a desktop.*

---

[7] Nonce similes are discussed in Section 2.5.3.

(6) *Pam looks like my sister.*

If we compare literally, the reciprocity should allow reversibility of the two entities. Thus, simple comparisons are generally considered reversible.[8] (5) can be reversed to *A desktop is as good as a laptop*, and (6) can become *My sister looks like Pam*. The reversibility of the target and source is not possible in (3) and (4) because the target cannot function as a good bearer of the salient feature (i.e. the ground).

To conclude, the reversibility test is a simple and reliable tool for determining literal, ad hoc comparisons. On the other hand, similes are irreversible for two reasons. Firstly, the target-source relationship is not reciprocal. The source illustrates the (sometimes implicit) ground of the comparison, a function the target cannot supply. Secondly, the source in similes functions as an intensifier of the ascribed quality or feature.[9] This function does not occur in literal comparisons.

2.4 Figurativeness in similes

Conventionally, similes are interpreted as figurative expressions. However, the figurativeness of similes is not merely a question of acceptance or rejection and requires a thorough examination. All the more so, as Geeraerts (2010: 283) points out that "lexical semantics has not yet come up with an adequate, operational definition of figurativeness." Moreover, the difficulty does not pertain only to the figurative meaning since the precise definition of literalness may be problematic, too (cf. Cruse, 2011; Fishelov 1993). I will avoid attempting a detailed analysis of literalness here and focus on the figurativeness in similes using the conventional dichotomy.[10]

Initially, we should revise the premise that all similes are figurative by considering the following:

> What makes a simile figurative is that it prompts one to search for similarities where one would not expect to find them, and to make connections across concepts which seem otherwise unconnected.

> (Israel et al., 2004: p.126)

---

[8] According to Functional Sentence Perspective (FSP), one of the entities must always be rhematic, which is contextually dependent. However, that does not invalidate the premise that the items are reversible without making the comparison infelicitous or altering the truth conditions.

[9] The intensification function is explored in Section 2.6.1.

[10] The dichotomy of literal and figurative meaning, as it is widely used in current linguistics.

In the traditional sense, similes are considered figures of speech, but that does not necessarily render them semantically figurative. Similes such as (AS) HARD AS ROCK often remain within a single domain (material toughness), such as *The bread is hard as rock.* The feature expressed by the ground applies to both the target and the source literally. However, the simile (AS) HARD AS ROCK remains a conventionalised expression (with some variability) used by speakers to describe literal or metaphorical toughness while comparing two different things. Thus, the simile is considered a figure of speech.

Consequently, in order to understand why similes are figures of speech but need not be figurative in their meaning, we need to distinguish between two different phenomena: 'figurative mode of reference' and 'figurative meaning'.

2.4.1 Figurative mode of reference

When Norrick (1986) mentions analogical description, he describes the figurative[11] mode of reference, in other words, using a substitute expression for an otherwise digital description. This makes similes figurative because they are conventional expressions (selectional preferences) associated with specific intended analogies. For example, when we want to use a simile to express that someone is (being) foolish, we could use the simile ACT/LOOK LIKE A CLOWN. This is not to say that no other options are available, as the language, by design, allows for multiple ways of expressing the same concept. Nevertheless, speakers generally tend to retrieve an established simile rather than design their own.[12]

Let us return to the argument that similes need not be semantically figurative, i.e. they need not employ figurative similarity. Many examples of similes may utilise literal similarity while still belonging to the figure of speech category, such as CRY LIKE A BABY or (AS) FLAT AS A PANCAKE. This essential aspect of similes illustrates the fine distinction between a figure of speech and figurative meaning.

---

[11] "The rhetorical use of *figure*, "peculiar use of words giving meaning different from usual," dates to late 14c.; hence **figure of speech** (1550s)." Harper, D. (n.d.). Etymology of figure. Online Etymology Dictionary. Retrieved August 12, 2023, from https://www.etymonline.com/word/figure

[12] This is a pragmatic motivation as conventional similes are unlikely to be misunderstood or regarded as infelicitous. It may well be interpreted as "playing by the rules" of a language. However, it is not uncommon for speakers to design nonce similes, either for the humorous effect or simply because they were unable to retrieve an existing simile.

2.4.2 Figurative meaning

Some similes, such as (AS) WHITE AS SNOW, can be somewhat literal, and others more figurative, such as (AS) BRIGHT AS A BUTTON. No fine line exists between literal and figurative, but standard similes tend to become perceived as figurative over time (Čermák, 2007: 386). As no sufficient, generally accepted definition of figurativeness exists, the operational definition of figurativeness adopted in this work is determined by the target-source relation. The definition is simple: 'target ≠ source'. This definition considers all similes figurative, regardless of the level of idiomaticity, allowing us to ignore the theoretical distinction between the figurative mode of reference and figurative meaning. Some authors offer elaborate semantic classifications and meticulously describe the differences between literal and figurative meanings in similes (e.g. Norrick, 1986) and domain (in)congruence (e.g. Ortony, 1979). However, such theoretical considerations are beyond the scope of this work.

2.5 General classification of simile types

This section deals with the general classification of simile types. It is not an in-depth analysis of all possible (sub)types as the main focus of this thesis is adjectival similes, but rather a presentation of the underlying principles shared by similes regardless of their structural variation.

2.5.1 Verbal vs adjectival similes

One of the most basic ways of classifying language elements is according to their part of speech. As similes are phraseological units with meaning, the essential constituents must be lexical parts of speech: nouns, adjectives, verbs, and adverbs. However, the distribution is not equal. The left-hand side of similes generally tends to contain verbs or adjectives, while the right-hand side strongly prefers nouns (Čermák, 2007: 387). This is hardly surprising, as the original purpose of similes is to compare one entity to another. Nouns are linguistic signs representing concrete (physical) and abstract entities, and thus, they occupy the right-hand side. The verb/adjective in the left-hand side slot represents the (behavioural) feature or quality shared by the target and the source.[13] Using the part-

---

[13] This does not apply to implicit similes, whose verbal slot is frequently occupied by common verbs followed by a comparator, such as *be like* or *look like*.

of-speech criterion, we postulate two main types of similes according to the leftmost lexical item, as it is generally considered the head element of the simile:[14]

(7) *Harry drinks like a fish.*

(8) *Iliza is bright as a button.*

Example (7) contains the relator *drink* and the source *fish*. Since *drink* is a verb, the string is classified as a 'verbal simile'. In (8), *bright* functions as the ground, which makes (8) an 'adjectival simile'. However, the part of speech is not the only difference between *drink* and *bright* here, as they are also different semantic components within their respective similes. This brings us to the topic of semantic components of similes.

2.5.2 Explicit vs implicit similes

The semantic classification of similes can be approached mainly from two different perspectives. Firstly, we can analyse the relations between the individual components and design a classification that primarily works with the literal-figurative dichotomy. Secondly, we can look at the explicitness of the ground.

The default simile structure comprises five semantic components, each with its own specifics and function. While various terms are used for the individual components, perhaps the most accurate nomenclature can be found in Figure 2.

comparandum (target)	relator	tertium comparationis (ground)	comparator	comparatum (source)
Robin	*was*	*high*	*as*	*a kite*
Lenny	*cried*	*(-)*	*like*	*a baby*

Figure 2. The semantic components of similes.

Naturally, not all similes contain all five components explicitly. Nevertheless, two prototypical forms illustrated in Figure 2 serve as the base forms for all structural variants. These simile forms present a target (*Robin* and *Lenny*) and a source (*kite* and *baby*), the formally mandatory elements of the simile. However, *Robin was high as a kite* also

---

[14] For a detailed structural classification with possible permutations in Czech and beyond see Čermák (2007: 404-407).

contains an explicit ground *high*, making the simile 'explicit' (Čermák, 2007: 401). The explicitness means there is no free ground interpretation, as the simile states the ground. This is not the case in *Lenny cried like a baby*; therefore, such similes are 'implicit' (ibid.). In implicit similes, the ground is either implied by the relator or is not required for the interpretation.

Consequently, the question of the ambivalence of implicit similes is in order. According to Tartakovsky & Shen (2019: 204), "a prevalent assumption with regard to similes that was developed for open similes is that the ground is a salient feature of the source."[15] Therefore, it is expected to be easily retrievable. Furthermore, Čermák (2007: 396) points out that the relator may inherit the implied ground's function, thus rendering the explicit ground redundant or even ineffective. Consider the following examples:

(9) *Madison runs like the wind.*

(10) *Their lack of communication made me feel like a mushroom.*

Regardless of the physical technicalities, it is easy to recover the implicit ground *fast* in (9). Explicitly providing the ground would be redundant, as the relator *run* inherits and implies the notion of velocity.

In the case of (10), we need to know something about mushrooms to recover the intended ground or be familiar with the expression. Here, the essential concept is that mushrooms thrive in the dark. The simile is closely associated with the idiom *to be left in the dark*, whose meaning is [to be uninformed]. Expressing the ground *to be uninformed* explicitly would result in an infelicitous construction:

(11) *Their lack of communication made me feel uninformed like a mushroom.*

The obvious flaw in (11) is that we are comparing an animate target (*me*) with an inanimate source (*mushroom*) by committing to a ground that is only compatible with animate entities. The explicitness of the ground inhibits the intended interpretation, i.e. the meaning *left in the dark*. The result is the incompatibility of *uninformed* and *mushroom*.

Lastly, the understanding of implicit similes can vary. Qadir et al. (2016) argue that multiple properties (grounds) can be inferred in implicit similes.

---

[15] The term 'open similes', originally used by Beardsley (1958: 137), describes implicit similes, i.e. similes without an explicit ground.

(12) *John sounds like a politician*.

In (12), the source *politician* is associated with several possible grounds, none of which is primed exclusively outside a specific communicative situation. Therefore, the implicit ground can be *deceitful, charismatic, diplomatic,* and so on. This is not to argue that all implicit similes are necessarily underdetermined. Qadir et al. (ibid.) divide implicit similes into three groups according to 'interpretive diversity' (low, medium, high) based on the number of properties inferred by their human interpreters.

To summarise, the explicit-implicit typology of similes is based on the explicitness of the ground. This should be viewed as an underlying classification directly corresponding to the verb/adjective distinction. Furthermore, different transformations might occur in other languages according to the morphosyntactic standards of each language.

2.5.3 Nonce vs standard similes

Similes are sometimes judged by their aesthetic effect. If the construction feels sufficiently *poetic*, it will be marked as a simile, which is directly related to figurativeness. Gargani (2016) and Oleniak (2018) use the term 'poetic comparison' as a synonym for simile and 'non-poetic comparison' for literal comparison. This terminology suggests that all similes are poetic, which does not hold for many examples of standard similes, such as (AS) SOFT AS A BABY'S BOTTOM or (AS) BLACK AS COAL. Fishelov (1993: 2-3) distinguishes between 'poetic' and 'non-poetic similes', with the former being frequently, but not exclusively, found in poetry. As poetic similes are predominantly one-off expressions that aim to trigger specific aesthetic effects (beyond all the other effects commonly conveyed by similes), they can be described as nonce (or author) similes. Consider the following:

(13) *She was staring at him like a forgotten torch trying to pierce through the night, desperately searching for its long-gone owner.*

Example (13) illustrates that there are similes designed for a single occasion whose meaning goes beyond a mere literal comparison. The reproduction of (13) would eventually dampen the original aesthetic effect and render the simile ineffective. Naturally, this contradicts the recurrence principle, one of the main criteria used in idiom (and simile) recognition.

To conclude, nonce similes are primarily the subject of literary analysis and are of little interest to linguistic enquiry for two reasons. Firstly, they are often purposeless (or

even ineffective) for standard communication, as they are created for a particular occasion. Secondly, as the source slot is frequently extensive, they are difficult to track down, making their collection arduous.[16] Therefore, this work does not focus on nonce similes beyond the general remarks. Conversely, standard similes are conventionalised, recurrent expressions that are not limited to specific situations. They are acquired by speakers of the language and relisted in dictionaries for the purpose of description and preservation (hence standard). The recurrence makes them possible to be traced in corpora and justifies the attention, as they can be used in communication to convey established and readily recognised meanings.

2.6 Simile functions and effects

As figures of speech, similes are associated with several functions, some of which are more prominent and frequent than others. Superficially, the purpose of a simile is to compare two entities. However, the target and source comparison is sometimes so bizarre that it has a humorous effect; for instance, *He was sprinting like a legless ostrich*. Furthermore, the usage of adjectival similes suggests that the process of likening one thing to another is diminished in standard similes, resulting in the source functioning as an intensifier of the quality or feature ascribed to the target. Some other theoretical phenomena, such as 'the C/C debate' (Barnden, 2016) or 'affective polarity' (Qadir et al., 2015), could be discussed concerning similes and their understanding and function. Nevertheless, this subchapter explores just the two primary functions of similes: intensification and humour.

2.6.1 The intensification function

Intensification is a defining feature of adjectival similes.[17] Regardless of whether speakers rely on comparing the target and ground when interpreting a particular simile, the intensification function is inherently present.

(14) *In this region, ravens are as unique as a fingerprint.*
(15) *Stacy is mad as a hornet.*

---

[16] This is especially true of nonce similes that occur in literary text, where the aesthetic function is often tied to creativity and originality.
[17] Generally, intensification is not found in verbal similes.

Both (14) and (15) are standard similes that use established ground-source lexical elements. A *fingerprint* is *unique* to every individual, and *hornets* are often perceived as aggressive and frantic (*mad*). Consequently, both *unique* and *mad* can be viewed as prototypical features of their respective sources. However, the similes can also be paraphrased as *very/completely unique* and *very/completely mad*. Moon (2008: 7) describes adjectival similes as "pragmatically emphasizing the degree of the quality denoted by the adjective." The intensification function is undisputed, but the question remains as to how much of the comparison between the target and the source remains active in the interpretation of these similes.

2.6.2 The humorous effect

Veale (2013: 4) notes that "the simile form continues to be a frequent, varied and ingeniously crafted vehicle for expressing humorous intent." This does not pertain to nonce similes only. Norrick (1986: 48-51) discusses humour in similes and elaborates on the phenomena contributing to particular similes being perceived as humorous. The following examples illustrate the four principal causes:

(16) *Her face was smooth as a baby's bottom.*
(17) *He was sprinting like a legless ostrich.*
(18) *Sometimes I like to get high as the sky.*
(19) *When I was a teenager, I was flat as a pancake.*

(16) is an example of an incongruent source *baby's bottom.* The idea of comparing someone's *face* to a *baby's bottom* is often perceived as humorous. This is caused by our notion of *bottom*, which is associated with a set of features different from *face*. Example (17) illustrates an ironic simile. The simile initially contains the relator *sprint* and then uses a bizarre source *legless ostrich* to contradict the relator. The humorous effect is, once again, triggered by the incongruent source. Additionally, the literal and intended meaning imbalance reinforces the humour. (18) represents punning, where the literal meaning remains feasible but cannot be considered the intended meaning (i.e. *intoxicated*).[18] Lastly, both (18) and (19) are examples of a euphemism for a taboo concept, with the latter also being rude in this particular case. *Flat* refers to breast size, and this type of

---

[18] Norrick (1986: 50) labels standard similes conventionally used for punning "*typically* punning stock similes".

language may be humorous in some situations, primarily among male speakers when describing women.

To conclude, humour is a complex and subjective phenomenon. There is no clear consensus on which similes are humorous, and the context plays a pivotal role in evaluating humour in individual instances.

3 Adjectival similes

This section deals with the formal and structural properties of adjectival similes. The individual structural components are explored, followed by a description of various classifications of what is considered an adjectival simile type. Another subsection is dedicated to a brief presentation of prosodic features relevant to adjectival similes. This section concludes with notes on the familiarity of some relatively infrequent similes.

3.1 The tagging of adjectival similes in corpora

Using corpus data is imperative for this work, but particular issues related to tagging need to be addressed first. According to Stefanowitsch (2020: 121), "it is necessary to understand that the categorisation of corpus data is an interpretative process in the first place. This is true regardless of the kind of category." The primary purpose of corpora is to allow large volumes of language data to be available and, ideally, easily retrievable for research. Therefore, there is little benefit in flooding the annotation with various possible interpretations, and usually, only one is selected. It is then up to every user to consider the presented annotation and decide whether to accept or refuse it. Some have even advocated the so-called *clean-text policy* (Sinclair, 1991: 21), but that significantly hinders corpus-based research. Unsurprisingly, Lindquist (2009: 45) argues that "the majority of corpus linguists generally prefer tagged corpora when they are available." The following point regarding tagging provides a fair argument as to why annotated corpora should not be viewed as something *tainted*:

> (…) corpus annotation is the manifestation within the sphere of corpus linguistics of processes of analysis that are common in most areas of linguistics. To identify problems with accuracy and consistency in corpus annotation is, in principle at least, to identify flaws with analytical procedures across the whole of linguistics. It is because of the issues of accuracy and consistency, in particular, that some linguists prefer to use unannotated corpora. But this does not mean to say that such linguists do not analyse the data they use; rather, it means that they leave no systematic record of either their analysis or their errors which can easily and readily be tied back to the corpus data itself.
>
> McEnery & Hardie (2012: 14)

Especially the lack of retraceable analysis in studies may negatively impact replicability, which is generally considered one of the main assets of corpus linguistics (McEnery & Hardie, 2012: 16; Stefanowitsch 2020: 133-136). To summarise, accepting

a specific interpretation in a corpus and revising it seems much more convenient than having none to help us in our corpus-based mining. Furthermore, relying solely on our interpretation in dealing with large datasets would inevitably result in more inconsistency and errors.

Bearing in mind that the success rate of automatic tagging is roughly between 95% to 97% (Lindquist, 2009: 47, Stefanowitsch, 2020: 89), we have to be prepared to encounter words that are wrongly or ambiguously tagged. Arguably, potential tagging issues are inevitable but may sometimes be used for our benefit, as even incorrect tagging allows us to design a query that will return the desired data. The only limitation is that we need to explore the incorrect tag manually.

Despite being very general, these considerations are highly relevant in corpus investigations related to adjectival similes. Since the adjectival simile is a theoretically established and structurally relatively fixed phraseological unit, the data investigation is corpus-based. Consequently, this allows us to search for specific similes and explore the tagging policy. Such an investigation reveals that the part-of-speech tagging of individual components shows some variation in different slots of the frame. Naturally, this must be accounted for when constructing queries, and while it is impossible to design an ideal query that would return similes only, there are still elements worth eliminating from the final adjectival simile search to reduce the number of items for manual reviewing. These issues are explored in Section 4.2 in more detail.

3.2 The structure and diagnosis of adjectival similes

Before subjecting adjectival similes to scrutiny, we need to establish a reliable method of their diagnosis, which is necessary for the query design and the subsequent corpus extraction. The formalised frame '*(as)* ADJECTIVE *as* NOUN PHRASE' is considered the prototypical form, as it is represented in most standard adjectival similes. Each component of adjectival similes is explored individually in the following lines, and comments are made regarding its optionality and corpus tagging. This is necessary for the standardised queries used for simile extraction in various corpora (in Section 4).

3.2.1 The initial *as*

The first component of an adjectival simile is *as*, which functions as a degree adverb. It is sometimes considered a correlate of the comparator *as*, for example, *as green as grass* or *as many as we wanted*, but that is somewhat problematic, as they differ in part

of speech.[19] The corpora usually use the general adverb tag for the initial *as*.[20] Other comparative constructions allow *so* as a variant of *as* in the initial position, especially negative assertions, for example, *the instructions are not so clear as we believed*, *he is not so good as a lover*, or idiomatic constructions, for example, *as/so long as* or *as/so far as*. However, initial *so* is not attested in similes. Therefore, its presence in a string can be used as a definitive argument against the simile interpretation.

Perhaps the most crucial fact for corpus research is that the initial *as* is not obligatory in standard similes (e.g. Huddleston and Pullum, 2002: 1130). Consequently, it cannot be used as a determining element in simile diagnosis, but it would still appear relatively frequently. In her study of similes in the Bank of English (BoE), Moon (2008: 5) reports that "more than half [of the adjectival similes] occur more commonly without [initial *as*]."

Lastly, the initial *as* is somewhat detrimental to the alphabetical organisation of entries in dictionaries or word lists. Therefore, idiom dictionaries usually list adjectival similes without the initial *as* alongside other idiomatic expressions (e.g. Makkai et al., *Handbook of Commonly Used American Idioms*, 2013), while other dictionaries choose to put them in a separate simile section (e.g. Seidl and McMordie, *English Idioms and How to Use Them*, 1988). Finally, online dictionaries completely eliminate alphabetical organisation, at least in the user interface.[21]

3.2.2 The adjective

The adjective is considered the head element of adjectival similes, both semantically and syntactically. Therefore, its presence is the primary determining factor in diagnosing similes. In most adjectival similes, the adjective is primarily a morphologically simple monosyllabic or disyllabic word; however, compound adjectival simile (CAS for short) transformations are also fairly common. Furthermore, very frequent transformations of similes (AS) LARGE AS LIFE or (AS) GOOD AS SEX, namely *larger than life* and *better than sex*, contain a comparative adjective.

One potential issue is caused by the fuzzy boundary between adjectives and adverbs, especially in American English, which may result in the adjective being tagged as an adverb. Furthermore, many adjectival similes function as adverbials of manner, further

---

[19] The *as…as* combination is often described as a correlative construction (e.g. Quirk, 1985), which is likely caused by its frequency and many idiomatic expressions utilising the string.

[20] Occasionally, initial *as* is tagged as a conjunction, but such instances are rare and best considered as marginal tagging errors.

[21] The internal database is likely to contain alphabetical ordering of entries.

complicating the part-of-speech issue. This needs to be considered when designing the query, as the recall of similes is likely to be worse without including both the adjective and the adverb tag. However, larger corpora may return too many irrelevant items with the adverb tag inclusion. Consequently, the decision to include the adverb tag is tied to the simile recall of the query without it.

3.2.3 Comparator *as*

An equally important element in simile diagnosis is the comparator and its form. Invariably, we can observe two comparator forms: *like* in open similes and *as* in closed similes. Both forms are best interpreted as prepositions in similes but are sometimes identified as subordinate conjunctions. This can cause problems, especially in the case of *as*, which may sometimes be tagged in the corpus as a conjunction (see Section 6.2.3). The conjunction classification is directly related to the initial *as*, as the familiar *as…as* string invites the correlative conjunction interpretation. Regardless of the inaccuracy of such an interpretation, it affects the tagging of the comparator *as*.

The comparator *as* is obligatory, making it a criterial component in adjectival similes. However, including the lemma *as* in the query eliminates all the possible transformations, such as the aforementioned *larger than life* and *better than sex*.

Lastly, the comparator *as* can be replaced by *like* in some adjectival similes, as in *hard like a rock*. Although non-standard, examples of this substitution repeatedly occur in larger corpora and may indicate a developing trend.

3.2.4 The noun phrase

The final element of the adjectival simile is the noun phrase. The number of components within the noun phrase varies from a single noun, for example, (AS) THICK AS [THIEVES], to an expanded noun phrase, for example, (AS) ALIKE AS [TWO PEAS IN A POD]. Most noun phrases within standard similes have the '(determiner) + noun' structure, but the inclusion of less frequent examples remains desirable. The part-of-speech pattern of the noun phrase comprises up to three components (Table 1).[22]

---

[22] The noun phrase can still be post-modified by additional elements such as relative clauses, for example, *(as) happy as a manager who's on a business trip with their assistant*, or preposition phrases, for example, *(as) alike as two peas in a pod*. The data suggest that this usually occurs in literary similes and the post-modifier can still be tracked manually.

Adjective + comparator	Article or cardinal number	Adjective	Noun
clear as	-	-	*day*
sharp as	*a*	-	*razor*
tough as	-	*old*	*boots*
pure as	*the*	*driven*	*snow*
thick as	*two*	*short*	*planks*

Table 1. The components of a noun phrase in adjectival similes.

Naturally, the grammatical category of the number affects the determiner, particularly the article. In terms of countability of the head noun, all the inflectional forms are possible, for example, countable singular ((AS) SHARP AS A RAZOR), countable plural ((AS) CHEAP AS CHIPS) or uncountable singular ((AS) CLEAR AS CRYSTAL). Countable nouns in similes primarily occur with the indefinite article, but examples with the definite article are also possible with both singular and plural, for example, (AS) BLUE AS THE SKY, (AS) BRIGHT AS THE SUN, (AS) OLD AS THE HILLS. The definite article results from a contextually unique phenomenon serving as the simile source.[23] Generally, similes do not utilise lexical units from one-item sets, so the frequency of the definite article in similes remains fairly low. Overall, it is necessary to keep the noun tag inclusive [pos="N..*"] to recall both uncountable and countable nouns, along with their inflected forms.

3.3 Classifying simile types

Before proceeding to simile lists, we need to establish what is considered a unique simile type. According to Moon (2008: 12), there are three main approaches to classifying an independent simile.

The 'conceptual approach' allows us to cluster many synonymous similes into a single conceptual simile, such as (AS) HARD/TOUGH AS ROCK/NAILS/OLD BOOTS. This is theoretically appealing as it significantly reduces the number of simile items while compiling simile lists. However, such an approach makes it difficult to account for the frequency of the individual similes and presents them all as equally viable. Furthermore, many of the synonymous similes are regional preferences or simply differ in meaning (both conceptual and associative). Therefore, deciding when two similes should be treated as a single conceptual simile is challenging.

---

[23] *Sun* and *sky* are unique in the context of the Earth and the overall perception of our existence. However, such context may change in the area of space exploration or (space) science fiction storytelling.

The 'dictionary approach' simplifies the clustering. Either the noun or adjective become the fixed element with the other slot listing possible transformations, for instance, (AS) WHITE AS A SHEET/A GHOST/DEATH/MARBLE/MILK/PAPER/SNOW or (AS) PALE/WHITE AS A GHOST. This approach is functionally motivated as it requires a head word used in searches and alphabetical lists. It also allows the dictionary to cluster the simile with the head word itself and other expressions containing the same lexical item.

The 'type approach' (or corpus approach) disregards clustering completely. Strings such as *(as) sharp as a razor* and *(as) sharp as a tack* are treated as two separate entities. This approach is convenient, for no decision must be made regarding usage differences between transformations. The drawback is an extensive list of all possible transformations or synonymous sets that are sometimes functionally identical. This can be mitigated by setting a frequency cut-off or attaching a commentary explaining the observed frequency of each individual simile.

All three approaches are used in this work but for different purposes. The conceptual and the dictionary approaches are used in the description of the most frequent similes (Section 7), while the type approach is used in quantitative analyses in Sections 6 and 9.

3.4 Prosodic features

Observations can also be made regarding the prosodic features of adjectival similes. As Moon (2008: 5) points out, "many similes are alliterative, assonantal, or both." This makes them easier for speakers to remember and reproduce. According to Moon's (2008: 35) BoE sample, 23% of simile types exhibit at least one prosodic feature. Arguably, phonetic motivation might have been the primary factor in some similes with striking prosodic features.

3.4.1 Alliteration

Often considered a literary device, alliteration frequently occurs outside literature, for example, in proper names (see Coard, 1959; Bush, 2020) or multi-word units, such as binomials (*done and dusted, sticks and stones* or *trick or treat*) and similes ((AS) BUSY AS A BEE or (AS) RIGHT AS RAIN). Broadly, alliteration can be described as a repetition of a letter or sound. Letter alliteration often corresponds to sound alliteration, but examples such as (AS) SURE AS THE SUN show that it is not always reliable. The operational definition of alliteration adopted in this work is "the repetition of word-initial sounds" (Roper, 2011: 1). This includes both word-initial consonants and vowels.

3.4.2 Assonance

Assonance is the repetition of a vowel, for instance, the repetition of /æ/ in (AS) FLAT AS A PANCAKE. Some similes are assonantal, which makes them more accessible and memorable. The repetition of vowels usually does not occur in the word-initial position, as the frequency of adjectives with an initial vowel is scarce. Assonance can also be accompanied by alliteration, for example, (AS) FIT AS A FIDDLE. Compared to alliteration, assonance is intuitively a less prominent prosodic feature, especially in similes with multisyllabic constituting elements, such as (AS) MISERABLE AS SIN.

3.4.3 Rhyme

Rhyme is a phenomenon occurring when the nucleus and coda (together termed 'rhyme') are identical in two or more words. Examples of similes utilising rhyme are (AS) DRUNK AS A SKUNK or (AS) THICK AS A BRICK. These similes have an exceptionally memorable form, but they tend to sacrifice the semantic aptness, as the source's primary purpose is to rhyme with the ground rather than provide us with an 'ideal' bearer of the described feature. Rhyme can arguably be considered the main factor motivating rhyming similes.

3.4.4 Other formal properties

Other formal properties found in similes relate to the morphological structure of the individual elements. Both the ground and the source tend to be single-word phrases. The lexical elements in similes are mostly of Germanic origin, which is reflected in the number of syllables. Moon (2008: 5) observes that "[m]ost adjectives in *as*-similes are monosyllabic or if disyllabic end in -y." The ground also tends to be morphologically simple. This further reinforces the idea that similes are not used as precise semantic descriptions. Their functions are rather to emphasise, entertain or compare the target entity to something generally familiar. A detailed exploration of the morphology of similes in the dataset is presented in section 9.1.

3.5 Notes on simile familiarity

Despite being sparsely used, many adjectival similes may look and sound familiar to speakers. This is primarily caused by transformations, especially the CAS, which are significantly more frequent than the simile form or lack a default simile variant altogether.

It is therefore at least possible that some similes believed to be institutionalized are either over-reported or linger on in the lexicon because of interference from other structures with the same lexical collocates.

<div align="right">(Moon 2008: 33)</div>

This can be illustrated by the colour similes and their common CAS transformations. The CAS forms JET(-)BLACK, SEA(-)BLUE and RUBY(-)RED are fairly frequent expressions whose existence is rooted in corresponding similes. However, the simile forms (AS) BLACK AS JET, (AS) BLUE AS THE SEA, and (AS) RED AS A RUBY, respectively, are either scarcely used or do not occur in the language anymore. Other examples of CAS transformations rarely seen in their original simile form are, for instance, WAFER(-)THIN, ROCK(-)SOLID or LIME(-)GREEN.

III. RESEARCH PROJECT

4 Data mining and hypotheses

This section presents the corpora chosen for this research and the process of designing the queries, starting from a general query and considering the specifics of each corpus concerning its user interface, annotation style and search options. Following the description of the data mining process, the hypotheses are formulated.

4.1 Researched corpora

Four corpora of varying sizes and data samples were used in this study: the spoken section of the British National Corpus 2014 (Spoken BNC2014; 11,422,617 tokens), the British National Corpus (BNC; 112,102,325), Corpus of Contemporary American English (COCA; 1,001,610,938 tokens), and the English Web 2015 (EW15; 13,190,556,334 tokens). The Spoken BNC2014 exclusively contains spoken data, the BNC and COCA combine written and spoken data (with spoken data amounting to approximately one-tenth of the overall sample size), and the EW15 comprises only written data. For this reason, any comparisons must be carefully considered, as data diversity can and will distort the results. This work primarily does not compare particular periods, dialects, genres, and age or gender demographics; it focuses on the frequency and tagging of adjectival similes in corpora. Consequently, dialect or demographic comparisons are made only sporadically to illustrate specific issues.

One of the reasons for choosing heterogeneous corpora was the diversity of data. The majority of the data come from written texts, as corpora of spoken language are less frequent because of the difficulties associated with gathering data (Sinclair, 1991; Lindquist, 2009; McEnery & Hardie, 2012; Stefanowitsch, 2020; and many others). Nevertheless, the inclusion of spoken language data is essential for diversity. Varying dialects and modes of communication are likely to contain different examples of adjectival similes, and the purpose of this work is to collect as many simile types as possible. Although listed for illustration, comparing frequencies of similes in the respective corpora is not the primary goal of this research.

4.2 The general query design

First, the preferred query mode is CQL (Contextual Query Language). Other simplified modes do not provide enough flexibility to research phraseological units with multiple variables. Since a simile is an established phraseological construction with fixed

essential part-of-speech elements, it lends itself nicely to a corpus-based investigation. However, other word combinations utilise an identical formal frame, and thus, we must find ways to eliminate the irrelevant word strings from the search results. Working through a cluttered list of items manually with corpora such as the Spoken BNC2014 may be manageable, but it would not be feasible with larger corpora, such as the COCA or the EW15. Resorting to simple queries with specific lemmas only returns established items. This is neither time-efficient nor result-effective.

4.2.1   Refining the general query

The general query represents the prevalent formal pattern of adjectival similes. It contains the optional initial lemma *as*, the obligatory adjective (JJ), the obligatory comparator lemma *as*, the optional article (AT) or the cardinal number (MC) determiner, the optional adjective modifier (JJ), and the obligatory noun (N).

(Q1)   [lemma="as"]? [pos="JJ.*"] [lemma="as"] [pos="AT.*|MC"]? [pos="JJ.*"]?
        [pos="N..*"]

This serves as a point of departure for further refinement. Since the research initially began with the Spoken BNC2014, this default query (Q1) was designed using the C6 tagset.[24] The Spoken BNC2014 proved a suitable gateway to simile research due to the easily manageable data sample size. The query returned just 504 matches (44.123 i.p.m.) with 428 different types.

No.	Query result	No. of occurrences
1	as bad as --ANONnameF	7
2	as good as --ANONnameM	5
3	common as muck	5
4	high as a kite	5
5	as far as --ANONplace	4
6	as good as --ANONnameF	4
7	as bad as --ANONnameM	3
8	as good as --ANONplace	3
9	as nice as the ones	3
10	awkward as fuck	3

Figure 3. The top ten results of the default query in the Spoken BNC2014.

---

[24] C6 tagset (UCREL CLAWS6 tagset) is used in the Spoken BNC2014 annotation.

The frequency breakdown in Figure 3 immediately exposes several drawbacks. First, the adjective slot is cluttered with words frequently occurring in the comparative frame, for example, *as far as*, *as good as* or *as bad as*. Similarly, words not found in similes can occupy the noun slot, namely *fuck*, *shit* and *hell*.[25] Furthermore, in larger corpora, the adjective part-of-speech tag can return frequent combinations that are also better to exclude, for example, *such as*, *available as*, or *same as*. It is best to use the NOT operator (!) to exclude specific lemmas in the query to avoid getting a clogged query result. However, the excluded lemmas need to be carefully considered; otherwise, we might be eliminating potential similes from the search.

Another simple and effective step in refining the query is to remove the optional initial *as*, as its absence does not result in missing similes in the list. The benefit is that the result will no longer show items with the initial *as* separately from those without it, meaning that *cheap as chips* and *as cheap as chips* are merged into a single item in the frequency list. On the other hand, we lose track of the initial *as*, but overall, it still seems a worthwhile trade-off.

(Q2)   [pos="JJ.*"&lemma!="bad|far"] [lemma="as"] [pos="AT.*|MC"]?
        [pos="JJ.*"]? [pos="N..*"&lemma!="fuck|shit|hell"]

The semi-refined query (Q2) no longer contains the initial *as* and excludes the frequent lemmas that appear in the frequency list in both the adjective and the noun slot. However, three more adjustments mentioned in the previous sections are necessary for the general query to recall most of the similes. First, the adjective-head tag needs to be expanded to include the adverb tag because of the transitional nature between adjectives and adverbs. Second, the noun tag likewise needs to be expanded to include the adjective tag, as simile sources are sometimes tagged as adjectives. Consequently, the number of excluded lemmas must increase considerably; otherwise, the precision of the simile recall drops significantly. Once again, this is achieved by browsing the frequency lists and pinpointing the most frequent items irrelevant to a simile inquiry.

The exclusion process can be divided into two subsequent steps. The first step is to exclude unwanted adjective and adverb lemmas from the adjective slot, including lemmas with the suffix *-ly*. The *-ly* lemmas were made part of the exclusion after careful consideration. Although the suffix *-ly* may occur in adjectives, the initial corpus searches

---

[25] These are instances of intensifying post-modification (IP), which is a phenomenon different from similes (Emmer, 2020).

did not return any simile with a *-ly* adjective as its head. Although Google search reveals the simile (AS) FRIENDLY AS A PUPPY, it is not attested in the researched corpora. However, some general adjectives such as *long* and *good* must be kept as they are attested in similes, for example, (AS) LONG AS (ONE'S) ARM and (AS) GOOD AS GOLD. Here, the second step helps us eliminate some remaining undesired items, this time by excluding lemmas from the noun slot, for example, *as long as possible* and *as good as the other*. The finalised general query (Q3) serves as the basis for further simile investigation in various corpora.

(Q3) [pos="JJ.*|RR.*"&lemma!=".*ly|almost|bad|far|fine|just|like|much|only|quite|
      soon|twice|well"] [lemma="as"] [pos="AT.*|MC"]? [pos="JJ.*"]?
      [pos="N..*|JJ.*"&lemma!="fuck|hell|other|possible|shit|well"]

One final comment should be made regarding the expanded tags. Larger corpora provide enough examples of similes using the semi-refined query (Q2), so the expanded tag may feel redundant. The recall of (Q2) would not be as high, but its precision would remain unaffected by the sudden flood of irrelevant adverbs and adjectives caused by (Q3). Ultimately, it remains a judgement call for each individual, and the list of excluded lemmas can be expanded or shrunk.

4.3 Standard adjectival similes: queries

The following subsections present the queries designed for adjectival simile extraction from the researched corpora and describe the potential annotation idiosyncrasies and related issues. The absolute frequencies of simile types in the tables are based on the orthographic form; therefore, *sharp as a razor* and *sharp as razors* are listed as two different types despite semantically representing the same conceptual simile (AS) SHARP AS A RAZOR.

4.3.1 The Spoken BNC2014 query

The Spoken BNC2014 was the first corpus used for simile mining. One of the original ideas was to get data from a spoken language corpus to assess whether similes are used in everyday language. Additionally, it is the smallest corpus used in this research, so it serves as an ideal starting point for lemma elimination. The lemmas excluded from the adjective and noun slots in (Q4) were chosen based on two criteria: generally known frames, for example, *as soon/much/well as*, and lexical items from the frequency list of

(Q3) that after a manual revision of the data, cannot be considered part of similes. A notable exclusion is the lemma '--.*', representing proper-name lemmas. The Spoken BNC2014 replaces proper names with ANONnameF (female names), ANONnameM (male names), and ANONplace (place names) for the sake of anonymity. These irrelevant lemmas ranked within the top 15, which was clogging the results. Since all proper names in this corpus are anonymised, distinguishing potential similes, such as (AS) RICH AS CROESUS, from literal comparisons is virtually impossible.[26] This makes not excluding proper names from the search counterproductive.

(Q4) is the finalised query for the Spoken BNC2014 simile extraction.

(Q4) [pos="JJ.*|RR.*"&lemma!=".*ly|almost|bad|far|fine|just|like|much|only|quite|
soon|twice|well"] [lemma="as"] [pos="AT.*|MC"]? [pos="JJ.*"]?
[pos="N..*|JJ.*"&lemma!="--.*|fuck|hell|other|possible|shit|well"]

4.3.2 The BNC query

The BNC uses a C5 tagset (UCREL CLAWS5 tagset) for annotation, which is the variant preceding the C6 tagset used by the Spoken BNC2014. Consequently, the query must reflect the different tags and positional attributes (p-attributes). Table 2 summarises the key differences between the two corpora for the simile query.

	BNC	Spoken BNC2014
p-attribute for *lemma*	Hw	Lemma
adjective tag	AJ	JJ
adverb tag	AV	RR
cardinal number tag	CRD	MC

Table 2. The key differences between the BNC and Spoken BNC2014 tagging.

Including the adverb tag in the query is problematic but necessary due to corpus tagging inconsistencies. It improves the recall, but the number of irrelevant items becomes overwhelming. Therefore, the best course of action is, once again, to exclude specific lemmas from the query. All the lemmas excluded in the Spoken BNC2014 query were incorporated into the BNC query, except the unique '--.*', which does not occur in

---

[26] Standard similes with sources containing proper names can be searched for individually, for example, by using the query 'rich as [proper name]'. The censored name could then be guessed based on the linguistic context. However, this approach does not work for non-standard or novel similes.

the BNC. The inventory of excluded lemmas was then expanded on the basis of the frequency list.

(Q5) [pos="AJ.*|AV.*"&hw!=".*ly|about|almost|available|bad|far|fine|just|least|like|much|off|on|only|out|quite|so|soon|together|twice|up|well"] [hw="as"] [pos="AT.*|CRD"]? [pos="AJ.*"]? [pos="N..*|AJ.*"&hw!="fuck|hell|likely|other|possible|shit|well"]

(Q5) represents the final query used for simile mining from the BNC. The query could be refined even further, but the result was deemed satisfactory and required no additional changes to (Q5).

Lastly, we can expand (Q5) by an additional obligatory p-attribute [pos="N..*"]. Naturally, this eliminates most of the similes, but it also allows us to conveniently retrieve similes such as (AS) THICK AS TWO SHORT PLANKS, (AS) POOR AS A CHURCH MOUSE or (AS) TOUGH AS OLD BOOTS. (Q5) recalls these similes, but they appear at the bottom of the list due to their relatively low frequency. Furthermore, they appear incomplete in the list, such as (AS) POOR AS A CHURCH MOUSE being listed as *poor as a church*, because the query does not return similes with open compound 'noun+noun' heads in their complete form.

4.3.3 The COCA query

The COCA corpus was initially considered one of the primary data sources for this simile research, but it poses several issues related to the web interface design.[27] Perhaps the most significant disadvantage is the impossibility of using a CQL query in the web interface, which compromises the use of complex queries utilising NOT and OPTIONAL operators. This does not preclude us from searching for similes, but we remain more dependent on the accuracy of the annotation. Furthermore, due to not having the possibility to use the OPTIONAL operator, multiple queries must be used to extract variants of adjectival similes. Another inconvenience lies in the frequency list itself, as it is impossible to access the tagging of recalled items effectively. Many of the restrictions could be tied to the type of account; however, the interface design seems to target general public inquiries rather than complex linguistic research.[28]

---

[27] These issues existed at the time of this research.
[28] The COCA account interface lists restrictions related to the type of license, which mostly involve daily limits for the number of searches, KWIC lines, text analyses and other features. No mentions of data-management interface are explicitly stated.

With the restrictions mentioned, four queries designed for the extraction of similes from COCA are presented below.

(Q6)   [j*] as [n*]
(Q7)   [j*] as [a*] [n*]
(Q8)   [j*] as [a*] * [n*]
(Q9)   [j*] as [mc*] * [n*]

(Q6) returns similes with a bare noun phrase solely made of the head noun, such as (AS) CLEAR AS DAY or (AS) COLD AS ICE. (Q7) comprises the article tag, thus expecting the most frequent simile type returns, such as (AS) HIGH AS A KITE or (AS) CLEAN AS A WHISTLE. (Q8) represents the marginal type of similes whose noun phrases contain a premodifier of the head, such as (AS) PURE AS THE DRIVEN SNOW or (AS) SERIOUS AS A HEART ATTACK. Similes such as (AS) THICK AS TWO SHORT PLANKS were expected to return with (Q9).

4.3.4 The EW15 query

The EW15 is the most extensive corpus used in this research.[29] At the time of writing this thesis, more recent versions of the EW corpus are already available, but the sheer size of these variants is overwhelming; therefore, the 2015 variant was used as a compromise between size and recentness.[30]

The EW15 fully supports the CQL query type, which allows us to return to the refined query used in the BNC corpora. Therefore, the explicitly excluded lemmas were inherited and complemented by some other frequent lemmas from the EW15. The tagging differences between the EW-type and the BNC-type corpora are marginal; the only notable change is the 'tag' attribute (EW) as opposed to the 'pos' (BNC) attribute. The 'word' attribute was used instead of the 'lemma' attribute to avoid irrelevant lemmas in the noun slot, such as *[number]*, that would otherwise rank very high on the frequency list. Additionally, the adjective tag was removed from the rightmost slot due to the overwhelming return, as its presence was rather detrimental.

(Q10)   [tag="JJ.*|RB.?"&word!=".*ly|about|almost|available|bad|early|English|far|fi
          ne|interesting|just|many|much|nice|off|on|only|out|possible|quite|same|such|so|

---

[29] The corpus is hosted by Sketch Engine.
[30] The English Web 2021 contains 52,268,286,493 tokens.

soon|together|twice|up|well"] [word="as"] [tag="DT.*|CD.*"]?
[tag="N..*"&word!=" fuck|hell|likely|other|possible|shit|well"]

The final complex query (Q10) was used for a general adjectival simile search in the EW15. However, since the size of the corpus caused several issues with the frequency list, an additional query was used to complement the original one (Q10).

(Q11) [tag="JJ.*"&word!=".*ly|about|almost|available|bad|early|English|far|fine|int eresting|just|many|much|nice|off|on|only|out|possible|quite|same|such|so|soon| together|twice|up|well"] [word="as"] [tag="DT.*|CD.*"]?
[tag="N..*"&word!=" fuck|hell|likely|other|possible|shit|well"]

(Q11) does not include the adverb tag for the ground component, resulting in less frequent items being returned within the 1000-slot limit. All adjectival similes were extracted from the frequency lists of (Q10) and (Q11) returns.

4.4 Research questions and hypotheses

The purpose of this research is to map adjectival similes in corpora. This includes their frequencies, tagging, and meanings and functions. The general task is complemented by several other goals. The process of directing the research started with the inductive approach by formulating the research questions. Then, using the deductive approach and drawing on general observations, the expected answers to the research questions were stated as hypotheses. The hypotheses are formulated as statements of tendencies rather than universalities, which is a standard approach in corpus linguistics (Stefanowitsch, 2020: 68).

- **Research question No.1:** *How effectively do dictionaries utilise corpus evidence regarding adjectival similes?*
- **Hypothesis No.1:** *Higher corpus frequencies of adjectival similes tend to result in their better dictionary coverage.*

Research question number one has an obvious answer regarding old dictionaries. These had to consider what to include due to the lack of space. Additionally, such dictionaries were becoming dated with each passing day, as they could not adapt to new developments (other than new updated editions). Even idiom dictionaries could not possibly present exhaustive lists of idioms and had to select what to include carefully. However, space restrictions are no longer an issue with online dictionaries; therefore, it

remains a question of meticulous inclusion of frequent, previously unlisted adjectival similes. This process is undoubtedly governed by specific criteria for extraction and determining what is recognised as established. Needless to say, the criteria will differ in each dictionary. While predictions can be made regarding the nature of those criteria, the process generally is not transparent.

Hypothesis number 1 is simple and can easily be tested. The presence or absence of adjectival similes in dictionaries can be determined by directly searching for them. Individual entries and mentions in entries of the constituting elements count as 'presence' in the dictionary. For instance, if the simile (AS) CHEAP AS CHIPS does not have its own entry in a specific dictionary but appears in the example section of *cheap* as an existing fixed expression, it is considered present in the dictionary.[31]

- **Research question No.2:** *Are adjectival similes easy to extract from corpora using general queries?*
- **Hypothesis No.2:** *The recall-precision ratio of adjectival similes is effective when using CQL pattern-queries.*

Research question number two reviews the conception that "[s]imile in general lends itself to empirical corpus studies because of the presence of predictable lexical signals" (Moon, 2008: 35). Limited formal variability is convenient for a corpus-based investigation.

The related hypothesis requires further explanation, namely two expressions it contains. Firstly, the term *effective* varies based on the preset criteria. In this work, a recall rate above 90% is considered sufficient. As for precision, anything above 50% can be considered sufficient for revision, as several other multi-word units use the formalised frame '(*as*) ADJECTIVE *as* NOUN PHRASE'. Consequently, expecting higher precision is unrealistic. Secondly, *pattern-queries* represent the standardised queries designed for extracting adjectival similes based on the general formalised frame.

- **Research question No.3:** *Do prominent prosodic features make the adjectival simile more frequent?*
- **Hypothesis No.3:** *Prosodically marked adjectival similes do not occur more frequently than prosodically unmarked ones.*

---

[31] Only examples where the adjectival simile is listed as an existing element count as present.

Research question number three and the related hypothesis build on the premise that a striking (marked) form results in a higher frequency than an unmarked form. This includes rhythm, alliteration and assonance. Prominent prosodic features make a simile easier to remember and, possibly, to retrieve from the mental lexicon. However, prosodically marked adjectival similes are relatively marginal and are not expected to be more frequent than other prosodically unmarked types.

- **Research question No.4:** *How frequent are animal sources in adjectival similes?*
- **Hypothesis No.4:** *Adjectival similes with 'animal' sources constitute a large group within the dataset.*

Research question number four stems from the popularity of 'animal' multi-word units. Adjectival similes with 'animal' sources are easily noticeable and fall into a clearly delineated category. Therefore, they appear to be frequent within the adjectival simile stock, as suggested by, for example, Parizoska and Petrović (2017: 350), but also in implicit *like*-similes (e.g. Hanks, 2004: 35).

Hypothesis number four requires an explanation of what is considered large. Norrick (1986: 40) notes that "[i]n the ODEP animals make up by far the largest class of stock simile vehicles: animals appear 138 times or in almost 38% of the total 366 entries for stock similes."[32] In this work, a frequency of around 20% is considered high enough to be labelled a large group.

- **Research question No.5:** *Is there a significant lexical overlap between English and Czech adjectival similes?*
- **Hypothesis No.5:** *Frequent English adjectival similes have lexical equivalents in Czech.*

Research question number four assumes that the languages draw from the same European cultural heritage. Naturally, this may not apply to novel similes, especially those primarily associated with the American variant. Nevertheless, many similes are either borrowed or inherited from other languages (typically Latin, French, and German), or simply a result of a shared conceptualisation of various features.

Section 8 is dedicated to the comparison of English and Czech sets of similes. Since the Czech sample (*n=886*) of adjectival similes is significantly larger than the English

---

[32] ODEP stands for Oxford Dictionary of English Proverbs (1970).

one (n=309), the absence of a lexical equivalent in Czech can be considered reliable, albeit not definitive.

5 Transformations of adjectival similes in the dataset

This section is dedicated to exploring transformations of adjectival similes. The pattern '*(as)* ADJ *as* NOUN PHRASE' is the prototypical form of adjectival similes. As Čermák (2007: 390) points out, it is sometimes challenging to establish the base form of the simile since the standardised form does not necessarily represent the original one. Etymology can provide a definitive answer, but it does not work for all similes. Without etymological evidence, we should consider empirical evidence and rely on the frequency of the simile variants as the main factor in a synchronic approach.[33] Furthermore, the observed variation can be divided into three major types of transformation: quantitative, lexical, and grammatical. This typology is based on Čermák's (2007: 389) classification of general simile variation in Czech and adjusted for the adjectival similes in English.

Some transformation subtypes can be standardised and recur frequently. Other subtypes are non-standard and considered deviations from the standard form. Table 3 illustrates the overview of transformation types, subtypes, and representative examples from the dataset.

Transformation type	Example
QUANTITATIVE	
Expansion	*sharp as a razor → sharp as a fucking razor*
Reduction	*pure as the driven snow → pure as snow*
LEXICAL	
Adjective slot variation	*hard as nails ↔ tough as nails*
Noun slot variation	*white as a sheet ↔ white as a ghost*
GRAMMATICAL	
Adjective comparison	*good as sex → better than sex*
Noun number change	*sharp as a knife → sharp as knives*
Lexicalised phrase	*solid as a rock → rock-solid*

Table 3. Overview of adjectival simile transformation types.

5.1 The quantitative transformation

The most notable quantitative transformation involves the initial *as*, namely its omission. Multiple factors influence whether the initial *as* occurs in an adjectival simile, including communication channel, type of text, syntactic position, or personal preference.

---

[33] The frequency may not be reflecting the base form historically, but that is irrelevant for present-day use.

Despite hints that the initial *as* may be less frequent in spoken communication, sweeping generalisations should be made only carefully. Other cases of quantitative transformation almost exclusively affect the noun phrase and have two forms. It is impossible to list all potential transformations as they are unique to every standard simile. Therefore, the following paragraphs present examples to illustrate general tendencies rather than providing an exhaustive list of all possible variants.

The first form is the expansion of the noun phrase. The base forms of standard similes such as (AS) DUMB AS A BRICK or (AS) SHARP AS A RAZOR can be expanded by inserting an expletive (usually *fucking*), *(as) dumb as a fucking brick* and *(as) sharp as a fucking razor*, respectively. Furthermore, examples such as (AS) BLACK AS A RAVEN can occur in the form *(as) black as a raven's feather/wing*, the same as (AS) LARGE AS LIFE would sometimes transform into *(as) large as life and twice as natural/repulsive/beautiful*. Another possibility is the blending of two simile types resulting in an expanded form of one of them; for instance, the blending of (AS) POOR AS A CHURCH MOUSE and (AS) QUIET AS A MOUSE sometimes results in *(as) quiet as a church mouse* (*n*=14 in the COCA; *n*=32 in the EW15).

The second form is the reduction of the noun phrase (or ellipsis), where one of the components of the noun phrase is omitted, such as (AS) AMERICAN AS APPLE PIE represented by *(as) American as pie*. An interesting example is the extensive simile (AS) NERVOUS AS A LONG-TAILED CAT IN A ROOM FULL OF ROCKING CHAIRS, which is often realised simply as *(as) nervous as a cat*.[34] However, the reduced form is sometimes also expanded as a result of blending with the verbal simile BE LIKE A CAT ON A HOT TIN ROOF, resulting in *I'm nervous as a cat on a hot tin roof*. Other instances of reduction are *(as) pure as snow*, representing the original form (AS) PURE AS THE DRIVEN SNOW, or *(as) crooked as a dog's leg*, whose base form is (AS) CROOKED AS A DOG'S HIND LEG. Next, we can encounter cases of the 'genitive + noun' reduced to a simple nominative, such as *(as) cute as a bug*, instead of the original (AS) CUTE AS A BUG'S EAR. Similarly, noun phrases with the *of*-phrase postmodifier get reduced, such as *(as) dumb as a rock* or *(as) dumb as rocks* used instead of the base form (AS) DUMB AS A BOX OF ROCKS.[35] Lastly, the loss of determiner can be observed in *(as) cold as stone*, originally (AS) COLD AS ANY STONE.

---

[34] In COCA, the historically original simile (AS) NERVOUS AS A LONG-TAILED CAT IN A ROOM FULL OF ROCKING CHAIRS (*n*=5) is less frequent than its reduced variant *(as) nervous as a cat* (*n*=17).

[35] The variant *(as) dumb as a rock* also includes the grammatical transformation from *rocks* to *rock*.

Generally, we are facing two issues associated with quantitative transformation. First, deciding whether the string is an original simile or a transformation is sometimes challenging, such as the two variants *(as) dark as night* and *(as) dark as the night sky*. Dictionaries often list specific forms and explicitly comment on possible transformations, but they seldom provide reliable data regarding the origin of the simile. Therefore, it remains a judgement call to decide between observed frequency in corpora (or search engines) and lexicographic evidence. Native speaker experience and intuition might also play a part in determining the original form, but it cannot be the sole source of information. This directly affects the second issue, which is classifying the type of quantitative transformation. If we fail to establish the original form of the simile, we cannot determine whether *(as) dark as night* is an instance of reduction or *(as) dark as the night sky* is an instance of expansion. Ultimately, it is best to treat both instances as unique standard similes that are synonymous.

5.2 The lexical transformation

The second major type of transformation concerns the lexical substitution in the adjective or noun slot. Therefore, we can observe two subtypes of lexical transformation. However, it can sometimes be challenging to establish whether two strings are just transformations of the same standard simile or whether there is a semantic or functional difference between the two examples. Furthermore, varying forms are likely to result in usage differences over time despite initial functional identity.

The first type is the transformation of the adjective slot (ground), illustrated by (AS) TOUGH AS NAILS and (AS) HARD AS NAILS. Both forms are represented in online dictionaries, and neither is marked as the original form. In such cases, the frequency can become the deciding factor. A simple Google Search suggests that the string *tough as nails* is roughly four times more frequent than *hard as nails*. The corpus data further support this, as illustrated in Table 4.

	Spoken BNC2014	BNC	COCA	EW15	#
TOUGH AS NAILS	0	2	200	993	**1195**
HARD AS NAILS	7	13	42	361	**423**

Table 4. (as) tough/hard as nails in corpora.

The COCA ratio also suggests that the transformation could be attributed to regional preferences, as the BNC corpora show a preference for (AS) HARD AS NAILS, whereas

COCA favours (AS) TOUGH AS NAILS. The EW15 is a large corpus of Internet English, so the regional differences are difficult to trace. Consequently, it would be misguided to consider the internet data as conclusive evidence regarding regional preferences, but we cannot dismiss the regional factor altogether.

The second type of lexical transformation concerns the noun slot (source), for instance, (AS) WHITE AS A SHEET and (AS) WHITE AS A GHOST. In this case, the corpus data do not suggest any regional preference (see Table 5).

	Spoken BNC2014	BNC	COCA	EW15	#
WHITE AS A SHEET	0	21	95	217	**333**
WHITE AS A GHOST	0	3	51	120	**174**

Table 5. (as) white as a sheet/ghost in corpora.

Compared to (AS) WHITE AS A SHEET, the simile (AS) WHITE AS A GHOST is overall less frequent. However, the synonymous (AS) PALE AS A GHOST (*n*=25 in the COCA; *n*=80 in the EW15) brings it more in line with the frequency of (AS) WHITE AS A SHEET.

A borderline example of lexical transformation is (AS) OLD AS [GENERAL CONCEPT SHARED BY MANKIND THAT HAS BEEN AROUND FOR GENERATIONS] (ITSELF). This example illustrates the intensification function as the noun component becomes semantically bleached and is mostly reduced to the meaning 'very', for example, *old as time/humanity/mankind/civilization/history (itself)*. Similes such as (AS) OLD AS DIRT or (AS) OLD AS THE HILLS are standard similes and are not considered transformations of that string. The simile (AS) OLD AS TIME (ITSELF) is slightly problematic because the Farlex idiom dictionary lists it as a standard simile[36]; however, it does occur with the optional *itself*, which is typical of the general (AS) OLD AS [GENERAL CONCEPT SHARED BY MANKIND THAT HAS BEEN AROUND FOR GENERATIONS] (ITSELF) string.

Ultimately, the lexical transformation is often a result of non-standard semantic representation, at least initially.[37] The speaker selects a lexical item from the same semantic field instead of the conventional component. The varying lexical items are typically near-synonyms (but also other semantically related lexemes), such as now-standard transformations (AS) HARD/TOUGH AS NAILS, (AS) QUICK/FAST AS LIGHTNING or (AS) SHARP AS A RAZOR/TACK. The transformation can sometimes be attributed to

---

[36] be (as) old as time. (n.d.) Farlex Dictionary of Idioms. (2015). Retrieved April 15 2023 from https://idioms.thefreedictionary.com/be+(as)+old+as+time
[37] The term 'semantic representation' is used to describe the process of selecting words to represent the speaker's mental concepts (cf. Leech, 1981: 11).

regional or cultural differences, but the variants may also coexist in the same environment with no marked distinction in specific uses.

To conclude, the lexical transformation, according to Čermák, is somewhat problematic. This work considers formally different similes as unique types, regardless of the semantic and functional overlap. Therefore, it is best to treat these similes as synonymous instead of transformations of a single simile.

5.3 The grammatical transformation

Cases of grammatical transformation involve the comparison of the adjective or the noun component of the adjectival simile, which is accompanied by the necessary syntagmatic changes to the comparator or the determiner. Another notable grammatical change is a formal lexicalisation of the simile as a single word in terms of spelling. Therefore, we divide grammatical transformation into three main subtypes.

The first subtype is the comparison of the adjective, which is followed by the relator change (from *as* to *than*), for example, (AS) LARGE AS LIFE (*n*=18 in the BNC[38]) and (AS) GOOD AS SEX (*n*=16 in the COCA) realised as *larger than life* (*n*=71 in the BNC) and *better than sex* (*n*=99 in the COCA), respectively. These cases illustrate that the comparative form can further reinforce the intensification, often accompanied by the emphasiser *even* (e.g. *this cake is even better than sex*). Additionally, the comparative form tends to be more frequent than the original simile, as it reinforces the emphasis.

The second subtype of grammatical transformation is counting the noun. The singular-to-plural transformation with the determiner loss is observed in *sharp as knives* (*n*=53 in the EW15) as a variant of the base form (AS) SHARP AS A KNIFE (*n*=115 in the EW15). This transformation is mainly caused by the target being plural; for example, *her words were sharp as knives*. The plural-to-singular transformation with the addition of the determiner, as seen in *(as) tough as a nail* representing the base form (AS) TOUGH AS NAILS, is possibly also the result of number matching since the target is singular. Alternatively, the determiner loss or addition could be perceived as an instance of reduction or expansion, albeit conditioned by the inflexion of the head element. This highlights the fact that quantitative and grammatical transformations are often intertwined.

---

[38] The frequency includes the expanded forms such as *(as) large as life and twice as sane.*

The third common variant of grammatical transformation is a lexicalised phrase. The simile becomes a lexicalised adjective used in the attributive position as a premodifier, such as *a hard-as-nails chap* or *her dark-as-night hair*. This third subtype also combines with the first one and results in examples such as *a larger-than-life character* or *a better-than-sex experience*. The transition to the adjective part of speech triggers hyphenation, which is common in examples such as *a five-year-old child* or *a slap-in-the-face punishment*.[39] Compound adjectival similes (CAS)[40], such as PITCH(-)BLACK or RAZOR(-)SHARP, also belong in this transformation. Generally, CAS are significantly more frequent than their original similes. This is illustrated in Table 6, which lists the EW15 data for some of the most frequent CAS transformations. The table presents the absolute frequency of each form with the relative frequency in the brackets.

Conceptual simile	Simile form	CAS open	CAS hyphenated
CLEAR AS CRYSTAL	466 (0.035)	4244 (0.322)	2231 (0.169)
COLD AS ICE	680 (0.052)	27270 (2.067)	4942 (0.375)
HARD AS (A) ROCK	1012 (0.077)	5702 (0.432)	4121 (0.312)
RED AS BLOOD	278 (0.021)	6946 (0.527)	1716 (0.130)
SOLID AS A ROCK	442 (0.034)	7283 (0.552)	4575 (0.347)

Table 6. Frequent CAS compared to their simile forms in the EW15.

CAS are not the only transformations contributing to similes' familiarity, but they might be solely responsible in some examples, especially when the simile form is relatively infrequent, such as (AS) RED AS BLOOD.

---

[39] These expressions are often referred to as lexicalised phrases (Plag, 2003: 175).
[40] Norrick (1987) terms them Comparative Noun-Adjective Compounds (CNACs), which is also adopted by Novoselec and Parizoska (2012), who describe them as Noun-Adjective Compounds or, more broadly, cognate adjectival forms.

IV. DATA ANALYSIS

6 A quantitative survey of findings

This section is dedicated to a quantitative survey of the adjectival similes extracted from the respective corpora using the methodology outlined in Section 4. The analysis in this section is quantitative; statistical overviews are presented individually for each corpus, followed by comments on other phenomena related to the query results.

6.1 An overview of the data extracted from the individual corpora

The following subsections present statistical overviews for each corpus. A 'token' represents any adjectival simile occurrence, and a 'type' is a formally unique adjectival simile, regardless of its frequency. Therefore, five instances of *deaf as a post* and three instances of *high as a kite* equal eight tokens and two types. The token-type ratio illustrates an average number of tokens per simile type. The relative frequency represents the number of similes per million words within the corpus.

Another important aspect of this study is the treatment of adjectival simile types. This work considers *sharp as a razor* and *sharp as razors* two different simile types. This is because neither singular nor plural sources can be treated as the general default for adjectival similes. While some similes appear with both singular and plural sources, others are restricted to one form. For example, (AS) HARD AS NAILS, (AS) HARMLESS AS DOVES or (AS) WIDE AS SAUCERS do not have a singular source variant attested in the dataset. Similarly, many similes are attested in the dataset only with a singular source, such as (AS) CLEAR AS A BELL, (AS) HIGH AS A KITE or (AS) QUIET AS A TOMB. The ability to count the source may be attributed to some similes, but it does not work for many others. As this is directly related to the semantic content of each simile, its preferred targets, and often also the context, it is best to treat all plural-source similes as unique types.

6.1.1 The Spoken BNC2014 data

Corpus size (number of tokens)	11,422,617
Original cut-off	not applied
Absolute frequency of simile tokens (retrieved)	96
Absolute frequency of simile types (retrieved)	56
Token/type ratio (tokens per simile)	1.714
Relative frequency of simile tokens (i.p.m.)	8.404

Table 7 represents the general statistics for the Spoken BNC2014. Since this is a small corpus, the decision was to use no cut-off. The query for the Spoken BNC2014, (Q4), returns 578 matches (50.601 i.p.m.) with 510 unique types. The overall number of adjectival similes in this corpus is not surprising, with a total of 96 tokens representing 56 unique types (1.714 tokens per simile), which translates to a precision of 16.61%. The relative frequency of adjectival similes in the corpus amounts to 8.404 instances per million words. No cut-off is one of the reasons why this corpus's relative frequency of adjectival similes is higher than in the other, larger corpora.

For illustration, the cut-off at the total frequency of *n*=3 results in a list of 15 items, five of which are not similes (Figure 4).

No.	Query result	No. of occurrences	Percent
1	hard as nails	7	1.21%
2	common as muck	5	0.87%
3	high as a kite	5	0.87%
4	cheap as chips	4	0.69%
5	good as good	4	0.69%
6	guilty as sin	4	0.69%
7	so as far	4	0.69%
8	big as marbles	3	0.52%
9	blind as a bat	3	0.52%
10	even as a kid	3	0.52%
11	fresh as a daisy	3	0.52%
12	good as gold	3	0.52%
13	good as the original	3	0.52%
14	nice as the ones	3	0.52%
15	thick as pig	3	0.52%

Figure 4. The Frequency breakdown of the top 15 query results in the Spoken BNC2014.

This illustrates the cut-off sensitivity when researching phraseological units with generally lower frequencies, especially in smaller corpora like the Spoken BNC2014. The corpus size makes it possible to sift through the query result manually. Conversely, larger corpora generate vast lists cluttered with irrelevant word strings, so setting a cut-off becomes necessary with larger datasets.

6.1.2 The BNC data

Corpus size (number of tokens)	112,102,325
Original cut-off	4
Absolute frequency of simile tokens (retrieved)	788
Absolute frequency of simile types (retrieved)	115

Token/type ratio (tokens per simile)	6.852
Relative frequency of simile tokens (i.p.m.)	7.029

Table 8. The overall BNC stats.

Table 8 shows the statistics for the BNC. It comprises spoken data (11,983,120 tokens) and written data (100,119,205 tokens). Due to the relatively large corpus size, the original cut-off was set to an absolute frequency of 4 instances. The number of adjectival simile tokens retrieved from this corpus is 788, represented by 115 unique types (6.852 tokens per simile). The relative frequency of adjectival similes in the corpus amounts to 7.029 instances per million words, which is less than that of the Spoken BNC2014 (8.404 i.p.m.).

This raises two questions that can be answered by comparing the spoken section of the BNC with the Spoken BNC2014. First, we can ask whether the relative frequency of adjectival similes is higher in spoken or written corpora (or written sections of the corpora). The second question is related to potential developments in the frequency of adjectival similes in speech, in other words, if a significant change can be observed.[41] The analysis of the spoken section of the BNC yielded 79 simile tokens represented by 43 types, which translates to 6.593 instances per million words. This short inquiry shows that the written section of the corpus contains more adjectival similes (according to the relative frequency), albeit only a marginal difference. Regarding language change, a simple chi-square test indicates that the higher result in the Spoken BNC2014 is not statistically significant ($\chi^2 = 1.0197$, df = 1, $p = .31259$).

Let us also explore Wikberg's (2008) investigation into similes in the BNC, which illustrates some of the issues related to corpus-based mining. In search of adjectival similes, Wikberg (2008: 134) explains that since "the search for the 'as Adj/Adv as' pattern in the BNC yielded as many as 34,224 occurrences, the number of hits had to be reduced." The tricky part is that Wikberg never discloses the query used for the search.[42] The CQL query '[hw="as"] [pos="AJ.*|AV.*"] [hw="as"]' is equivalent to Wikberg's pattern.

First, the query returns 52,700 matches, which does not correspond to Wikberg's number of hits (34,224). Furthermore, the initial *as* is obligatory; therefore, many similes

---

[41] The BNC corpora mainly represent British English, but they offer mere fragments of a complex whole. Consequently, no sweeping generalisations can be made based on these data.

[42] Unfortunately, this is a common phenomenon in studies that employ corpus linguistics. I believe that corpus investigations should be transparent with the queries so that it is possible to assess the validity and completeness of the presented data.

are not included in the result.[43] For comparison, removing the initial *as* from the query provides us with 119,144 matches, which is more than twice the size of the original return. Naturally, such a number of matches is excessive and contains an overwhelming number of irrelevant items.

However, using this 'incomplete' query remains beneficial for query refinement. Since the '*as* ADJ/ADV *as*' pattern lacks the source element (i.e. the noun), the query's frequency breakdown lists strings such as *as fast as the previous winner* or *as fast as planes* under the same item: *as fast as*. This makes identifying the adjectives and adverbs best excluded from the search easier (see Figure 5), as the dataset itself presents us with the irrelevant items in a clearer form and under the same entry.

No.	Query result	No. of occurrences	Percent
1	as well as	18243	34.62%
2	As far as	5689	10.80%
3	as soon as	5445	10.33%
4	as long as	3933	7.46%
5	as good as	1052	2.00%
6	as early as	843	1.60%
7	as quickly as	635	1.20%
8	as fast as	557	1.06%
9	as high as	518	0.98%
10	as bad as	427	0.81%
11	as simple as	373	0.71%
12	as important as	344	0.65%
13	as late as	327	0.62%
14	as often as	309	0.59%
15	as big as	273	0.52%
16	as hard as	266	0.50%
17	as low as	261	0.50%
18	as easy as	190	0.36%
19	as strong as	188	0.36%
20	as large as	182	0.35%

Figure 5. The frequency breakdown of the 'as ADJ/ADV as' query results in the BNC.

Regarding precision, the refined BNC query (Q5) used in this work returns 21,491 matches with 18,581 different types, which is significantly less than the token return of the unrefined query with the obligatory initial *as*. Furthermore, the recall of similes is also much higher due to the absence of the obligatory initial *as*.[44] Nevertheless, the precision of the return remains under 5%.

---

[43] The ratio of tokens (*n*=52,700) to types (*n*=2,736) further illustrates the issue with the query.
[44] Keeping the obligatory initial *as* in the BNC returns just 6,731 matches, which is 31.32% of the original return. Many similes frequently occur without the initial *as*; these would not be recalled within the 31.32%.

6.1.3 The COCA data

Corpus size (number of tokens)	1,001,610,938
Original cut-off	8
Absolute frequency of simile tokens (retrieved)	7194
Absolute frequency of simile types (retrieved)	189
Token/type ratio (tokens per simile)	38.063
Relative frequency of simile tokens (i.p.m.)	7.182

Table 9. The overall COCA stats.

Table 9 shows the statistics for the COCA. At the time of this research (30/01/2023), the COCA comprised 1,001,610,938 tokens collected from 1990 to 2019. The original cut-off was set to an absolute frequency of 8 instances due to the query recall. The total number of similes retrieved from the corpus is 7194, represented by 189 unique types, which means an average of 38.063 tokens per simile. The relative frequency of adjectival similes in the COCA is 7.182 instances per million words, which is relatively similar to the BNC (7.029 i.p.m.). If the COCA expands and maintains its 'source diversity' policy, the corpus size and absolute frequency of adjectival similes will change, but the comparison with the smaller BNC suggests that the relative frequency can be expected to remain close to its current value.

ON CLICK: ▤ CONTEXT ⊕ TRANSLATE (??) Ⓖ GOOGLE 🖾 IMAGE ▶ PRON/VIDEO 📖 BOOK (HELP)

HELP	WORDS	SEE FULL LIST [?]	FREQ	TOTAL 3,711	UNIQUE 159 ✦
1	☐	HIGH AS A KITE	119	
2	☐	CLEAN AS A WHISTLE	94	
3	☐	CLEAR AS A BELL	93	
4	☐	BIG AS A HOUSE	82	
5	☐	WHITE AS A SHEET	82	
6	☐	FREE AS A BIRD	73	
7	☐	DEAD AS A DOORNAIL	73	
8	☐	SOLID AS A ROCK	73	
9	☐	CUTE AS A BUTTON	70	
10	☐	SHARP AS A TACK	66	
11	☐	HEALTHY AS A HORSE	65	
12	☐	SICK AS A DOG	61	
13	☐	SMART AS A WHIP	60	
14	☐	AVAILABLE AS A SUPPLEMENT	59	
15	☐	OLD AS THE HILLS	58	
16	☐	BLIND AS A BAT	55	
17	☐	STIFF AS A BOARD	55	
18	☐	HAPPY AS A CLAM	51	
19	☐	STRONG AS AN OX	51	
20	☐	WHITE AS A GHOST	49	

Figure 6. COCA frequency list interface.

Figure 6 illustrates the COCA frequency list interface. The top bar contains links to Google Translator, Google Search, and other linguistically less relevant phenomena.

One phenomenon should be noted regarding the simile retrieval in the COCA. Both BNC-type corpora require lemma exclusion from both the adjective and the noun slots to filter out the non-simile results. That is not the case with the COCA, where the (Q8) top 20 results are similes only (Figure 3), with some non-simile expressions occurring further down the frequency list. This is a result of the simplicity of the COCA queries with no OPTIONAL elements. The noticeable upside is the lack of irrelevant elements in the frequency list, whereas the definite downside is the hampered recall. As a result of (Q8), the simile (AS) DRUNK AS A SKUNK returned nine times. However, a simple word-string search for *drunk as a skunk* returned a total of 59 instances of the simile, which is more than 6.5 times the (Q8) result. Upon closer inspection, *drunk* is tagged "VVN@_JJ" in the instances not returned with (Q8). This effectively means that compound tags compromise the recall of part-of-speech queries in the COCA.

To deal with the excessively low recall rate, I compiled a list of all the similes from the frequency lists of (Q6-Q9) and searched them individually. The number of discrepancies returned was often significant enough to justify the additional effort. Table 10 illustrates the difference between the (Q6-Q9) and individual search results, sorted by absolute frequency.

Simile	Absolute frequency	Q6-Q9 frequency	Q6-Q9 recall effectiveness
CLEAR AS DAY	273	203	74.36%
TOUGH AS NAILS	200	186	93.00%
PLAIN AS DAY	184	115	62.5%
RIGHT AS RAIN	156	63	40.38%
WHITE AS SNOW	149	140	93.96%
CLEAN AS A WHISTLE	144	94	65.28%
HOT AS HELL	138	126	91.30%
HIGH AS A KITE	130	119	91.54%
COLD AS ICE	129	112	86.82%
GOOD AS GOLD	128	93	72.66%
CLEAR AS A BELL	123	93	75.61%
HARD AS A ROCK	122	28	22.95%
OLD AS TIME (ITSELF)	121	116	95.87%
LIGHT AS A FEATHER	118	42	35.59%
FREE AS A BIRD	112	73	65.18%
EASY AS PIE	106	90	84.91%

AMERICAN AS APPLE PIE	105	99	94.29%
WHITE AS A SHEET	95	82	86.32%
BIG AS A HOUSE	92	82	89.13%
	2625	**1956**	**74.51%**

Table 10. Absolute vs Q6-Q9 frequency of top 20 similes in the COCA.

The overall recall effectiveness of the top 20 similes (74.51%) is almost identical to the recall of the whole list of similes collected in the COCA (74.68%). This varies for individual similes; for example, (AS) DRUNK AS A SKUNK (15.25%), (AS) HARD AS A ROCK (22.95%), or (AS) LIGHT AS A FEATHER (35.59%) exhibit extremely low recall effectiveness, whereas other similes are above 90%, which could be considered satisfactory, but corrective searches were done for all the similes regardless.

The last COCA query, '[j*] as [mc*] * [n*]', was expected to return examples such as (AS) THICK AS TWO SHORT PLANKS, but that was not the case.[45] Overall, the recall of similes with a numeral ('mc' tag) in the COCA seems unreliable (e.g. (AS) ALIKE AS TWO PEAS (IN A POD)). Fortunately, such similes are generally marginal and can be searched individually. The drawback is that potentially new similes with numerals cannot be retrieved using this predesigned query.

The COCA is a highly user-friendly corpus with many built-in functions but little query-design freedom. The restrictions make it difficult to construct queries that would return complex multi-word units with an acceptable recall rate. All the similes had to be searched individually as lemmas, which made working with the corpus laborious.

6.1.4 The EW15 data

Corpus size (number of tokens)	13,190,556,334
Original cut-off	40
Absolute frequency of simile tokens (retrieved)	38251
Absolute frequency of simile types (retrieved)	218
Token/type ratio (tokens per simile)	175.763
Relative frequency of simile tokens (i.p.m.)	2.900

Table 11. The overall English Web 2015 stats.

Table 11 provides an overview of the EW15 statistics. The EW15 is a large corpus containing 13,190,556,334 tokens. Naturally, the number of adjectival simile tokens

---

[45] (AS) THICK AS TWO SHORT PLANKS occurs in the COCA just twice, and neither of the cases was listed, possibly due to complex tagging of the individual elements somehow interfering with the simple tags in the query (as was the case in the BNC corpora).

extracted from this corpus (n=38251) exceeds the combined number of the similes from all the other corpora (n=8078) investigated in this research. The absolute frequency of the retrieved simile types is 218, which equals an average of 175.763 tokens per simile.

The first query (Q10) returned a list of 1,693,011 items (109.85 i.p.m.), but the query had two significant issues. As the account type allowed only 1000 items listed in the frequency list, the initially imposed cut-off was the absolute frequency of 76.[46] Furthermore, many items on the list were not adjectival similes, which meant that the irrelevant items pushed less frequent adjectival similes out of the frequency list. Another issue was the recall effectiveness of individual similes, which generally ranged from 60% to 95%. However, (Q10) exhibited an unacceptable recall rate for some similes, namely (AS) BRIGHT AS THE SUN (14.83%), (AS) COOL IAS ICE (34.68%), (AS) LIGHT AS A FEATHER (45.87%), (AS) DARK AS NIGHT (49.57%) or (AS) PALE AS DEATH (49.61%).

The first issue was partly diminished by removing the adverb tag from the ground slot. The return of the altered query (Q11) was 1,007,115 items (65.35 i.p.m.). The absence of the adverb tag in the leftmost slot lowered the imposed cut-off to 40. Ultimately, this cut-off was accepted for the investigation in this corpus, as it is considerably larger than the COCA or the BNC-type corpora. The low relative frequency of similes in the EW15 is likely owing to the imposed cut-off. Similes are relatively infrequent, and many similes with lower frequencies (n<40) might remain undiscovered. Consequently, the relative frequency (2.900 i.p.m.) is considerably lower compared to the smaller corpora.

After investigating all the similes retrieved from the EW15 using the two queries (Q10) and (Q11), the top ten similes were searched manually to check their recall. This revealed an unsatisfactory recall rate for some similes. Consequently, a corrective search was done for every simile from the collected list, as no simile had a perfect recall (100%). This was extremely time-consuming but necessary to get the actual frequencies.

Lastly, one crucial feature of the EW15 is that it uses web-crawler data mining. Consequently, it cannot always be established if the data are produced by native or non-native speakers. This may result in more ungrammatical constructions compared to corpora with exclusively native-speaker data. For similes, this mainly concerns grammatical transformations, and the result is typically the absence of the indefinite article, for instance, *high as kite*. Furthermore, the texts produced by non-native speakers

---

[46] 76 was the absolute frequency of the 1000th item in the frequency list of the first EW15 query.

are likely to contain fewer idiomatic expressions. This is perhaps another factor contributing to the low relative frequency of adjectival similes in the EW15.

6.1.5 Corpora results comparison

Comparing the frequencies of marginal multi-word units in corpora of different sizes might seem ill-advised, but several observations can be made regardless.

Corpus	Types	T/T ratio	i.p.m.
Spoken BNC2014	56	1.714	8.404
BNC	115	6.852	7.029
COCA	189	38.063	7.182
EW15	218	182.995	2.900

Table 12. Corpora results comparison.

As Table 12 illustrates, the number of extracted types grows with the size of the corpus. This is unsurprising, as the plethora of source texts offers a greater variety of contexts where less common similes can appear. Similarly, the token-type ratio increases with the corpus size since the most common similes tend to recur.

However, the relative frequency of simile tokens seems to decrease with the increasing size of the corpus. The relative frequency difference between the BNC and the COCA is negligible, but the EW15 relative frequency suggests that adding billions of tokens from heterogeneous source texts reduces the 'simile density'. One aspect is the relatively high cut-off for the EW15 ($n=40$), but the difference between the COCA and the EW15 in the number of extracted types does not hint at a significant simile extraction loss. The frequencies of similes not retrieved from the EW15 would not be high. Moreover, most of the similes not extracted from the EW15 were likely extracted from the other researched corpora, so the simile types were not lost, just their EW15 frequencies.

Another important aspect is the type of source texts. Similes are generally associated with 'embellished language'. Therefore, they are expected to occur in literary texts, perhaps also journalistic texts and other types of language designed to impress. Nevertheless, the relative frequency of adjectival similes is the highest in the Spoken BNC2014, even if we factor in a minor extraction loss of similes in the larger corpora.

6.1.6 Cut-offs and assessing the significance

The larger corpora limit the number of returns, often due to server load. Consequently, an imposed cut-off must be accepted when using generalised queries. The imposed cut-off can be lowered by specifying the query, but this means sacrificing the recall of formally marginal simile strings, such as *(as) thick as two short planks* or *(as) poor as a church mouse.*

Ideally, no cut-off should be used for simile investigation, as it is a relatively infrequent multi-word unit compared to other constructions. However, the sheer amount of data often makes not applying any cut-off impossible. The original corpus-imposed cut-offs are presented in Table 13.

Corpus	AF cut-off	RF cut-off
Spoken BNC2014	0	0
BNC	4	0.036
COCA	8	0.008
EW15	40	0.003

Table 13. The original corpus-imposed cut-offs.[47]

A hybrid approach was used for data mining in the larger corpora. The original corpus-imposed cut-offs (n=4 for the BNC, n=8 for the COCA and n=40 for the EW15) were lowered by refining the queries as much as possible. Additionally, the top 60 similes from the EW15 frequency list were searched manually in the other three corpora, which sometimes generated similes with frequencies under the original cut-offs. These were added to the final frequency lists even if their frequencies within the individual corpora were below the cut-off. Similarly, similes ranking high in the BNC-type corpora and the COCA were searched manually in the EW15, adding more items to the EW15 simile list with an absolute frequency of n<40. While this means that some similes with lower frequencies were mined and others were not, this approach still results in more data, which is desirable. Furthermore, some adjectival similes with frequencies below the cut-offs were manually extracted while compiling various tables.

The cut-offs in this work do not function as a significance threshold. The arguments presented by Moon (2008: 20) regarding setting a significance threshold are valid, but the main goal of this work is to extract as many adjectival simile types as possible. Examples identified as nonce expressions, such as *sick as a god*, are not included in the simile lists.

---

[47] AF stands for absolute frequency, RF for relative frequency.

Since this work explores data from multiple corpora, instances of adjectival similes with low frequencies in the smaller corpora can be compared to those from the larger corpora to eliminate nonce or creative expressions reliably. Dispersion is also reviewed to ensure that a single source text does not 'bloat' any simile's frequency.[48] The relative frequency is somewhat unreliable, though, as 0.014 i.p.m. is relatively low for an adjectival simile in the COCA but reasonably high in the EW15. This indicates that relative frequency cannot be used uniformly in all four corpora, and setting an absolute frequency cut-off for each corpus remains arbitrary.

Another important aspect is the token size of adjectival similes. Setting cut-offs for single-token and multi-token lexical units differs. Five instances per million of a single-token item in a corpus is represented by five tokens per million words. The same relative frequency (5 i.p.m.) of a three-token item is represented by fifteen tokens per million words. The data show that the prototypical length of an adjectival simile is from three to four tokens.

While the question of frequency significance cannot be dismissed entirely, any frequency threshold needs to be carefully considered. As has been stated many times, similes are relatively infrequent compared to other multi-word units. For instance, if the cut-off is set to an absolute frequency of $n=6$ (0.054 i.p.m.) in the BNC, the list of adjectival similes shrinks to 53 types, which is less than half of all the types extracted from that corpus. Similes such as (AS) EASY AS PIE or (AS) CLEAR AS CRYSTAL would be eliminated from the BNC list. However, these similes rank very high in the EW15 (#3 and #13, respectively), suggesting both are significant. Furthermore, they typically appear in traditional printed dictionaries, working against the argument that they have only recently become more prominent, as could be suggested by the age of the BNC data.

To summarise, a cut-off is a convenience tool used to avoid the necessity of browsing through long lists of data. Additionally, it is often used to prevent less frequent units from negatively affecting statistical calculations. However, the massive drawback is the omission of many less common multi-word units that are nevertheless firmly established in the lexicon. As a result, the cut-offs in this work are only used when imposed by the technical limitations or size of the corpus.

---

[48] This is generally not expected with striking multi-word units (such as similes), whose functions are to embellish, jest or otherwise enliven the language. Consequently, they are unlikely to be repeated in a single communicative event.

6.2 Individual components of adjectival similes in the dataset

The individual components of adjectival similes are described in Section 3.2. This subsection explores their observed behaviour in the researched corpora.

6.2.1 Initial *as*

One of the expected phenomena was the absence of initial *as* in the more recent data. Table 14 illustrates the initial *as* distribution in the ten most frequent similes within the original BNC corpus.

Simile	(+) initial *as*	(-) initial *as*	AF
GOOD AS GOLD	21 (61.76%)	13 (38.24%)	34
RIGHT AS RAIN	12 (42.86%)	16 (57.14%)	28
QUICK AS A FLASH	3 (12.00%)	22 (88.00%)	25
PLEASED AS PUNCH	4 (18.18%)	18 (81.82%)	22
WHITE AS A SHEET	13 (59.09%)	9 (40.91%)	22
COLD AS ICE	11 (57.89%)	8 (42.11%)	19
LARGE AS LIFE	9 (50.00%)	9 (50.00%)	18
WHITE AS SNOW	9 (50.00%)	9 (50.00%)	18
SAFE AS HOUSES	5 (31.25%)	11 (68.75%)	16
OLD AS THE HILLS	13 (81.25%)	3 (18.75%)	16
	100 (45.87%)	118 (54.13%)	218

Table 14. The ten most frequent similes in the BNC.

In general, the BNC shows a relatively uniform distribution regarding the initial *as*. However, upon closer inspection, some similes show a significant preference towards omitting initial *as*, for example, (AS) PLEASED AS PUNCH (81.82%) or (AS) QUICK AS A FLASH (88.00%). Conversely, other similes prefer to retain the initial *as*, for example, OLD AS (THE) HILLS (81.25%). Naturally, the syntactic position can play a role in omitting the initial *as*; for example, (AS) QUICK AS A FLASH occurs in the initial position of the sentence in 64.00% of the cases with just a single instance of the initial *as*. In general, the data suggest that the initial *as* is significantly less frequent in the initial position of the sentence, but the evidence is inconclusive.

Furthermore, the spoken language data show an even scarcer usage of the initial *as*. Table 15 illustrates the five most frequent similes in the BNC (the spoken section) compared to the Spoken BNC2014 (Table 16).

Simile	(+) initial *as*	(-) initial *as*	AF
GOOD AS GOLD	**7** (70.00%)	**3** (30.00%)	**10**
HIGH AS A KITE	**4** (80.00%)	**1** (20.00%)	**5**
RIGHT AS RAIN	**3** (60.00%)	**2** (40.00%)	**5**
DAFT AS A BRUSH	**0** (0%)	**4** (100%)	**4**
LARGE AS LIFE	**2** (66.66%)	**1** (33.33%)	**3**
	16 (59.26%)	**11** (40.74%)	**27**

Table 15. The BNC (the spoken section) similes.

Simile	(+) initial *as*	(-) initial *as*	AF
HARD AS NAILS	**1** (0%)	**6** (100%)	**7**
COMMON AS MUCK	**0** (0%)	**5** (100%)	**5**
HIGH AS A KITE	**0** (0%)	**5** (100%)	**5**
CHEAP AS CHIPS	**1** (25.00%)	**3** (75.00%)	**4**
BLIND AS A BAT	**0** (0%)	**3** (100%)	**3**
	2 (8.33%)	**22** (91.67%)	**24**

Table 16. The Spoken BNC2014 similes.

Considering that the number of similes collected from the two corpora is insufficient for any definitive conclusion, we can still observe a growing tendency to drop the initial *as* in spoken language, as is illustrated by the drop from 59.26% to 8.33%. This is best attributed to the phenomenon of the speech economy.

It was illustrated in Section 6.1.2. that including the initial *as* in the query is detrimental to simile extraction. However, it might be used to simplify the query return, especially in larger corpora. Assuming that every adjectival simile type potentially occurs with the initial *as*, the initial type extraction could be done with a query containing the initial *as*. Consequently, all the extracted types would need to be searched individually to get the actual frequencies of the similes, including the instance without the initial *as*.

6.2.2 The adjective

The researched data reveal that tagging of the adjective is relatively unambiguous, but instances of compound tags AJ0-AV0 or JJ-RR can be found in BNC and COCA, respectively.[49] These instances demonstrate the transitional nature between adjectives and adverbs in some regional variants, where examples such as *come here quick* or *I expressed myself clear enough* blur the part-of-speech boundary. The compound tag usually

---

[49] The compound tag includes both the adjective (AJ0, JJ) and the adverb tag (AV0, RR).

comprises the adjective, making the query easier to construct, but this is not always reliable. To list just a few examples for illustration, the BNC contains the following tagging inaccuracies:

(1) *You chased it round the sink with water **cold**_NN1 as_PRP ice_NN1* (CAV 102)

(2) *we'll have you **right**_AV0 as_PRP rain_NN1 in a moment* (H82 344)

(3) *He was a **white**_NN1 as_PRP-CJS a_AT0 sheet_NN1.* (A73 1396)

Like many other words in the English lexicon, both *cold* and *right* have the potential of being different parts of speech other than the adjective. (1) shows *cold* tagged as a singular noun, possibly due to the preceding *water* causing it to be misinterpreted as a nominal compound *water_NN1 cold_NN1*. (2) shows *right* tagged as an adverb, which is likely the result of *right* being misinterpreted as an adverbial modifier of *have*. (3) contains a typo, leading to the adjective *white* tagged as a noun due to the preceding indefinite article. Ultimately, the reason for the incorrect tag is irrelevant as long as we remain aware of potential inconsistencies and work with them during the data mining. These tagging issues are relatively infrequent but will affect the results when ignored. Therefore, it is worth including the adverb tag, such as [pos="JJ.*|RR.*"] in COCA, to improve the simile recall, as the boundary between adjectives and adverbs can become quite fuzzy. Conversely, including the noun tag in the adjective slot would be counterproductive because it would result in an overwhelming number of irrelevant expressions.

6.2.3 The comparator *as*

The comparator *as* is a stable component in adjectival similes. However, its tagging is often peculiar. A brief inspection of corpus tagging in similes reveals the following (see Table 17).

Corpus	*like*	*as*
Spoken BNC2014	II	II, CSA
BNC	PRP	PRP-CJS, PRP, CJS
COCA	II, CS	II, CS(A)
EW15	IN	IN

Table 17. Tagging of like *and* as.[50]

---

[50] II and PRP stand for general preposition. CS and CJS represent subordinating conjunctions, and CSA and CS(A) are unique tags for *as* as a conjunction. IN is a tag used by the Sketch Engine for prepositions or subordinating conjunctions (i.e. subordinators).

The ambivalence of tagging perhaps results from the varying theoretical interpretations of *as*. In traditional grammar, *as* is classified primarily as a conjunction (e.g. Quirk et al., 1985) unless it occurs in comparative constructions—then it is treated as a preposition. However, some theoretical works consider it a preposition regardless of its function, namely CaGEL (Huddleston & Pullum, 2002).[51] As illustrated in Table 4, the corpora approach the tagging of *like* somewhat uniformly. When it comes to *as*, the tagging is much less stable, which results from its functional potentiality. The compound tag is a practical solution that allows queries to return a similar result, regardless of the tagging decision, but the tagging variability is not always solved using the compound tag. Therefore, if we opt for a part-of-speech tag, the query must contain two separate tags to extract all the adjectival similes, including those where *as* is inaccurately tagged as a conjunction only. Ultimately, the best solution for extracting the standardised form is to use the lemma attribute and avoid the part-of-speech attribute whenever possible.

6.2.4 The noun

Perhaps unsurprisingly, the noun component is not tagged unequivocally, as some nouns often prove problematic in the automatic annotation. These are chiefly morphologically simple lexical units whose form potentially represents multiple part-of-speech categories.[52] This concerns especially words that can be both nouns and adjectives. The corpus data show tagging inaccuracies in similes such as (AS) GOOD AS GOLD. Outside the simile, the lexeme *gold* can be either a noun or an adjective, which is probably why its tagging is somewhat inconsistent in the BNC.

(4) *One lot are as good*_AJ0 *as*_CJS ***gold*_NN1**. (BLW 1610)

(5) *She's as good*_AJ0 *as*_CJS ***gold*_NN1-AJ0**. (KDY 828)

(6) *they're as good*_AJ0 *as*_CJS ***gold*_AJ0-NN1** *those are*. (KC1 1157)

The standard tagging of *gold* can be seen in (4). The compound tag in (5) lists the noun tag as the first (primary) tag, which is reversed in (6), where the adjective tag assumes the primary role. Although seemingly trivial, the tagging inconsistency may affect the query recall. In fact, the three BNC instances with the compound tag in (6) are not returned in a search when the query contains just the [pos="N..*"] tag. In this case, using the lemma attribute for the noun component is impossible; hence it is best to use

---

[51] This work does not aim to assess the theoretical validity of the conflicting approaches to conjunctions.
[52] Words that are the base or resulting items of conversion.

the OR operator (|) in the tag [pos="N..*|AJ.*"]. However, the inclusion of the adjective tag will recall many non-simile constructions with adjectives, the most frequent of which are best excluded from the simile query. This 'unclogging' is achieved by browsing the frequency list and pinpointing the unambiguous adjectives irrelevant to a simile search.

6.3 The representative sample of adjectival similes

The essential part of this research was mining adjectival similes from the selected corpora and creating a list of all the unique types within each corpus. These lists were put together, and any duplicates were removed. After a thorough revision based on the criteria outlined in sections 4 and 5, the final list comprises 309 adjectival simile types collected from all four corpora (Appendix 1). This sample provides a picture of the current use of adjectival similes in English and serves as a basis for analysis in the following sections of this work.

7   Frequency-based core of adjectival similes

This section analyses the 60 (19.42%) most frequent adjectival similes from the whole set of 309 adjectival similes collected, as analysing every simile from the sample individually is impossible due to space limitations.

The analysis of each simile presented in this section is descriptive, with quantitative data for illustration. An example from one of the corpora is always provided for each simile. The structure and content in the example texts from the corpora are unedited – any typos or spelling mistakes are preserved. Each simile is interpreted in terms of meaning and its typical interaction with frequent targets. Where relevant or otherwise significant, etymological comments are included to explain the meaning or function of the simile. Semantically and structurally related similes and comments concerning prominent formal properties are included.

The simile analysis is divided into five subsections, each comprising 12 adjectival similes. Each subsection contains an overview frequency table presenting the absolute frequency of each simile with the relative frequency in the brackets.

7.1   1st list (similes ranking #1-12)

Simile	EW15	COCA	BNC	SBNC2014	Total
CLEAR AS DAY	1341 (0.101)	273 (0.273)	8 (0.071)	0 (0)	**1622**
WHITE AS SNOW	1150 (0.087)	149 (0.148)	18 (0.161)	0 (0)	**1317**
TOUGH AS NAILS	993 (0.075)	200 (0.200)	2 (0.018)	0 (0)	**1195**
EASY AS PIE	1061 (0.080)	106 (0.106)	4 (0.036)	0 (0)	**1171**
HARD AS A ROCK	805 (0.061)	122 (0.122)	1 (0.009)	0 (0)	**928**
GOOD AS GOLD	757 (0.057)	128 (0.128)	34 (0.303)	3 (0.263)	**922**
LIGHT AS A FEATHER	787 (0.060)	118 (0.118)	9 (0.080)	0 (0)	**914**
COLD AS ICE	680 (0.052)	129 (0.129)	19 (0.169)	0 (0)	**828**
PLAIN AS DAY	631 (0.048)	184 (0.184)	2 (0.018)	0 (0)	**817**
OLD AS TIME	675 (0.051)	121 (0.121)	4 (0.036)	0 (0)	**800**
SMOOTH AS SILK	695 (0.053)	90 (0.090)	7 (0.062)	0 (0)	**792**
AMERICAN AS APPLE PIE	534 (0.040)	105 (0.105)	0 (0)	0 (0)	**639**

Table 18. Frequencies of #1-12 similes in the corpora.

The first list (Table 18) contains the top 12 most frequent adjectival similes in the dataset. They all exceed the frequency threshold of *n=500* in the EW15 and *n=600* overall. Interestingly, only (AS) GOOD AS GOLD occurs in the Spoken BNC2014. In the COCA, (AS) SMOOTH AS SILK is the only simile from this list that does not reach *n>100* (i.p.m. *n>0.1*).

One striking aspect of this list is that it contains no simile with an animal source. As described in Section 10.2.1, 'animal' similes are quite frequent; 54 similes (17.48%) in the dataset contain an animal source. Another common adjectival simile group are 'colour' similes, represented by just a single type in this list: (AS) WHITE AS SNOW.

This simile list also illustrates the highest relative frequencies for the EW15 and the COCA. The relative frequency of (AS) CLEAR AS DAY in the EW15 (0.101 i.p.m.) shows that while it is the most frequent simile in that corpus, its relative frequency is still much lower than that of (AS) GOOD AS GOLD in the BNC (0.303 i.p.m.).

#1 (AS) CLEAR AS DAY

*However, many of the other common traits were **clear as day**: petty theft, vandalism and setting things on fire, inability to feel remorse, excessive lying, defiance of authority, lack of empathy.*

(EW15; 256958)

The most frequent simile in the dataset is (AS) CLEAR AS DAY. The meaning of this simile is 'easy to see or understand' or 'obvious'. A common target is *memories* and *laws* but also a visual experience.

The simile (AS) PLAIN AS DAY appears to be semantically and functionally identical, but dictionaries do not list it as a variant of (AS) CLEAR AS DAY.[53] The data suggest that (AS) CLEAR AS DAY is more common in American English. Moreover, it is the most frequent simile in both the COCA (*n=273; 0.273* i.p.m.) and the EW15 (*n=1341; 0.101* i.p.m.). Additionally, it is the only simile in the EW15 to reach the relative frequency of 0.10 instances per million words.

Another simile with the meaning 'easy to see or understand' is (AS) CLEAR AS CRYSTAL. However, this simile is not used with the meaning 'obvious'. (AS) CLEAR AS

---

[53] This is most likely caused by the dictionary organisation of entries, where the focus remains on the leftmost element for the purpose of alphabetical ordering.

DAY also has a humorous counterpart (AS) CLEAR AS MUD, which is used as irony with the meaning 'not easy to understand' or 'not obvious'.

Lastly, the simile (AS) BRIGHT AS DAY shares the same source, but its meaning differs. The source in (AS) CLEAR AS DAY alludes to clarity in comprehension, whereas in (AS) BRIGHT AS DAY, it refers to light.

#2 (AS) WHITE AS SNOW

Susan's face was as white as snow and she was shaking with cold.

(BNC; GUL 180)

The second most frequent simile in the dataset, (AS) WHITE AS SNOW, means '(extremely) white or pale'. Typical targets are *skin, hair* or *clothes*.

The similes (AS) WHITE AS A SHEET and (AS) WHITE AS A GHOST are sometimes listed as variants of (AS) WHITE AS SNOW.[54] However, they often suggest that the whiteness results from an illness or shock. Consequently, the two forms cannot be considered variants of (AS) WHITE AS SNOW. Both (AS) WHITE AS A SHEET and (AS) WHITE AS A GHOST have related similes with the ground *pale*, such as (AS) PALE AS A GHOST or (AS) PALE AS DEATH, which further supports the 'white/pale due to illness or shock' interpretation.[55]

The Cambridge and Oxford dictionaries do not list (AS) WHITE AS SNOW, arguably because snow is less common in the climate of the United Kingdom, making it a far less prominent concept. Nonetheless, the simile is attested in the BNC (n=18) with a significant relative frequency (0.161 i.p.m.), which is slightly higher than its relative frequency in the COCA (0.148 i.p.m.).

#3 (AS) TOUGH AS NAILS

*They made fine conscripts, **tough as nails** and accustomed to doing as ordered.*

(COCA; FIC: Bk:LordsEmptyLand)

---

[54] white as a sheet. (n.d.) McGraw-Hill Dictionary of American Idioms and Phrasal Verbs. (2002). Retrieved October 17 2022 from https://idioms.thefreedictionary.com/white+as+a+sheet
[55] Semantically related similes with the ground *white*, (AS) WHITE AS DEATH and (AS) WHITE AS CHALK, also refer to illness, fear or shock.

This simile is closely related to (AS) HARD AS NAILS. It is also used with the meanings 'physically or mentally tough', 'determined', and 'showing little sympathy'. The target is predominantly a person or their behaviour or attitude.

Similes related to (AS) TOUGH AS NAILS that contain the ground *tough* are (AS) TOUGH AS OLD BOOTS, (AS) TOUGH AS AN OLD BOOT, (AS) TOUGH AS LEATHER, (AS) TOUGH AS OLD LEATHER or (AS) TOUGH AS OLD SHOE LEATHER. All share the same range of meanings and functions, but the sources containing *old* might have negative connotations.

Many other similes can mean 'physically tough', mostly with the target *hard*, such as (AS) HARD AS NAILS, (AS) HARD AS STONE, (AS) HARD AS IRON, (AS) HARD AS GRANITE, (AS) HARD AS DIAMOND, (AS) HARD AS ROCK, (AS) HARD AS A ROCK or (AS) HARD AS CONCRETE. However, these tend to differ in the connotative meanings.

The simile that is the closest in overall usage to (AS) TOUGH AS NAILS is doubtless (AS) HARD AS NAILS. They have identical meanings and a range of functions, with the only difference being regional preference. The researched data suggest that (AS) TOUGH AS NAILS is preferred in the American dialect over the simile (AS) HARD AS NAILS, which is more common in British English data.

#4 (AS) EASY AS PIE

*You just learned how to subtract fractions, and it was **easy as pie**!*

<div align="right">(COCA; TV: When Calls the Heart)</div>

This simile describes 'something easy, requiring little skill or effort'. The target is usually an activity, such as sports, manual work or cooking, but it is not restricted to anything specific.

An expanded variant (AS) EASY AS APPLE PIE can be found in some dictionaries; however, only three instances of this variant appear in the COCA and a single instance in the BNC. Therefore, this variant can be considered marginal, possibly influenced by the simile (AS) AMERICAN AS APPLE PIE. Some dictionaries also note that the reference is to eating the pie, not making it.[56] The idiom *piece of cake* is historically related.

The simile (AS) EASY AS ABC contains the same ground and is sometimes listed as a variant conveying a similar meaning.

---

[56] easy as pie. (n.d.) The Dictionary of Clichés. (2013); Collins COBUILD Idioms Dictionary, 3rd ed.. (2012). Retrieved October 17 2022 from https://idioms.thefreedictionary.com/easy+as+pie

#5 (AS) HARD AS A ROCK

*I was **hard as a rock** and wondering why I had never thought to do any of this before.*

<div align="right">(EW15; 267969)</div>

This simile is somewhat peculiar. Dictionaries list it with the meaning 'very hard' and provide targets such as *pillow* or *cake*. However, the data show that the simile (AS) HARD AS A ROCK is often used to describe male sexual arousal. A closely related simile with the same ground and source is (AS) HARD AS ROCK. As the EW15 data suggest, the range of meanings greatly overlaps with (AS) HARD AS A ROCK. Potentially near-synonymous similes, such as (AS) HARD AS IRON, (AS) HARD AS STONE, (AS) HARD AS STEEL or (AS) SOLID AS A ROCK, usually do not refer to sexual arousal and have different associative meanings.[57]

The meaning of (AS) HARD AS A ROCK is likely reinforced by a song by the band AC/DC. Moreover, the simile repeats in the song's chorus, significantly affecting its overall frequency in the internet corpora that use web crawlers to collect data. For illustration, the overall frequency of the simile in the COCA is *n*=122, with 28 instances (22.95%) coming from the AC/DC song. The simile occurs only once in the BNC-type corpora, as they generally do not contain music lyrics or sexually explicit language. Additionally, the BNC data are older than the song, which was released in 1995.

Lastly, the CAS transformation ROCK(-)HARD is significantly more frequent than the simile form: 8662 instances (*n*=1716 hyphenated) in the EW15, 477 instances (*n*=320 hyphenated) in the COCA, 43 instances (*n*=27 hyphenated) in the BNC, and 6 instances in the Spoken BNC2014.

To summarise, (AS) HARD AS A ROCK tends to become fairly frequent with the inclusion of sexually explicit content and song lyrics (often referring to sexual scenarios). Once these sources are eliminated, the frequency of this simile decreases significantly.

#6 (AS) GOOD AS GOLD

*oh (.) oh I'm (.) shattered after looking after --ANONnameM again (.) neck ache back ache (.) and he's **as good as gold** but*

<div align="right">(Spoken BNC2014; SPLU 44)</div>

---

[57] Occasionally, some of the similes are attested with the 'sexual arousal' meaning but it is not prevalent.

The simile (AS) GOOD AS GOLD occurs in all the researched corpora. It usually means 'good (in behaviour) or non-conflicting'. The target is typically a child; however, references to adults are also relatively common.

(AS) GOOD AS GOLD is the most frequent simile in the BNC (n=34; 0.303 i.p.m.) and is one of the few similes that occur more than once in the Spoken BNC2014 (n=3; 0.263 i.p.m.).

Formally, the simile contains alliteration of both the initial and final consonants.

#7 (AS) LIGHT AS A FEATHER

*She had been **as light as a feather** to carry and her small hands were as cold as ice.*

(BNC; EVC 1364)

Meaning '(exceptionally) light in weight', the simile typically refers to the physical weight or figurative lightness of texture (mainly used for food). The simile (AS) LIGHT AS AIR shares the ground *light* and appears to be synonymous.

The simile has a CAS transformation FEATHER(-)LIGHT. In the BNC, the CAS (n=24 hyphenated; n=3 open) is more frequent than the simile (n=9). However, the simile (n=118) is more frequent than the CAS (n=80 hyphenated; n=12 open) in the COCA. Overall, this simile's CAS is not as frequent as some of the CAS transformations of other adjectival similes, such as CRYSTAL(-)CLEAR or ROCK(-)HARD, which tend to outnumber their standard simile forms significantly.

This simile (AS) LIGHT AS A FEATHER often occurs together with the simile (AS) STIFF AS A BOARD. The string *light as a feather, stiff as a board* commonly refers to a children's slumber party game or a levitation trick in popular culture.[58]

The origin of this simile can be traced back to 1548, eventually motivating the name of a weight class (*featherweight*) in combat sports.[59]

#8 (AS) COLD AS ICE

*From then on his manner towards me was **as cold as ice**, which caused me great pain.*

(BNC; FR6 2534)

---

[58] https://en.wikipedia.org/wiki/Light_as_a_feather%2c_stiff_as_a_board
[59] as light as a feather. (n.d.) The Dictionary of Clichés. (2013). Retrieved March 20 2023 from https://idioms.thefreedictionary.com/as+light+as+a+feather

The basic meaning of this simile is 'very cold', usually referring to water, but also other physical objects, such as body parts. When the target is a person, the simile means 'unfriendly or unsympathetic' or even 'hostile', as illustrated by the example.

The simile is generally popular in song lyrics and literature, making it cliché. When used in speech, it marks a higher style but also humour and may attract mockery. The simile has a CAS transformation ICE(-)COLD, which is significantly more frequent. Compare *ice cold* (*n*=5702 in the EW15; *n*=473 in the COCA) and *ice-cold* (*n*=4121 in the EW15; *n*=573 in the COCA) with *cold as ice* (*n*=680 in the EW15; *n*=129 in the COCA).

A related simile is (AS) COLD AS (A) STONE, once again frequently used in its CAS transformation STONE(-)COLD, typically describes people, their eyes or their behaviour. It is rare in the simile form and is not used to describe literal coldness.

#9 (AS) PLAIN AS DAY

*Ezekiel 18:20 makes it **plain as day** that we are responsible only for our own sins.*

(COCA; BLOG: theblaze.com)

The simile (AS) PLAIN AS DAY means 'obvious or noticeable' or 'easy to understand'. Based on the corpus data, this simile appears identical in meaning to (AS) CLEAR AS DAY. However, the dictionaries do not link these two similes, possibly due to the ground being a different word. Furthermore, (AS) PLAIN AS DAY is not attested in the BNC corpora, which suggests that it may be a variant of (AS) CLEAR AS DAY in American English.

Other related similes listed in dictionaries are (AS) PLAIN AS THE NOSE ON ONE'S FACE and (AS) PLAIN AS A PIKESTAFF, but their frequencies in the dataset are significantly lower compared to (AS) PLAIN AS DAY. Formally, the simile is assonantal.

#10 (AS) OLD AS TIME

*Every star has a story. Some are **as old as time**, faint and almost forgotten.*

(EW15; 23231)

The simile (AS) OLD AS TIME is used to describe something 'timeless or eternal'. The simile sometimes occurs with an expanded source, as illustrated by *This city is old as time itself*, which is typical of similes of this type, such as (AS) OLD AS HUMANITY, (AS) OLD AS MANKIND or (AS) OLD AS CIVILIZATION. The source in these similes is a general

concept that is considered 'extremely old' or even 'eternal' within our culture and *itself* functions as an emphasiser. By definition, it is a hyperbole, often used humorously.

Of the 675 instances of this simile in the EW15, 131 include the *itself*. The frequency of this simile is affected by the film The Beauty and the Beast, which features a song with the lyrics *a tale as old as time*. Consequently, corpora using web crawlers to gather data contain more tokens of this simile due to the song lyrics' repetition.

#11 (AS) SMOOTH AS SILK

*The familiar raspiness is still there, yet refined, mature and **smooth as silk.***

(COCA; WEB: amazon.com)

This simile is used with several meanings. Firstly, the primary meaning, 'very smooth and soft', refers to the surface of an object, the vocal quality, or a part of the body, such as the *face*. Similes that also express the meaning 'very smooth' are (AS) SMOOTH AS GLASS, frequently with the target *water*, or (AS) SMOOTH AS BUTTER, which is not found in dictionaries but is relatively frequent in the EW15 (n=223; 0.017 i.p.m.).

Secondly, the meaning of (AS) SMOOTH AS SILK may be 'without obstacles or hindrance', said of *negotiations, progress* or any general activity. This meaning can also be expressed by (AS) SMOOTH AS BUTTER. Thirdly, it can describe a person, highlighting the slippery nature of silk with the meaning 'polite but insincere'.

The third meaning shares features expressed by similes (AS) SLIPPERY AS A SNAKE or (AS) SLIPPERY AS AN EEL. However, (AS) SMOOTH AS SILK also suggests deception and charm. Once again, (AS) SMOOTH AS BUTTER is also used to describe a charismatic person with questionable intentions.

Another related simile sharing the source is (AS) SOFT AS SILK, whose meaning is 'soft or smooth to the touch'.

Formally, (AS) SMOOTH AS SILK is alliterating.

#12 (AS) AMERICAN AS APPLE PIE

*That is **as American as apple pie**, it is what our country was founded upon.*

(COCA; WEB: athensnews.com)

The simile (AS) AMERICAN AS APPLE PIE is attested only in the COCA and English Web 2015. Consequently, the implication is that it is predominantly used in the American

dialect. The meaning of this simile is 'exemplary or typical of the American way of life'. Americans often use it as a cliché when referring to something typically American – either with pride or disdain. Popular culture media, such as the song American Pie by Don McLean or the film series American Pie, further reinforce the idea that *apple pie* represents American culture.

Formally, the simile is somewhat peculiar. Firstly, the source is a compound, which is relatively uncommon in similes. Secondly, the ground contains four syllables, effectively making it the longest ground among the most frequent simile examples.

7.2   2nd list (similes ranking #13-24)

Simile	EW15	COCA	BNC	SBNC2014	Total
FREE AS A BIRD	444 (0.034)	112 (0.112)	12 (0.107)	1 (0.088)	**569**
CLEAR AS A BELL	416 (0.032)	123 (0.123)	6 (0.054)	2 (0.175)	**547**
SOLID AS A ROCK	442 (0.034)	85 (0.085)	13 (0.116)	0 (0)	**540**
CUTE AS A BUTTON	464 (0.035)	72 (0.072)	0 (0)	0 (0)	**536**
RIGHT AS RAIN	340 (0.026)	156 (0.156)	28 (0.250)	2 (0.175)	**526**
CLEAR AS CRYSTAL	466 (0.035)	41 (0.041)	5 (0.045)	0 (0)	**512**
COOL AS A CUCUMBER	381 (0.029)	75 (0.075)	5 (0.045)	0 (0)	**461**
STRAIGHT AS AN ARROW	371 (0.028)	76 (0.076)	10 (0.089)	1 (0.088)	**458**
QUICK AS A FLASH	397 (0.030)	30 (0.030)	25 (0.223)	1 (0.088)	**453**
CLEAR AS MUD	413 (0.031)	33 (0.033)	3 (0.027)	0 (0)	**449**
BLACK AS NIGHT	364 (0.028)	74 (0.074)	8 (0.071)	0 (0)	**446**
BRIGHT AS THE SUN	391 (0.030)	48 (0.048)	4 (0.036)	0 (0)	**443**

Table 19. Frequencies of #13-24 similes in the corpora.

The second list (Table 19) presents a group of similes ranking #13 to #24 in the dataset. This list shows a more even distribution across the corpora than the first. Just a single adjectival simile is absent from the BNC, and five types appear in the Spoken BNC2014. The absolute frequency in this list ranges from *n=443* to *n=569*.

The list contains a single 'animal' simile, (AS) FREE AS A BIRD, and one 'colour' simile, (AS) BLACK AS NIGHT. The simile (AS) RIGHT AS RAIN is the least frequent in the EW15 from this list, but its high frequency in the other corpora makes up for it, especially in the COCA (0.156 i.p.m.) and the BNC (0.250 i.p.m.).

#13 (AS) FREE AS A BIRD

*And here you are **free as a bird**, for once, provided you come back to your cage in time.*

(BNC; G0M 622)

(AS) FREE AS A BIRD is one of the similes occurring in all the researched corpora. The meaning is 'free, without hindrance or limitations' or 'free of responsibilities or obligations', and the simile predominantly describes people.

The simile (AS) FREE AS (THE) AIR is sometimes listed in dictionaries as a variant.[60] However, it is not attested in the dataset. Another variant, (AS) FREE AS THE WIND, is not found in any dictionary but appears both in the COCA (n=14) and the EW15 (n=84). Semantically, the data suggest that (AS) FREE AS THE WIND is interchangeable with (AS) FREE AS A BIRD.

#14 (AS) CLEAR AS A BELL

*(...) I got there at one I mean London was **as clear as a bell** yeah I mean if you'd been a pedestrian (...)*

(Spoken BNC2014; SN64 2051)

This simile occurs in all the researched corpora. Its meaning is either 'clearly audible' or 'easy to understand or navigate'. The original (more literal) meaning alludes to the clarity of a bell sound, which is also expressed by the dated (AS) CLEAR AS A WHISTLE. The meaning 'easy to understand' overlaps with the meanings of (AS) CLEAR AS CRYSTAL, (AS) CLEAR AS DAY and (AS) PLAIN AS DAY, with possibly different connotations in some cases.

Typical targets of (AS) CLEAR AS A BELL are various sounds, *words* or language in general, but also situations or places. The first recorded use of the simile dates back to 1670.[61]

#15 (AS) SOLID AS A ROCK

*But I really like the bass; it feels **as solid as a rock** and represents a more worthwhile investment for not a lot more money.*

---

[60] free as air/as a bird. (n.d.) *Collins COBUILD Idioms Dictionary, 3rd ed.*. (2012). Retrieved November 21 2022 from https://idioms.thefreedictionary.com/free+as+air%2fas+a+bird

[61] (as) clear as a bell. (n.d.) The American Heritage Dictionary of Idioms by Christine Ammer. (2003, 1997). Retrieved January 31 2023 from https://idioms.thefreedictionary.com/(as)+clear+as+a+bell

The simile (AS) SOLID AS A ROCK conveys two established meanings: 'extremely solid or sturdy' or 'reliable or dependable'. Seemingly similar to (AS) HARD AS A ROCK and (AS) HARD AS ROCK, this simile is actually quite different, as it refers to *stability* rather than *toughness*.[62] Additionally, (AS) SOLID AS A ROCK is not used to describe male sexual arousal. Typical targets are *relationship, individual's behaviour* or *performance*. The EW15 also contains 45 instances of *solid as rock*. Given the nature of internet data, it may be interpreted as an uncountable use of *rock*, or a mistake. If treated as a proper simile, (AS) SOLID AS ROCK shares both the meaning and function with (AS) SOLID AS A ROCK.

The simile has a CAS transformation ROCK(-)SOLID, which is significantly more frequent than the original simile form; an overwhelming 11858 instances (*n*=4575 hyphenated) in the EW15, 948 instances (*n*=570 hyphenated) in the COCA, 46 instances (*n*=27 hyphenated) in the BNC, and 11 instances (all open) in the Spoken BNC2014. Consequently, the CAS transformation might make the simile feel significantly more familiar to speakers than its frequency would suggest.

Formally, the simile is assonantal, but the repetition of /ɒ/, or /ɑ/ in General American, is less prominent due to three other vowels. However, it is more prominent in ROCK(-)SOLID.

#16 (AS) CUTE AS A BUTTON

*(…) his movies make me laugh and he's **cute as a button**.*

The simile (AS) CUTE AS A BUTTON means 'charming or dainty', especially about children. The allusion to the button suggests that it is mainly applied to little things, but the simile is also frequently used when talking about attractive people.

In the dataset, the simile does not appear in the BNC corpora, supposedly making it exclusive to the American variant, as some dictionaries suggest. The word *cute* further reinforces this assumption, as it is used predominantly in American English with the

---

[62] In some cases, (AS) SOLID AS A ROCK still refers to *toughness* in the dataset, but it is not the prevalent ground.

meaning 'pretty'.[63] This might also make the simile (AS) PRETTY AS A PICTURE synonymous.

Similes with similar meanings are (AS) CUTE AS A BUG or (AS) CUTE AS A BUG'S EAR, with the former being a shortened (and usually not listed in dictionaries) variant of the latter.

#17 (AS) RIGHT AS RAIN

*no I think it'll be fine and by Monday hopefully I'll be **right as rain** (…)*

<div align="right">(Spoken BNC2014; S9YG 96)</div>

The simile (AS) RIGHT AS RAIN occurs in all the researched corpora. It can mean 'perfectly healthy', 'correct or truthful', 'in working order' or 'satisfactory or desired state'. With the first meaning, the target is typically a person who has recently suffered from an illness or injury, and this simile indicates a complete recovery. The second meaning also takes people as targets and describes correctness or veracity, for example, *Ain't that right, Johnny? Right as rain, sir* (COCA). Interestingly, this meaning is often not listed in major dictionaries. The third meaning describes tools, machines and similar concepts that work as intended or designed. The fourth meaning is used to assess a state of affairs positively. It can also be used to describe one's mental state or preparedness, such as a response to the question *Are you alright?*

The origin of (AS) RIGHT AS RAIN is unclear. One interpretation is that *right* means *dependable*, which would explain its origin as a reference to the notoriously rainy British weather.[64] Some dictionaries list (AS) RIGHT AS A TRIVET or (AS) RIGHT AS A GLOVE as similes with nearly identical meanings, but neither occurs in the researched corpora with any significant frequency.

Formally, the simile is alliterating.

#18 (AS) CLEAR AS CRYSTAL

*Things which had bewildered him were **clear as crystal**, problems which had daunted and defied him gave like locks opening to the right key.*

---

[63] Harper, D. (n.d.). Etymology of cute. Online Etymology Dictionary. Retrieved June 13, 2023, from https://www.etymonline.com/word/cute

[64] Right as Rain. (n.d.) The Dictionary of Clichés. (2013). Retrieved December 25 2022 from https://idioms.thefreedictionary.com/right+as+rain

Another frequent simile with the ground *clear* is (AS) CLEAR AS CRYSTAL. Similar to (AS) CLEAR AS DAY and (AS) PLAIN AS DAY, the meaning of this simile is either 'easy to understand' or 'visually clear, transparent'. The latter makes the simile functionally different, and the data suggest that common targets of this simile are *water, river* or *view* – hence, referring to visual clarity. However, it is still used, even if less frequently, with *instructions, explanations* or *ideas* with the meaning 'understandable'.

The relative frequency of (AS) CLEAR AS CRYSTAL is similar in all three corpora where it is attested, which is unusual given the size and mode differences.

The simile also has a CAS transformation CRYSTAL(-)CLEAR, which is significantly more frequent in the dataset. In the BNC, the simile form occurs five times, while the CAS search returns 107 open and 52 hyphenated instances. The EW15 reveals the true prevalence of the CAS form: 27270 open and 4942 hyphenated instances. The CAS transformation frequency might make the simile more familiar to speakers than its actual simile form would suggest.

Formally, this simile is alliterating.

#19 (AS) COOL AS A CUCUMBER

*He settled right in and resumed his breakfast, and was **as cool as a cucumber**.*

This simile means 'extremely calm' or 'self-possessed'. The figurative meaning alludes to a cool temperature representing a calm and collected demeanour (as opposed to the volatile, emotional states). The simile almost exclusively describes human targets, with animals occasionally taking the place of the target as well, especially in cases of animal anthropomorphisation.

Formally, this simile is alliterating. Additionally, the source *cucumber* comprises three syllables, which is unusual. Most simile sources contain either one or two syllables. However, since the ground is monosyllabic, the simile remains rhythmical with the double dactyl form.[65]

---

[65] In a dactyl foot, the stressed syllable is followed by two unstressed syllables. In a double dactyl, the foot repeats.

#20 (AS) STRAIGHT AS AN ARROW

His cheekbones were high, his nose straight as an arrow (…)

(BNC; JY6 104)

This simile has two main meaning. Firstly, it means either 'very straight' or 'erect', with typical targets such as *roads* and *tunnels*, or body parts (e.g. *nose* and *back*). The second meaning is 'honest and truthful', ascribed almost exclusively to *person* targets to describe their character and behaviour. Some dictionaries consider this simile to occur primarily in British or Australian variants[66], but the data do not support that claim. What is more, the Cambridge Dictionary does not even list the simile. A simile with the opposite meaning is (AS) SLIPPERY AS AN EEL.[67]

A related simile with an identical function is (AS) STRAIGHT AS A DIE. It describes the same targets with the same meanings. It is also often reported as primarily occurring in British and Australian variants. However, this simile is significantly less frequent in the dataset, with its frequency not exceeding the cut-off in either the COCA or the EW15. Another simile sharing the ground *straight* is (AS) STRAIGHT AS A RAMROD, but its meaning alludes to an erect or stiff posture, often referring to strict and rigid behaviour (also (AS) STIFF AS A RAMROD).

#21 (AS) QUICK AS A FLASH

Quick as a flash and game for a good laugh, I said (…)

(BNC; H9Y 492)

This simile occurs in all the researched corpora. Its meaning is 'very quick', with typical targets being almost exclusively people's behaviour (*reaction, response, action*). However, the simile sometimes describes events as well. This simile predominantly functions as an adverbial and often occurs at the beginning of a sentence.

Nearly identical similes in terms of meaning include (AS) FAST AS LIGHTNING, (AS) QUICK AS A WINK, (AS) QUICK AS LIGHTNING, (AS) FAST AS THE WIND, (AS) SWIFT AS AN ARROW, or (AS) SWIFT AS THE WIND. The last two similes occur in the researched corpora

---

[66] straight as an arrow. (n.d.) Farlex Dictionary of Idioms. (2015). Retrieved December 12 2022 from https://idioms.thefreedictionary.com/straight+as+an+arrow
[67] straight as a die. (n.d.) Farlex Partner Idioms Dictionary. (2017). Retrieved December 12 2022 from https://idioms.thefreedictionary.com/straight+as+a+die

with low frequency or do not occur at all despite being listed in dictionaries. The meaning 'very quick' is general; therefore, different grounds and targets can express it.[68]

#22 (AS) CLEAR AS MUD

*My ticket was open-ended, and my purpose was **clear as mud**.*

(COCA; FIC: VirginiaQRev)

The simile (AS) CLEAR AS MUD means 'not clear at all' or 'difficult to understand'. It is a counterpart to similes such as (AS) CLEAR AS DAY, (AS) CLEAR AS CRYSTAL or (AS) CLEAR AS A BELL. The dissonance between *clear* and *mud* causes a humorous effect.[69] Typical targets of this simile are *explanations, speech* (or *language* in general) or *a visual experience*. The first recorded use of this simile dates back to 1796.[70]

#23 (AS) BLACK AS NIGHT

*He likes his eggs over-easy, and his coffee **black as night**.*

(EW15; 205440)

The simile (AS) BLACK AS NIGHT means 'completely black' or 'without light'. Generally, the target can be either something of black colour, such as *hair* or *sky*, or something without light or lighting, for instance, *street* or *room*. Many other similes contain the *black* ground: (AS) BLACK AS COAL, (AS) BLACK AS PITCH, (AS) BLACK AS JET, (AS) BLACK AS MIDNIGHT, (AS) BLACK AS THE ACE OF SPADES, (AS) BLACK AS THUNDER, (AS) BLACK AS A RAVEN or (AS) BLACK AS INK. The last three are also used with the meaning 'evil or mean-spirited'. Interestingly, the most frequent source *night* is a metonymy, while other sources (*pitch, coal, jet*) are concrete material objects. One possible reason is that *night* remains a relevant concept while the other sources have become obsolete.

The simile also occurs in the grammatical transformation (AS) BLACK AS THE NIGHT. This transformation is sometimes pragmatically conditioned, where the source *night* is contextually bound (as opposed to a generic reference).

---

[68] The word *quick* has many near-synonyms, such as *fast, hasty, speedy, swift, rapid*. Hence the number of possible lexical transformations.
[69] Referred to as *salience imbalance* by Norrick (1986).
[70] Harper, D. (n.d.). Etymology of mud. Online Etymology Dictionary. Retrieved November 25, 2022, from https://www.etymonline.com/word/mud

The source *night* also occurs in the simile (AS) DARK AS NIGHT and its transformations. This simile is often semantically and functionally identical to (AS) BLACK AS NIGHT.

#24 (AS) BRIGHT AS THE SUN

*Her eyes were **as bright as the sun**, her smile was endless!!!*

<div align="right">(EW15; 455679)</div>

The simile (AS) BRIGHT AS THE SUN means 'exceptionally bright' when referring to *light* or *reflection*. It also describes people and their emotions with the meaning 'positive, warm, welcoming', with targets such as *smile*, *love* or *person*. No major dictionary lists this simile despite its significant frequency compared to other, less frequent similes. One of the possible reasons is that the expression *as bright as the sun* occurs in astronomy texts, where it may be dismissed as a literal comparison. However, most instances of this simile in the researched corpora are from texts outside astronomy.

The source *sun* is also found in other strings, such as *(as) hot as the sun, (as) sure as the sun (comes up/rises/sets)* or *(as) clear as the sun.* These occur in the EW15 and are not classified as true similes in this work, as they are frequently literal comparisons. The string *(as) sure as the sun comes up* is listed in dictionaries as a set expression but is better interpreted as an adverbial idiom.

7.3 3ʳᵈ list (similes ranking #25-36)

Simile	EW15	COCA	BNC	SBNC2014	Total
OLD AS THE HILLS	347 (0.026)	65 (0.065)	15 (0.134)	0 (0)	**427**
HARD AS NAILS	361 (0.027)	42 (0.042)	13 (0.116)	7 (0.612)	**423**
THICK AS THIEVES	316 (0.024)	89 (0.089)	14 (0.125)	1 (0.088)	**419**
CLEAN AS A WHISTLE	262 (0.020)	144 (0.144)	7 (0.062)	1 (0.088)	**414**
HIGH AS A KITE	264 (0.020)	130 (0.130)	13 (0.116)	5 (0.438)	**412**
SHARP AS A TACK	322 (0.024)	82 (0.082)	5 (0.045)	0 (0)	**409**
WISE AS SERPENTS	377 (0.029)	27 (0.027)	1 (0.009)	0 (0)	**405**
SMOOTH AS GLASS	323 (0.024)	39 (0.039)	6 (0.054)	0 (0)	**368**
NATURAL AS BREATHING	316 (0.024)	39 (0.039)	8 (0.071)	0 (0)	**363**

SWEET AS HONEY	312 (0.024)	49 (0.049)	2 (0.018)	0 (0)	**363**
HAPPY AS A CLAM	282 (0.021)	71 (0.071)	0 (0)	0 (0)	**353**
LIGHT AS AIR	287 (0.022)	58 (0.058)	2 (0.018)	0 (0)	**347**

Table 20. Frequencies of #25-36 similes in the corpora.

The third list (Table 20) of adjectival similes ranking from #25 to #36 includes items with an absolute frequency between *n=347* and *n=427*. It contains four similes that occur in all the researched corpora, including the most frequent simile in the Spoken BNC2014: (AS) HARD AS NAILS (0.612 i.p.m.). Only a single adjectival simile in this list is not attested in the BNC corpus, (AS) HAPPY AS A CLAM, often described as typically American. The similes (AS) CLEAN AS A WHISTLE (0.144 i.p.m.) and (AS) HIGH AS A KITE (0.130 i.p.m.) are particularly frequent in the COCA while somewhat less frequent in the EW15 (both 0.020 i.p.m.) compared to the other similes in this list.

The list contains two 'animal' similes, (AS) WISE AS SERPENTS and (AS) HAPPY AS A CLAM. Furthermore, four adjectival similes in this list have generally less frequent plural-noun sources.

#25 (AS) OLD AS THE HILLS

*Mortgage scams were **as old as the hills.***

(BNC; FAB 4047)

This simile generally means 'very old' or even 'ancient', often used humorously. The target of this simile is usually people's behaviour, their approach, and the furnishing or design of living spaces. A related simile (AS) OLD AS METHUSELAH is used exclusively to describe people of advanced age. The ground transformation in (AS) ANCIENT AS THE HILLS potentially makes the simile even more humorous and evaluative than the original form with *old*.

Many other similes are used with a similar meaning to imply timelessness without necessarily being humorous: (AS) OLD AS THE REPUBLIC (ITSELF), (AS) OLD AS TIME (ITSELF), (AS) OLD AS THE BIBLE (ITSELF), (AS) OLD AS MANKIND (ITSELF), (AS) OLD AS HUMANITY (ITSELF), (AS) OLD AS HISTORY (ITSELF), (AS) OLD AS THE EARTH (ITSELF), (AS) OLD AS AMERICA (ITSELF), (AS) OLD AS CIVILIZATION (ITSELF), (AS) OLD AS WAR (ITSELF), (AS) OLD AS THE DEVIL or (AS) OLD AS THE UNIVERSE (ITSELF). The context of

the utterance usually determines the source. For instance, *America* is likely to occur in the American variant, and *republic* tends to occur in the political discourse.

#26 (AS) HARD AS NAILS

*Young Keith looks a great prospect, a fast and astute runner with the ball, **hard as nails** in driving play, a good kicker and passer.*

(BNC; CB2 121)

The simile (AS) HARD AS NAILS occurs in all the researched corpora. It conveys the meaning 'physically or mentally tough', 'determined', or even 'showing little sympathy'. The target of (AS) HARD AS NAILS is almost exclusively a person or their behaviour and attitude. It is closely related to the simile (AS) TOUGH AS NAILS, which, according to the researched data, is more common in the American variant.

Many other similes with the ground *hard* can mean 'physically tough', such as (AS) HARD AS STONE, (AS) HARD AS IRON, (AS) HARD AS GRANITE, (AS) HARD AS DIAMOND, (AS) HARD AS ROCK, (AS) HARD AS A ROCK or (AS) HARD AS CONCRETE. However, these tend to differ in the connotative meanings.

According to The American Heritage Dictionary, (AS) HARD AS NAILS is a variant of the original simile (AS) HARD AS FLINT STONE.[71]

#27 (AS) THICK AS THIEVES

*I suspect she and her mother are really **as thick as thieves**, eh?*

(EW15; 1032641)

The simile (AS) THICK AS THIEVES occurs in all the researched corpora. Its meaning is 'close, allied, intimate', with the target being people and their relationship or partnership. The source *thieves* affects the connotations of the simile and insinuates suspiciousness. The meaning of *thick* ('intimate') is now obsolete outside this simile.[72] The idiom *to be at daggers drawn* is sometimes listed as the opposite – an observation doubtless reinforced by the association of *dagger* with *thief*.

---

[71] hard as nails. (n.d.) The American Heritage Dictionary of Idioms by Christine Ammer. (2003, 1997). Retrieved February 21, 2022, from https://idioms.thefreedictionary.com/hard+as+nails
[72] Thick as Thieves. (n.d.) The Dictionary of Clichés. (2013). Retrieved January 1 2023 from https://idioms.thefreedictionary.com/thick+as+thieves

Two American films and a British sitcom named *Thick as Thieves* possibly contribute to the survival of this simile. Moreover, the simile is alliterating.

#28 (AS) CLEAN AS A WHISTLE

*(…) just a bit holier than thou at the moment you know **clean as a whistle***

<div align="right">(Spoken BNC2014; SUPV 942)</div>

The simile (AS) CLEAN AS A WHISTLE occurs in all the researched corpora and is used with several meanings. Firstly, it can mean 'clean from dirt', as in *I want the car clean as a whistle by tomorrow*, with typical targets being *clothes, car, room* or *house*. Secondly, it is used to describe people with the meaning 'neat in appearance'. Thirdly, the simile describes people's behaviour with the meaning 'not involved in anything suspicious or illegal'. Lastly, the simile may mean 'entirely', such as in *he tore down the wall, clean as a whistle.* In these cases, the simile functions as an adverbial of manner.

Historically, the simile is reported to have had a different form (AS) CLEAR AS A WHISTLE in the 18th century, possibly referring to the clarity of the sound.[73] This would make the simile functionally related to (AS) CLEAR AS A BELL at that time.

#29 (AS) HIGH AS A KITE

*(…) just been like popping them over four hours and I were **high as a kite***

<div align="right">(Spoken BNC2014; SMC2 618)</div>

This simile (AS) HIGH AS A KITE means 'very high' in terms of altitude, 'very happy or excited' or '(very) drunk or drugged'. Another possible meaning of this simile typically absent from dictionaries is 'very active or energetic', as in *the [team] were flying high as a kite*. Predominantly, it is used to describe intoxication or other forms of substance-induced states, with the other meanings being marginal. The target is almost exclusively a person or a group of people. The slang meaning of *high* ('intoxicated') dates back to the 1620s.[74]

---

[73] clean as a whistle. (n.d.) The American Heritage Dictionary of Idioms by Christine Ammer. (2003, 1997). Retrieved January 31 2023 from https://idioms.thefreedictionary.com/clean+as+a+whistle
[74] Harper, D. (n.d.). Etymology of high. Online Etymology Dictionary. Retrieved February 22, 2022, from https://www.etymonline.com/word/high

The simile (AS) HIGH AS THE SKY is perhaps the only related simile, and it shares all the meanings and functions with (AS) HIGH AS A KITE, with intoxication being the most frequent.

Formally, both (AS) HIGH AS A KITE and (AS) HIGH AS THE SKY contain vocalic assonance /aɪ/.

#30 (AS) SHARP AS A TACK

*He is now over 90 years old and his mind is **as sharp as a tack**.*

(EW15; 619654)

The simile (AS) SHARP AS A TACK means 'mentally acute' or 'very intelligent'. The target is almost exclusively a person or their mind. Like its predecessor (AS) SHARP AS A NEEDLE[75], this simile does not describe material sharpness. Other similes with the same ground are (AS) SHARP AS A RAZOR, (AS) SHARP AS A KNIFE (BLADE) and (AS) SHARP AS A BLADE, which describe both people and material objects.

Some dictionaries consider (AS) SHARP AS A TACK an American variant of (AS) SHARP AS A RAZOR, but the data in the researched corpora do not confirm that. Moreover, the sharpness is conceptualised differently in each simile. A pointy tip (*tack*) evokes jabbing or puncturing, whereas a sharp edge (*razor*) refers to cutting. The difference in conceptualisation results in (AS) SHARP AS A RAZOR also being used to describe material sharpness, broadening its range of functions. Consequently, the two similes must be treated as two independent units, both formally and semantically.

#31 (AS) WISE AS SERPENTS

*Or how about the delightful instruction to us to be "**wise as serpents**"?*

(EW15; 113559)

The simile (AS) WISE AS SERPENTS alludes to the ability of serpents to thrive in the world. The meaning of this simile is 'crafty and subtle', possibly also 'shrewd, cunning or slippery'. The target of this simile is almost exclusively a person.

---

[75] sharp as a tack. (n.d.) The Dictionary of Clichés. (2013). Retrieved December 28 2022 from https://idioms.thefreedictionary.com/sharp+as+a+tack

The simile allegedly originates in the Bible, followed by another simile (AS) HARMLESS AS DOVES.[76] While sometimes used as one phraseological unit in the string *wise as serpents, and harmless as doves*, (AS) WISE AS SERPENTS is significantly more frequent in the EW15 (*n*=377; 0.029 i.p.m.) than (AS) HARMLESS AS DOVES (*n*=162; 0.012).

This simile is strongly associated with religious discourse, which explains its absence in the BNC-type corpora. Furthermore, no major dictionary lists this simile. The simile also has a grammatical variant (AS) WISE AS A SERPENT, which is much less frequent in the English Web (*n*=56). In the COCA, only four instances come from spoken language source texts, all of which are films (i.e. scripted dialogues). Since the EW15 only contains written sources, it is reasonable to assume that the simile is somewhat exclusive to literary language or its oral adaptations.

Formally, the simile is alliterating the sibilants /z/ and /s/. Consequently, the extra /s/ in the plural form may contribute to the prevalence of the source *serpents*. This simile may also be considered partly onomatopoeic, as the repetition of /s/ corresponds to the hissing sound of serpents. However, the prosodic features might be less significant, as illustrated by the lack of spoken text evidence.

#32 (AS) SMOOTH AS GLASS

*An ocean **as smooth as glass** closing over things vast, alive and hateful.*

(BNC; CH0 3298)

This simile is used with two meanings. The first meaning is 'still or tranquil' and typically takes *water* as a target. The second meaning, 'very smooth', usually describes a solid surface that is exceptionally smooth and possibly slippery, often due to polishing. Typical targets are *black ice, floor* or any material. It can also describe human skin and animal fur, emphasising smoothness and shininess.

A related simile sharing the ground *smooth* is (AS) SMOOTH AS SILK. It is likewise used to describe the smoothness of the surface, but it also frequently means 'without obstacles or hindrance', as in *It went smooth as silk.* As indicated by the source, the connotations of the two similes differ. The smoothness in (AS) SMOOTH AS SILK typically

---

[76] The Gospel of Matthew in the New Testament (Matthew 10:16, p.1825).

results from the perfection of the material, while (AS) SMOOTH AS GLASS indicates polishing and shininess.

The meanings 'very smooth' and 'without obstacles' are also expressed by the simile (AS) SMOOTH AS BUTTER, but this simile is only attested in the EW15 ($n=223$; 0.017 i.p.m.).

#33 (AS) NATURAL AS BREATHING

*A few years ago, using credit was **as natural as breathing** for many consumers.*

(EW15; 1033034)

The meaning of this simile is 'very natural or innate'. It typically describes habits or skills (*swimming, singing, playing the guitar*) that have become automatic and require little conscious effort. The target is almost exclusively a person or an activity performed by people.

Despite being relatively frequent, the simile does not occur in any dictionary, possibly due to being interpreted as a literal comparison. This can be attributed to two oddities. Firstly, the source is morphologically unusual, as gerunds rarely function as sources in similes. Human activities generally do not occur as grounds in standard similes. Secondly, the ground contains three syllables, which is infrequent.

Overall, the reversibility test suggests that (AS) NATURAL AS BREATHING is a simile, and its usage further confirms that.

#34 (AS) SWEET AS HONEY

*(…) he said with a sanctimoniously false smile, his voice **as sweet as honey** laced with venom.*

(EW15; 4407170)

The simile (AS) SWEET AS HONEY means 'very sweet', 'friendly' or 'charming or kind'. The target is almost exclusively a person or their feature, such as *smile*, *lips* (kissing), *voice* or *words*. Literal sweetness is rare.

A related simile (AS) SWEET AS HONEY is also reasonably frequent in the COCA ($n=40$) and the EW15 ($n=136$) and has the same meaning and function. Another related

simile is (AS) SWEET AS PIE, which is sometimes listed as chiefly British[77], but the distribution of the simile in the data does not support that. Lastly, (AS) SWEET AS CANDY can also be considered identical in meaning and function. However, this simile is not listed in any major dictionary.

#35 (AS) HAPPY AS A CLAM

*She's just **happy as a clam** as long as she has some item not intended for babies in her possession.*

(COCA; WEB: windtraveler.net)

This simile means 'very happy'. It possibly refers to clams' safety (and therefore happiness) from predators during high tide.[78] The simile has a grammatical variant (AS) HAPPY AS CLAMS occurring in the EW15 (*n*=75). The simile is considered chiefly American and is not attested in the BNC-type corpora.[79]

The roots of (AS) HAPPY AS A CLAM date back to the 19th century. Bartlett (1848: 81-82) lists *as happy as a clam at high water* as "a very common expression in those parts of the coast of New England where clams are found." However, the expanded source is rare in present-day English, as evidenced by the COCA: only 3 out of the 71 instances include *at high tide*, and only one contains *in high water*. Dictionaries often list (AS) HAPPY AS LARRY and (AS) HAPPY AS A LARK as related similes with identical meanings and functions, with the only difference being a regional preference and possibly style.

Formally, the simile is assonantal.

#36 (AS) LIGHT AS AIR

*As I was turning to look, I felt static electricity so bad my skin started to itch like crazy and I felt naked like my clothing was **light as air** or completely being repelled by static.*

(EW15; 877340)

The simile (AS) LIGHT AS AIR has several meanings. Firstly, it can mean 'very light in weight', typically ascribed to material targets, such as *clothes* or *sunglasses*. It can also

---

[77] sweet as pie. (n.d.) Collins COBUILD Idioms Dictionary, 3rd ed.. (2012). Retrieved January 17 2023 from https://idioms.thefreedictionary.com/sweet+as+pie

[78] happy as a clam. (n.d.) Farlex Trivia Dictionary. (2011). Retrieved January 20 2023 from https://www.thefreedictionary.com/happy+as+a+clam

[79] This is further supported by the Cambridge Dictionary not listing this simile.

describe lightness of food, often used for various meals or desserts. Lastly, it can represent figurative lightness, meaning 'free from stress or difficulty'. This meaning is typically tied to person or body part targets, especially *head*.

A simile sharing the ground *light*, (AS) LIGHT AS A FEATHER, appears in the researched corpora more than twice as frequently. Its meaning and functions overlap with (AS) LIGHT AS AIR. Perhaps (AS) LIGHT AS AIR is a more recent simile, as is suggested by the source *air*, which is arguably more relatable than *feather* in today's world. Furthermore, (AS) LIGHT AS AIR does not have a CAS transformation, possibly indicating it has not been in the language for long.

The simile (AS) LIGHT AS AIR also has a transformation *lighter than air* which is generally more frequent and dates back to 1887.[80] The transformation may well be the original form that gave rise to the standard simile form.

7.4 4th list (similes ranking #37-48)

Simile	EW15	COCA	BNC	SBNC2014	Total
SICK AS A DOG	263 (0.020)	79 (0.079)	3 (0.027)	0 (0)	345
HARD AS STEEL	371 (0.024)	23 (0.023)	4 (0.036)	0 (0)	344
STIFF AS A BOARD	260 (0.020)	80 (0.080)	4 (0.036)	0 (0)	344
WHITE AS A SHEET	217 (0.016)	95 (0.095)	22 (0.196)	0 (0)	334
RED AS BLOOD	278 (0.021)	49 (0.049)	6 (0.054)	0 (0)	333
FLAT AS A PANCAKE	265 (0.020)	59 (0.059)	4 (0.036)	1 (0.088)	329
CHEAP AS CHIPS	318 (0.024)	6 (0.006)	0 (0)	4 (0.350)	328
OLD AS HUMANITY	282 (0.021)	21 (0.021)	0 (0)	0 (0)	303
QUIET AS A MOUSE	241 (0.018)	54 (0.054)	3 (0.027)	2 (0.175)	300
DRY AS A BONE	221 (0.017)	59 (0.059)	9 (0.080)	0 (0)	289
FAST AS LIGHTNING	267 (0.020)	21 (0.021)	0 (0)	0 (0)	288
DARK AS NIGHT	230 (0.017)	42 (0.042)	7 (0.062)	0 (0)	279

Table 21. Frequencies of #37-48 similes in the corpora.

---

[80] Harper, D. (n.d.). Etymology of light. Online Etymology Dictionary. Retrieved October 22, 2022, from https://www.etymonline.com/word/light

The list of adjectival similes ranking from #37 to #48 (Table 21) contains similes with an absolute frequency between *n=*279 and *n=*345. It includes only two similes attested in all four corpora: (AS) FLAT AS A PANCAKE and (AS) QUIET AS A MOUSE.

The list contains two 'animal' similes, (AS) SICK AS A DOG and (AS) QUIET AS A MOUSE, and three 'colour' similes, (AS) WHITE AS A SHEET, (AS) RED AS BLOOD and (AS) DARK AS NIGHT. Two similes do not occur in the BNC-type corpora: (AS) OLD AS HUMANITY and (AS) FAST AS LIGHTNING.

#37 (AS) SICK AS A DOG

*He was also **sick as a dog** through a lot of filming but it doesn't show.*

(EW15; 912690)

The simile (AS) SICK AS A DOG means 'very ill' or 'nauseated'. The first recorded use of this simile dates back to 1705.[81] A reasonable and convincing explanation for the source *dog* is lacking. The simile often suggests sickness related to stomach issues. The target of this simile is almost exclusively a person.

Other similes using the same ground include (AS) SICK AS A PIG and (AS) SICK AS A PARROT. However, their meaning differs significantly, as they mean 'annoyed' or 'displeased'. Both are regarded chiefly British, and the researched data corroborate that. (AS) SICK AS A PARROT occurs eight times in the BNC corpora, while only twice in the COCA and 29 times in the EW15.

#38 (AS) HARD AS STEEL

*His voice was **as hard as steel**.*

(EW15; 397151)

The simile (AS) HARD AS STEEL means 'very hard' or 'emotionless or hostile'. When the simile describes material toughness, the targets are usually durable items, such as *armour, sword* or *knife*, but also *(cold) water*. However, it is usually used for people, their body parts (*eyes*, *biceps, spine*) or even *voice*. This simile is often encountered in fiction, namely the fantasy subgenre. Other similes describing material toughness are (AS) HARD AS NAILS, (AS) HARD AS IRON, (AS) HARD AS GRANITE, (AS) HARD AS DIAMOND, (AS)

---

[81] sick as a dog. (n.d.) The American Heritage Dictionary of Idioms by Christine Ammer. (2003, 1997). Retrieved January 27 2023 from https://idioms.thefreedictionary.com/sick+as+a+dog

HARD AS ROCK, (AS) HARD AS A ROCK, (AS) HARD AS STONE or (AS) HARD AS CONCRETE. Many of these are used to describe people as well.

The source *steel* is also found in other similes. (AS) STRONG AS STEEL is closely related to (AS) HARD AS STEEL, but it frequently describes non-person targets (*relationship, wood, wire*). The simile (AS) TRUE AS STEEL means 'loyal or dependable' or 'completely true, correct', possibly referring to the reliability of weapons made of steel.

Lastly, (AS) HARD AS STEEL is sometimes expanded by stacking grounds, such as *hard and cold as steel* or *hard and sharp as steel*. Such instances might be treated as independent similes, but they remain marginal.

#39 (AS) STIFF AS A BOARD

*I'm **as stiff as a board**, climb out of the cab like an old man.*

(COCA; FIC: The Antioch Review)

The simile (AS) STIFF AS A BOARD has multiple meanings. The first meaning is 'very straight, inflexible' and usually describes human posture or motion, often suggesting clumsiness but also nervousness. The second meaning, 'rigid in behaviour', typically describes stubborn or uptight people. The third meaning is 'motionless and lifeless', which is often ascribed to dead bodies – human or animal. Interestingly, the third meaning is not listed in dictionaries but seems to be very frequent in the researched corpora. Functionally and semantically related similes are (AS) STIFF AS A POKER, (AS) STIFF AS A RAMROD and AS STIFF AS A STAKE. However, the researched data, as well as several dictionaries, suggest they are all obsolete.

This simile (AS) STIFF AS A BOARD often occurs alongside another simile (AS) LIGHT AS A FEATHER. The string *light as a feather, stiff as a board* commonly refers to a children's slumber party game or a levitation trick known from popular culture.[82]

#40 (AS) WHITE AS A SHEET

*'Goodbye!' he said in a low, hollow voice, his face **as white as a sheet**.*

(BNC; FR6 2278)

---

[82] https://en.wikipedia.org/wiki/Light_as_a_feather%2c_stiff_as_a_board

This simile means 'extremely pale', frequently alluding to illness or shock. Consequently, the target is typically a person. If we consider the grammatical transformation *white as sheets* in the Spoken BNC2014 (*n*=1), the simile occurs in all the researched corpora.

Functionally related similes are (AS) WHITE AS A GHOST, (AS) PALE AS A GHOST and (AS) PALE AS DEATH, all of which express paleness caused by negative circumstances. The relative frequency suggests that (AS) WHITE AS A SHEET is (or at least used to be) more common in British English.

#41 (AS) RED AS BLOOD

*Its eyes were **red as blood** and seemed to pierce you when they looked at you, seeing into your mind.*

(EW15; 42883)

The simile (AS) RED AS BLOOD means 'very red' or 'flushed'. The target is mostly a person, often a newborn child, or specific body parts such as *face, lips* or *eyes* (typically in fantasy fiction). Non-person targets usually include *sun* or various types of clothes.

Another reasonably frequent simile with the same ground is (AS) RED AS A TOMATO, occurring in the EW15 (*n*=56) and the COCA (*n*=4). Other related similes, such as (AS) RED AS A BEET, (AS) RED AS A BEETROOT[83], (AS) RED AS A POPPY, (AS) RED AS A RUBY, (AS) RED AS A CHERRY or (AS) RED AS A ROSE, share the meaning and function but their frequency in the researched data suggests that these are either obsolete or marginal.

This simile has a CAS transformation BLOOD(-)RED, which is more common than the simile form.[84] In the BNC, the CAS form (*n*=27 open; *n*=70 hyphenated) overwhelmingly exceeds the simile form (*n*=6). The COCA provides a similar picture, with the CAS form (*n*=314 open; *n*=456 hyphenated) being significantly more frequent than the simile form (*n*=49). Lastly, the EW15 only reinforces the CAS form's prevalence (*n*=4244 open; *n*=2231 hyphenated) compared to the simile form (*n*=278).

The simile (AS) RED AS BLOOD might be considered alliterating, but the final /d/ in *blood* will often remain unreleased in everyday speech.

---

[83] Dictionaries often list that *beetroot* is typically British while *beet* tends to occur in American English.
[84] Colour similes tend to occur in the CAS form almost invariably compared to other similes, where the CAS form does not have to exist or remains marginal.

#42 (AS) FLAT AS A PANCAKE

*And, yeah, Rosie's character was **as flat as a pancake**.*

<div align="right">(COCA; BLOG moviecitynews.com)</div>

The simile (AS) FLAT AS A PANCAKE occurs in all the researched corpora and generally has three meanings. Firstly, it means 'completely flat, level' when describing a surface. The target is often *terrain*, *city* or *country*, and body parts, such as the *belly*. The simile is also used to describe something destroyed by falling (related to *flatten*), such as in *Brad's computer lay on the barren ground, flat as a pancake*. Secondly, it can be a rude slang euphemism for 'a woman with small breasts'. Thirdly, it can describe a lack of depth, usually said of a *story* or a *character*.

The researched data show only one related simile (AS) FLAT AS A BOARD occurring in the EW15 (*n=55*; 0.004 i.p.m.). The meaning and function are identical.

Formally, (AS) FLAT AS A PANCAKE is assonantal.

#43 (AS) CHEAP AS CHIPS

*then we were like we all like put our our stuff in his car and he took us there and he was it was **cheap as chips** but he was like he was driving like a nutter*

<div align="right">(Spoken BNC2014; S3LE 1734)</div>

This simile means 'very cheap', often said of bargains or inexpensive items or services. The target can be anything purchasable.

Dictionaries list this simile as primarily British, which is further supported by the data in the researched corpora. The simile is among the most frequent in the Spoken BNC2014 (*n=4*; 0.350 i.p.m.) and appears marginal in the COCA (*n=6*; 0.006 i.p.m.). Additionally, all the COCA instances are from 2012. The EW15 frequency (*n=318*; 0.024 i.p.m.) suggests that it is relatively popular in present-day English. Its absence from the original BNC is attributed to the relative novelty of the simile.[85] It is also used in the Australian variant, possibly due to the *Cheap as Chips* retailer based in Australia.

The simile is alliterating – the repetition of the initial /tʃ/ and root-final /p/ makes it easily memorable.

---

[85] Cresswell (2021: 570) notes that the origin of the phrase dates back to at least 1850s. However, it had not become widely used until the end of the 20th century.

#44 (AS) OLD AS HUMANITY

*The problem of unwanted pregnancy is **as old as humanity** -- and human beings have always been smart enough to want to find solutions to it.*

(COCA; BLOG, rhrealitycheck.org)

The simile (AS) OLD AS HUMANITY describes something 'timeless or eternal'. It is usually a hyperbole and may be used humorously. Unlike the similar (AS) OLD AS TIME, none of the major dictionaries lists (AS) OLD AS HUMANITY.

The simile often occurs with an expanded source, as illustrated by *The search for the meaning of the afterlife is as old as humanity itself*, which is typical of this simile type. The component *itself* functions as an emphasiser. In the COCA, the simile occurs with *itself* in 9 out of 21 instances (42.86%). The target is usually a general concept or issue within our culture that is considered timeless, such as *the afterlife, religion* and many other (often philosophical) questions. Related similes with overlapping meaning and function are (AS) OLD AS TIME, (AS) OLD AS MANKIND or (AS) OLD AS CIVILIZATION.

#45 (AS) QUIET AS A MOUSE

*She'll be **quiet as a mouse**, won't cause no trouble to anything.*

(COCA; FIC LitCavalcade)

The simile (AS) QUIET AS A MOUSE occurs in all the researched corpora and has three primary meanings. The first meaning is 'very quiet or silent', usually with a person target. The second meaning is 'meek or gentle', mostly used positively to describe a humble, tolerant, and patient person who does not impose on anyone. The third meaning is 'submissive, easily manipulated', typically ascribed to a person without enough confidence or courage to exert their own will.

Related similes, including the ground *quiet*, are (AS) QUIET AS A LAMB and (AS) QUIET AS A/THE GRAVE. The similes (AS) SILENT AS THE GRAVE and (AS) SILENT AS THE DEAD are also similar in meaning. However, the connotations differ mostly due to the sources. *Lamb* implies submissiveness, whereas *grave* and *the dead* have ominous connotations.

Another related simile, (AS) QUIET AS A CHURCH MOUSE, is a blend of (AS) QUIET AS A MOUSE and (AS) POOR AS A CHURCH MOUSE. It only occurs in the COCA (*n*=14) and

the EW15 (*n*=32), suggesting that it might be more common in the American variant, as evidenced by the entry in the *Dictionary of American Idioms and Phrasal Verbs*.[86]

#46 **(AS) DRY AS A BONE**

*Her mouth felt **as dry as a bone** and her eyes were closed against the intrusive light.*

(BNC; FAB 1537)

The simile (AS) DRY AS A BONE is used with three primary meanings. Firstly, it can mean 'extremely dry', referring to the absence of water or humidity. The targets are typically *air, pond* and other material objects and places, especially those typically containing water. Secondly, the meaning can be 'very thirsty', primarily ascribed to a person target. Thirdly, the simile possibly means 'straight-faced or unamusing' when said of a particular person's (sense of) humour.

The simile has a CAS transformation, BONE(-)DRY, which is significantly more frequent than the simile form, as also listed by some dictionaries.[87] This is further corroborated by the data. In the COCA, the CAS forms (*n*=151 open; *n*=199 hyphenated) are overwhelmingly more frequent than the simile form (*n*=59). EW15 offers a similar picture, with the CAS forms (*n*=1194 open; *n*=716 hyphenated) also occurring more frequently than the simile form (*n*=221).

#47 **(AS) FAST AS LIGHTNING**

*If you are a fan of, well, kung fu fighting, then yes, these cats are **as fast as lightning**.*

(EW15; 495739)

The simile (AS) FAST AS LIGHTNING means 'very quick or speedy'. The target is usually some motion, especially when describing fighting (*kicks, punches*), but also a person.

Several other similes are nearly identical in meaning and function. The ground *fast* occurs in the simile (AS) FAST AS THE WIND; other related similes are (AS) SWIFT AS AN ARROW or (AS) SWIFT AS THE WIND. The source *lightning* appears in the simile (AS)

---

[86] quiet as a mouse and *quiet as the grave. (n.d.) McGraw-Hill Dictionary of American Idioms and Phrasal Verbs. (2002). Retrieved February 1 2023 from
https://idioms.thefreedictionary.com/quiet+as+a+mouse+and+*quiet+as+the+grave
[87] dry as a bone. (n.d.) Collins COBUILD Idioms Dictionary, 3rd ed. (2012). Retrieved April 25 2023 from
https://idioms.thefreedictionary.com/dry+as+a+bone

QUICK AS LIGHTNING, which is synonymous with (AS) FAST AS LIGHTNING due to the notorious synonymy of *fast* and *quick*.

The simile (AS) FAST AS LIGHTNING also has CAS transformation LIGHTNING(-)FAST, which by far exceeds the frequency of the simile form. In the BNC, this simile only appears in the CAS form ($n=5$ hyphenated; $n=3$ open). In the COCA, the CAS form ($n=177$ hyphenated; $n=113$ open) is more frequent than the simile form ($n=21$). Similarly, the CAS ($n=2059$ hyphenated; $n=3415$ open) is overwhelmingly more frequent than the simile form ($n=267$).

Lastly, (AS) FAST AS LIGHTNING sometimes functions as an adverbial phrase, such as in *Fast as lightning, she ducked and got out of harm's way*. Since *fast* can be either an adjective or an adverb, simple adjective queries may not return all the instances of this simile. As the data show, *fast* is sometimes tagged as an adverb even in clearly adjectival uses, making the annotation unreliable.

#48 (AS) DARK AS NIGHT

*Kafar looked up at a coal-black face distorted by rage, eyes **dark as night**, red flames burning deep within.*

(COCA; FIC: FantasySciFi)

This simile means 'very dark', usually in association with the black colour. Typical targets are human parts of body, such as *eyes* and *hair*, and also *room* or *outside*.

The simile (AS) DARK AS THE NIGHT SKY can be considered an expanded variant. Another interpretation is to consider (AS) DARK AS NIGHT a condensed variant (a metonymy). The two similes are fairly similar in meaning; however, they differ functionally, with (AS) DARK AS THE NIGHT SKY being rather literary. The variant (AS) DARK AS THE NIGHT is also attested in the researched data ($n=32$ in the EW15; $n=3$ in the COCA). Interestingly, none of the similes occur in dictionaries.

Another related simile sharing the source *night* is (AS) BLACK AS NIGHT. It is similar to (AS) DARK AS NIGHT in both meaning and function.

7.5 5th list (similes ranking #49-60)

Simile	EW15	COCA	BNC	S. BNC2014	Total
HARD AS STONE	236 (0.018)	38 (0.038)	2 (0.018)	0 (0)	**276**

STRONG AS STEEL	256 (0.019)	17 (0.017)	2 (0.018)	0 (0)	**275**
DEAD AS A DOORNAIL	185 (0.014)	79 (0.079)	5 (0.045)	0 (0)	**269**
BRIGHT AS DAY	211 (0.016)	49 (0.049)	6 (0.054)	0 (0)	**266**
SMART AS A WHIP	182 (0.014)	81 (0.081)	1 (0.009)	0 (0)	**264**
STRONG AS AN OX	191 (0.014)	59 (0.059)	6 (0.054)	1 (0.088)	**257**
BLIND AS A BAT	167 (0.013)	82 (0.082)	4 (0.036)	3 (0.263)	**256**
MAD AS A HATTER	206 (0.016)	42 (0.042)	6 (0.054)	1 (0.088)	**255**
PRETTY AS A PICTURE	185 (0.014)	56 (0.056)	14 (0.125)	0 (0)	**255**
SAFE AS HOUSES	216 (0.016)	20 (0.020)	16 (0.143)	0 (0)	**252**
FRESH AS A DAISY	202 (0.015)	37 (0.037)	5 (0.045)	3 (0.263)	**247**
HARD AS ROCK	207 (0.016)	30 (0.030)	2 (0.018)	0 (0)	**239**

Table 22. Frequencies of #49-60 similes in the corpora.

The last simile list (Tabel 22) ranking adjectival similes from #49 to #60 comprises similes with frequencies between *n=*239 and *n=*276. It is the only list where every single simile is attested in the BNC. The simile (AS) BLIND AS A BAT is the only one in the top 60 with a relative frequency under 0.014 i.p.m. in the EW15, but its relative frequencies in the other corpora make up for it by being fairly high.

Two similes in this list have an animal source, (AS) STRONG AS AN OX and (AS) BLIND AS A BAT. (AS) STRONG AS STEEL ranks #50 due to its frequency in the EW15, despite not being attested in the Spoken BNC2014 and occurring relatively infrequently in both the BNC and the COCA.

#49 (AS) HARD AS STONE

*Part of getting good at various circus skills is learning how to hyperextend your shoulders, arch your back, and make your abs **hard as stones**.*

(EW15; 3359356)

The simile (AS) HARD AS STONE has two primary meanings. The first meaning is 'very hard', used to refer to the physical toughness of material objects or stiff body parts. The target is commonly stale or dry food, *neck*, *back* or *muscles*. Occasionally, the simile is used to describe sexual arousal, with various terms for *penis* and *nipples* as targets. The second meaning, 'uncompromising or unflinching', describes people, their behaviour or their body parts (as metonymies for behaviour). Apart from a general reference to a

112

person, typical targets are *face, look* or *voice*. Interestingly, dictionaries only list the first meaning referring to material toughness, while the researched data suggest that (AS) HARD AS STONE is more frequently used to describe people rather than objects.

Other similes with *hard* as their ground include (AS) HARD AS A ROCK, (AS) HARD AS ROCK, (AS) HARD AS STEEL, (AS) HARD AS NAILS, (AS) HARD AS CONCRETE, (AS) HARD AS GRANITE or (AS) HARD AS DIAMOND. These similes may overlap in the meaning 'physically hard', but their connotations may significantly differ when used with person targets.

#50 (AS) STRONG AS STEEL

*It is the lubricant that ensures that our inevitable family entanglement will be composed of silken threads **as strong as steel**.*

(EW15; 147071)

This simile means 'durable or enduring' or 'stable'. It is closely related to the simile (AS) HARD AS STEEL. However, (AS) STRONG AS STEEL frequently describes non-person targets that can be abstract (*relationship, loyalty*) or concrete (*wood* or *wire*).

Other similes with the ground *strong* include (AS) STRONG AS IRON, (AS) STRONG AS AN OAK, (AS) STRONG AS A BULL, (AS) STRONG AS A HORSE, (AS) STRONG AS A LION, (AS) STRONG AS A BEAR and (AS) STRONG AS AN OX. The similes with animal sources may have identical uses and meanings, but their connotations usually differ from those with a material source.

Formally, (AS) STRONG AS STEEL is alliterating.

#51 (AS) DEAD AS A DOORNAIL

*Like this sea without sun, he wrote, without fish, without birds, **dead as a doornail** despite the goddamned swell that tosses the boat, wearing the sails out and wearing me down.*

(COCA; FIC: KenyonRev)

This simile is used with two meanings. The first meaning is 'dead beyond doubt,' which is typically ascribed to people or animal targets. The second meaning is 'no longer popular', usually describing forms of entertainment, musical and other genres of art, or ways of performing various activities. The component *doornail* is an example of an

incongruent source, as doornails are inanimate objects, putting extra emphasis on the ground *dead*.

Other similes containing the ground *dead* are (AS) DEAD AS A DODO and (AS) DEAD AS MUTTON. While also meaning either 'dead beyond doubt' or 'no longer popular', the similes have different connotations triggered by the sources. *Dodo* may allude to extinction, and *mutton* possibly refers to the act of slaughtering an animal for meat. Originally, both sources are animate entities, as opposed to *doornail*.

Formally, (AS) DEAD AS A DOORNAIL is alliterating.

#52 (AS) BRIGHT AS DAY

The yard was bright as day for a few seconds, then night crept back in around the flames.

(BNC; GVL 3523)

This simile is used almost exclusively with the meaning 'very bright, illuminated'. The target of (AS) BRIGHT AS DAY is typically an object or area that can be lit by artificial light or sunlight. It is also used to describe a person's *smile*.

Interestingly, the simile is not listed in any dictionary despite being attested in the researched corpora. One of the possible explanations for its absence from dictionaries might be that (AS) BRIGHT AS DAY is treated merely as a less frequent variant of (AS) CLEAR AS DAY. The latter is the most frequent simile in the dataset and shares the same source (*day*), but it is otherwise semantically and functionally unrelated.

#53 (AS) SMART AS A WHIP

He's smart as a whip, passionate, speaks the hard, uncomfortable truth and has called this disciple to deeper faithfulness.

(EW15; 1833323)

The simile (AS) SMART AS A WHIP means 'very intelligent' or 'quick-thinking'. The interaction between the ground *smart* and the source *whip* may seem incongruous in present-day English. However, one of the older meanings of *smart* is 'fast'[88], which is a salient feature of *whip*.

---

[88] Harper, D. (n.d.). Etymology of smart. Online Etymology Dictionary. Retrieved June 9, 2023, from https://www.etymonline.com/word/smart

Two synonymous similes contain the ground *sharp*: (AS) SHARP AS A RAZOR and (AS) SHARP AS A TACK. While the former can also refer to material sharpness, the latter is used exclusively with the meaning 'very intelligent' or 'quick-thinking'.

#54 (AS) STRONG AS AN OX

The voice belonged to the 'Trog', a short, thick-set man who resembled a character of the Stone Age and was as strong as an ox.

<div align="right">(BNC; A61 846)</div>

The simile (AS) STRONG AS AN OX occurs in all the researched corpora. Its meaning is 'extremely strong', exclusively referring to physical strength. The target of this simile is predominantly a person.

Some other similes contain the ground *strong*, typically in combination with an 'animal' source: (AS) STRONG AS A BULL, (AS) STRONG AS A LION, (AS) STRONG AS A BEAR and (AS) STRONG AS A HORSE. Additionally, the ground *strong* occurs with a 'material' source in (AS) STRONG AS STEEL and (AS) STRONG AS IRON. The animal and material similes can be considered synonymous with (AS) STRONG AS AN OX, but the latter's connotations might differ. Lastly, the ground *strong* occurs in (AS) STRONG AS AN OAK.

Formally, (AS) STRONG AS AN OX is assonantal.

#55 (AS) BLIND AS A BAT

*Everyone knows I am **blind as a bat** and really a pretty terrible driver, this being the case Kiel was in the driver seat I was ridding shotgun we were burning spliff after spliff we were Canada bound.*

<div align="right">(COCA; WEB: planet.infowars.com)</div>

This simile is used with three related meanings. Firstly, it can mean 'unable to see', often used as a hyperbole to describe a person's inability to find something. Secondly, it is used with the meaning 'having poor vision' and mostly with people targets. This meaning is more in accordance with the limited vision of bats. Thirdly, it can mean 'oblivious to something', utilising the figurative meaning of *blind*. The target is typically a person unaware of an issue or appropriate behaviour in a social situation.

Both constituting elements are unique to the (AS) BLIND AS A BAT and do not occur in any other standard adjectival similes.

Formally, (AS) BLIND AS A BAT is alliterating.

#56 (AS) MAD AS A HATTER

*From what I can gather he was **as mad as a hatter**, and really no good at all.*

<div align="right">(BNC; BN6 66)</div>

The simile (AS) MAD AS A HATTER is used with two meanings. The first and original meaning is 'crazy or deranged', alluding to the hatter profession and the heightened exposure to the chemicals they worked with, causing mental issues. Lewis Carroll's Alice in Wonderland is sometimes wrongly cited as the source of the simile, perhaps because of the story's popularity. The second meaning, 'very angry or cross', results from misunderstanding the simile based on the American use of the word *mad*, meaning almost exclusively *angry*.[89] With either of the two meanings, the target of the simile is almost exclusively a person or their behaviour.

The simile (AS) MAD AS A MARCH HARE is often listed in dictionaries as synonymous[90], but it is now reported as obsolete, which agrees with the data in the BNC (*n=2*) and the COCA (*n=6*, five date back to the 1990s). However, the EW15 returns 43 instances of the simile, suggesting that it may still be in use.

Formally, the simile is assonantal.

#57 (AS) PRETTY AS A PICTURE

*Well, we also know that Kate, the Duchess of Cambridge is **as pretty as a picture**.*

<div align="right">(COCA; SPOK: NBC Today Show)</div>

The simile (AS) PRETTY AS A PICTURE means 'very pretty' or otherwise 'visually appealing'. Typical targets include people, cities or towns, and landscapes. The simile alludes to the idea that pictures are beautiful, and people display them for admiration.

The similes (AS) PRETTY AS A PEACH and (AS) PRETTY AS A SPECKLED PUP are sometimes listed as synonymous, albeit perhaps with slightly different connotations. Similarly, (AS) CUTE AS A BUG and (AS) CUTE AS A BUG'S EAR are related, despite their meaning being 'adorable' rather than 'attractive because of beauty'.

Formally, the simile is alliterating.

---

[89] The Cambridge Dictionary does not list this meaning at all.
[90] mad as a hatter. (n.d.) Farlex Partner Idioms Dictionary. (2017). Retrieved March 6 2023 from https://idioms.thefreedictionary.com/mad+as+a+hatter

#58 (AS) SAFE AS HOUSES

*'Saul could have reported us, but now he's six feet under, so we're **safe as houses**,' concluded Josh.*

(BNC; FPM 319)

The simile (AS) SAFE AS HOUSES means 'very safe' or 'secure'. It usually appears in business or finance texts, often describing investment, money or property, but also job security. As the example illustrates, the simile is also used to mean 'safe from harm or unpleasant situation'.

The simile is listed as a UK idiom in the Cambridge Dictionary.[91] This may be reinforced by the Scottish property show *Safe as Houses* from 2007. The data further support the regional preference in the UK, as the simile's relative frequency in the BNC (0.143 i.p.m.) is significantly higher than in the COCA (0.020 i.p.m.).

#59 (AS) FRESH AS A DAISY

*You'll be (.) up **fresh as a daisy** at eight o'clock in the morning I know you will*

(Spoken BNC2014; SDZC 104)

This simile is generally used with two meanings. Firstly, it can mean 'alert, enthusiastic or well-rested', typically ascribed to a person target, especially after a period of rest or sleep. The second meaning is 'clean or well-kept', primarily used to describe places such as personal rooms and offices.

The simile (AS) FRESH AS A ROSE is related both semantically and functionally, but it is considered archaic, as its use dates back to Chaucer.[92] Some dictionaries consider even the simile (AS) FRESH AS A DAISY cliché. However, its presence in the Spoken BNC2014 suggests it is not entirely obsolete. Two speakers who used the simile were between 19 and 25 years of age, and the third was between 45 and 49.

#60 (AS) HARD AS ROCK

*The hearts of abortion supporters were **hard as rock** (...)*

---

[91] https://dictionary.cambridge.org/dictionary/english/as-safe-as-houses

[92] fresh as a daisy. (n.d.) The Dictionary of Clichés. (2013). Retrieved June 5 2023 from https://idioms.thefreedictionary.com/fresh+as+a+daisy

The simile (AS) HARD AS ROCK means 'physically hard or tough'. The target is usually *ice, soil* or various materials.

The lexemic overlap with (AS) HARD AS A ROCK raises several questions. Firstly, it may be interpreted as a grammatical mistake because the (standard) indefinite article is missing. These instances are possible, but the data refute such an interpretation to be generally applicable. Secondly, it possibly results from blending two related similes, (AS) HARD AS A ROCK and (AS) HARD AS STONE. The lexemes *rock* and *stone* are closely related but differ in countability. Nevertheless, lexemes are not uncommon to shift between countable and uncountable based on context. Thirdly, the simile may differ from (AS) HARD AS A ROCK in meaning and function. (AS) HARD AS A ROCK is mainly used to describe male sexual arousal, which is not necessarily true of (AS) HARD AS ROCK. However, the EW15 evidence suggests that (AS) HARD AS ROCK often refers to female *nipples* or *penis*.

Ultimately, (AS) HARD AS ROCK is best interpreted as a unique simile. It may share its origin with (AS) HARD AS A ROCK, but its meaning and function diverged. Other similes with the source *hard* include (AS) HARD AS NAILS, (AS) HARD AS STEEL or (AS) HARD AS IRON. These may share the meaning and function with (AS) HARD AS ROCK but often describe different targets.

Lastly, the CAS transformation ROCK(-)HARD is significantly more frequent than the simile form, which can be both (AS) HARD AS A ROCK or (AS) HARD AS ROCK.

Ultimately, (AS) HARD AS ROCK, same as (AS) HARD AS A ROCK, is frequent in sexually explicit descriptions and song lyrics (often referring to sexual scenarios). Therefore, it is infrequent in corpora not including sources with such content.

8 Frequency-based core of adjectival similes in dictionaries

This section is dedicated to testing the hypothesis that higher corpus frequencies correlate with better dictionary coverage. The 60 most frequent adjectival similes from section 7 are sought in the selected sources, and the findings are presented in five subsections for better clarity, with comments on their coverage. The section concludes with an overview of specific dictionary coverage rates and a tier distribution of adjectival similes based on dictionary recognition.

8.1 The review of dictionaries selected for simile coverage

Six dictionaries and one website dedicated to similes were selected for a simile coverage review. The first dictionary is Seidl and McMordie's *English Idioms and How to Use Them* (*English Idioms*; EI) published by Oxford University Press (OUP), representing a traditional printed dictionary of idioms. Given the time of its last edition (1988) and the general space limitations of hard copies, this dictionary only includes 66 adjectival simile types (according to this work's definition of a simile type).[93]

The second dictionary is the online *Cambridge Dictionary* (*Cambridge*; Cam) published by Cambridge University Press & Assessment 2023, which is widely considered one of the leading authorities on English for its extensive work in publishing, assessment (language certificates), and teaching. The dictionary includes data from American English, but they are somewhat limited.

Choice number three is the online version of the *Oxford Advanced Learner's Dictionary* (*Oxford*; Ox), published by OUP. OUP is another respected authority in the area of language research and teaching. However, this dictionary is primarily a learning tool. Therefore, its inclusion of idiomatic expressions is limited.

Dictionary number four is the *Collins online dictionary* (*Collins*; Col). Collins is one of the world's largest and most traditional dictionary publishers based in Glasgow. It was the first dictionary to base its contents on corpus research. Its primary focus is British English, but it also contains information about American English.

The fifth dictionary is the *Merriam-Webster online dictionary* (*Merriam-Webster*; MW). Merriam-Webster is a renowned American dictionary publisher and one of the

---

[93] For comparison, the 4th edition of *English Idioms and Howe to Use Them* from 1978 contains 166 adjectival similes.

leading dictionary authorities worldwide. The dictionary primarily covers American English but also includes data on British English.

Farlex's *The Free Dictionary* (TFD) is the last dictionary used for reviewing similes. It boasts one of the largest databases, and its entries are based on data from various associated dictionaries (online and printed).[94] This dictionary is the most inclusive of all the simile coverage review sources.

Lastly, the website *Simile Stack* (SiS) is part of the list. This source differs from all the others by not being a dictionary. SiS collects data by letting people from the public submit the similes themselves, which allows the site to include similes that might not be covered in the official dictionaries due to lack of evidence, regional preferences, or oversight. It describes its simile collection as the largest online but includes a lot of expressions unattested in the researched corpora. Additionally, some of the submitted entries are arguably creative similes that cannot be considered standard.

8.2 Coverage Tiers

Regarding dictionary coverage, the adjectival similes are divided into five tiers. Tier 5 items are covered in all the selected sources without exception. These can be considered the most standard based on the amount of dictionary recognition. Tier 4 items are covered in all but one source, which is typically *English Idioms*, but not exclusively. These might also be considered well-established. Tier 3 similes range from 3 to 5 source coverage. This is typically due to variant preferences of the particular dictionaries (*Cambridge*, *Collins*, *Merriam-Webster*), the specifics of *Oxford* or the datedness of *English Idioms*. Tier 2 examples are only covered in one or two sources, mostly *The Free Dictionary* and *Simile Stack*. Tier 1 items do not occur in any of the sources in any form.

8.3 1st list (#1-12)

Simile	EI	Cam	Ox	Col	MW	TFD	SiS
CLEAR AS DAY	x	✓	✓	✓	✓	✓	x
WHITE AS SNOW	x	x	x	✓	x	✓	✓
TOUGH AS NAILS	x	✓	✓	✓	✓	✓	✓
EASY AS PIE	✓	✓	✓	✓	✓	✓	✓

---

[94] Notably, it cross-references the *Collins* online dictionary also used for simile coverage review in this work. Consequently, adjectival similes occurring in the *Collins* also occur in *The Free Dictionary*.

Simile	EI	Cam	Ox	Col	MW	TFD	SiS
HARD AS A ROCK	x	x	x	x	✓	✓	x
GOOD AS GOLD	x	✓	✓	✓	✓	✓	✓
LIGHT AS A FEATHER	✓	✓	x	✓	✓	✓	✓
COLD AS ICE	✓	✓	x	✓	x	✓	✓
PLAIN AS DAY	x	x	✓	✓	✓	✓	✓
OLD AS TIME	x	x	x	x	x	✓	x
SMOOTH AS SILK	x	✓	x	x	✓	✓	✓
AMERICAN AS APPLE PIE	x	✓	✓	✓	✓	✓	x

Table 23. #1-12 simile coverage in dictionaries.

The first list (Tabel 23) contains similes ranking from #1 to #12 in this research (based on their frequency). The list contains a single Tier 5 simile, (AS) EASY AS PIE. The similes (AS) TOUGH AS NAILS, (AS) GOOD AS GOLD and (AS) LIGHT AS A FEATHER are Tier 4, making them appear well-established. The opposing trend can be observed with two Tier 2 similes, (AS) OLD AS TIME, only occurring in *The Free Dictionary*, and (AS) HARD AS A ROCK, which is covered just in *Merriam-Webster* and *The Free Dictionary*. The reason for their lack of coverage might be related to popular culture since both appear in songs. Consequently, the similes are possibly treated as expressions unique to their particular contexts. The former appears in the film The Beauty and the Beast song in the string *a tale as old as time.* The latter can be a popular song by the band AC/DC called *Hard as a rock*. However, the data show that the similes frequently appear in different texts with various targets, making their exclusion from dictionaries somewhat puzzling.

Overall, this list consists of adjectival similes with high frequencies in the dataset, and the dictionary coverage suggests that the corpus data used by dictionaries are somewhat limited. A perfect example is the simile (AS) WHITE AS SNOW, which is relatively frequent in the BNC, the COCA and the EW15 but remains unlisted by two major dictionaries, *Cambridge* and *Merriam-Webster*.

8.4 2nd list (#13-24)

Simile	EI	Cam	Ox	Col	MW	TFD	SiS
FREE AS A BIRD	x	✓	x	✓	✓	✓	✓
CLEAR AS A BELL	x	✓	x	✓	✓	✓	✓
SOLID AS A ROCK	x	✓	x	x	✓	✓	x

CUTE AS A BUTTON	x	x	x	x	✓	✓	✓
RIGHT AS RAIN	✓	✓	✓	✓	✓	✓	✓
CLEAR AS CRYSTAL	✓	x	x	✓	x	✓	✓
COOL AS A CUCUMBER	✓	✓	✓	✓	x	✓	✓
STRAIGHT AS AN ARROW	x	x	x	x	x	✓	✓
QUICK AS A FLASH	x	✓	✓	✓	✓	✓	✓
CLEAR AS MUD	✓	✓	✓	✓	✓	✓	✓
BLACK AS NIGHT	x	x	x	x	x	✓	x
BRIGHT AS THE SUN	x	x	x	x	x	x	✓

Table 24. #13-24 simile coverage in dictionaries.

The second list (Table 24) contains similes ranking from #13 to #24. It includes two Tier 5 adjectival similes, (AS) RIGHT AS RAIN and (AS) CLEAR AS MUD. The simile (AS) COOL AS A CUCUMBER is Tier 4 as it is, perhaps surprisingly, not listed in *Merriam-Webster*. Another Tier 4 simile is (AS) QUICK AS A FLASH, which does not appear in *English Idioms*.

The list also includes two Tier 2 similes, (AS) BLACK AS NIGHT only appearing in *The Free Dictionary* and (AS) BRIGHT AS THE SUN covered just in *Simile Stack*. The former is possibly considered less prominent than related similes with the ground *black*, such as (AS) BLACK AS COAL or (AS) BLACK AS PITCH. However, it can be considered traditional as it appears in Milton's *Paradise Lost* in the form *black it stood as Night*.[95] The simile (AS) BRIGHT AS THE SUN is perhaps treated as unique to scientific texts. It may also be erroneously interpreted as a literal comparison. Both similes are attested in the BNC, the COCA and the EW15 with significant frequencies (above the imposed cut-offs).

Another peculiarity is the (lack of) coverage of (AS) CLEAR AS CRYSTAL. *Cambridge*, *Oxford* and *Merriam-Webster* do not list this simile, but they all list its CAS transformation CRYSTAL(-)CLEAR. For comparison, the Tier 5 simile (AS) CLEAR AS MUD is less frequent in every corpus researched in this work. This suggests that empirical evidence may not be the only factor (or even the main factor) in determining the simile's inclusion in a dictionary.

---

[95] Book II, line 670.

8.5 3rd list (#25-36)

Simile	EI	Cam	Ox	Col	MW	TFD	SiS
OLD AS THE HILLS	✓	✓	✓	✓	✓	✓	✓
HARD AS NAILS	✓	✓	✓	✓	✓	✓	✓
THICK AS THIEVES	x	✓	✓	✓	✓	✓	✓
CLEAN AS A WHISTLE	x	✓	✓	✓	✓	✓	✓
HIGH AS A KITE	x	✓	✓	✓	✓	✓	✓
SHARP AS A TACK	x	x	x	x	x	✓	✓
WISE AS SERPENTS	x	x	x	x	x	x	x
SMOOTH AS GLASS	x	x	x	x	x	✓	x
NATURAL AS BREATHING	x	x	x	x	x	x	x
SWEET AS HONEY	x	x	x	x	x	✓	✓
HAPPY AS A CLAM	x	x	x	✓	✓	✓	✓
LIGHT AS AIR	x	x	x	x	x	✓	✓

Table 25. #25-36 simile coverage in dictionaries.

The third list (Table 25) comprises similes ranking from #25 to #36. It includes some diligently covered similes, as well as those seldom occurring in any dictionary. Two Tier 5 similes, (AS) OLD AS THE HILLS and (AS) HARD AS NAILS, along with the three Tier 4 similes, (AS) THICK AS THIEVES, (AS) CLEAN AS A WHISTLE and (AS) HIGH AS A KITE, can be considered well-established. The Tier 4 similes are only absent from *English Idioms*.

The list includes only a single Tier 3 simile, (AS) HAPPY AS A CLAM, not covered in *English Idioms, Cambridge* and *Oxford*. Three Tier 2 simile types are absent from all the sources but *The Free Dictionary* and *Simile Stack*: (AS) SHARP AS A TACK, (AS) SWEET AS HONEY and (AS) LIGHT AS AIR. The Tier 2 simile (AS) SMOOTH AS GLASS is covered just by *The Free Dictionary*.

Lastly, two similes, (AS) WISE AS SERPENTS and (AS) NATURAL AS BREATHING, are Tier 1. This is somewhat surprising since the data show that they are indeed adjectival similes. Their lack of coverage might be because the former is strongly associated with the biblical discourse. The latter is perhaps erroneously interpreted as a literal comparison. In addition to the corpus evidence, a Google Search also returns many examples of (AS) NATURAL AS BREATHING used to describe effortless activities or innate abilities that appear natural to the target.

8.6 4th list (#37-48)

Simile	EI	Cam	Ox	Col	MW	TFD	SiS
SICK AS A DOG	x	✓	✓	✓	✓	✓	✓
HARD AS STEEL	x	x	x	x	x	x	x
STIFF AS A BOARD	x	x	x	✓	✓	✓	✓
WHITE AS A SHEET	✓	✓	x	✓	x	✓	x
RED AS BLOOD	x	x	x	x	x	✓	✓
FLAT AS A PANCAKE	✓	✓	✓	✓	✓	✓	✓
CHEAP AS CHIPS	x	x	x	✓	x	✓	✓
OLD AS HUMANITY	x	x	x	x	x	x	x
QUIET AS A MOUSE	✓	x	x	✓	✓	✓	✓
DRY AS A BONE	✓	✓	x	✓	✓	✓	✓
FAST AS LIGHTNING	x	x	x	x	x	✓	x
DARK AS NIGHT	x	x	x	x	x	x	x

Table 26. #37-48 simile coverage in dictionaries.

List number four (Table 26) consists of similes ranking from #37 to #48. It contains a single Tier 5 simile, (AS) FLAT AS A PANCAKE. Two Tier 4 similes are included: (AS) SICK AS A DOG, absent from *English Idioms*, and (AS) DRY AS A BONE, not occurring in *Oxford*.

Four Tier 3 similes appear in this list, with two examples being rather peculiar. The Tier 3 simile (AS) WHITE AS A SHEET is not covered by *Oxford*, *Merriam-Webster* and *Simile Stack*. This might suggest that the simile is considered chiefly British, as *Merriam-Webster* and *Simile Stack* are primarily American sources. However, it is not *Merriam-Webster's* policy to exclude British English and the data from the COCA show that (AS) WHITE AS A SHEET occurs relatively frequently in American English. *Oxford's* lack of coverage can be attributed to streamlining the vocabulary for learners and omitting more advanced expressions. The second interesting simile is (AS) QUIET AS A MOUSE, which is not covered by *Cambridge* and *Oxford*. The dictionary survey shows that *Cambridge* usually includes similes listed in *English Idioms*. This would also be expected with *Oxford*[96], but its primary purpose may once again be the reason for excluding some standard similes.

---

[96] The publishing house for *Oxford* and *English Idioms* is identical.

The list also comprises three Tier 1 similes: (AS) HARD AS STEEL, (AS) OLD AS HUMANITY and (AS) DARK AS NIGHT. The reasons behind the exclusion of these similes from dictionaries cannot be their corpus frequencies, so the explanation lies in their simile status. (AS) HARD AS STEEL is one of many similes with the ground *hard*, which might make it less prominent than the other more frequent similes. However, (AS) HARD AS A ROCK, the most frequent simile with *hard* in the dataset, is Tier 2. This suggests that similes with *hard* might be considered lacking in idiomaticity despite the corpus evidence. (AS) OLD AS HUMANITY is one of many adjectival similes with the ground *old*. The different sources might have different connotations, but their meaning is essentially the same, 'very old'. It is perhaps that dictionaries only include the traditional (AS) OLD AS THE HILLS, and the other simile types with *old* are neglected. (AS) DARK AS NIGHT seems to be similar to its synonymous Tier 2 simile (AS) BLACK AS NIGHT in terms of dictionary recognition. The source *night*, together with the ground *dark* and *black*, is probably not considered idiomatic enough by the dictionaries to be interpreted as a simile, despite the simile (AS) BLACK AS NIGHT occurring with significant frequencies in the BNC, the COCA and the EW15.

8.7 5th list (#49-60)

Simile	EI	Cam	Ox	Col	MW	TFD	SiS
HARD AS STONE	x	x	x	x	x	✓	✓
STRONG AS STEEL	x	x	x	x	x	✓	x
DEAD AS A DOORNAIL	x	✓	✓	✓	✓	✓	✓
BRIGHT AS DAY	x	x	x	x	x	x	✓
SMART AS A WHIP	x	x	x	x	✓	✓	✓
STRONG AS AN OX	x	x	x	✓	x	✓	✓
BLIND AS A BAT	✓	✓	✓	✓	✓	✓	✓
MAD AS A HATTER	✓	✓	✓	✓	✓	✓	✓
PRETTY AS A PICTURE	x	x	✓	x	✓	✓	✓
SAFE AS HOUSES	✓	✓	✓	✓	x	✓	x
FRESH AS A DAISY	✓	✓	x	✓	✓	✓	✓
HARD AS ROCK	x	x	x	x	x	x	x

Table 27. #49-60 simile coverage in dictionaries.

The last list (Table 27) contains similes ranking from #49 to #60. The list includes two Tier 5 similes, (AS) BLIND AS A BAT and (AS) MAD AS A HATTER. Two Tier 4 similes, (AS) DEAD AS A DOORNAIL and (AS) FRESH AS A DAISY, are not covered by *English Idioms* and *Oxford*, respectively.

The list comprises three Tier 2 similes. (AS) HARD AS STONE is one of the *hard*-ground similes that are generally overlooked by the sources, and it is covered only by *The Free Dictionary* and *Simile Stack*. Another Tier 2 simile, (AS) STRONG AS STEEL, is less frequent than the related Tier 1 (AS) HARD AS STEEL, but it is at least covered by *The Free Dictionary*. The last Tier 2 simile, (AS) BRIGHT AS DAY, is only covered by *Simile Stack*, despite occurring frequently in the BNC, the COCA and the EW15.

Only a single simile in this list is Tier 5, (AS) HARD AS ROCK. Its absence from the dictionaries is understandable since it is almost identical to the more frequent Tier 4 simile (AS) HARD AS A ROCK, whose coverage is likewise poor. Nevertheless, (AS) HARD AS ROCK is attested in the BNC and occurs with significant frequencies in the COCA and the EW15.

8.8 Dictionary coverage summary

Figure 7 provides a statistical overview of the Tier distribution of the top 60 similes.
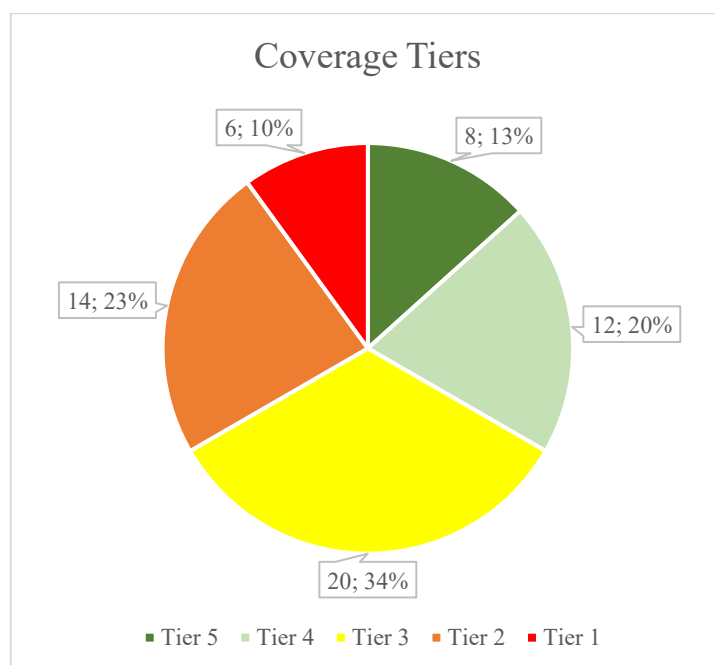
Figure 7. An overview of the adjectival simile Tiers regarding dictionary coverage.

The Tier distribution resembles a standard bell-shaped curve. The largest group are Tier 3 similes, represented by 20 types (34%). Tier 4 comprises 12 types (20%), and Tier

2 includes 14 examples (23%). Finally, eight similes (13%) get perfect coverage of 100% (Tier 5), while six similes (10%) do not appear in any source (Tier 1).

Table 28 provides an overview of the overall dictionary coverage of adjectival similes.

Dictionary	Coverage
English Idioms	28.33%
Cambridge	48.33%
Oxford	36.67%
Collins	58.33%
Merriam-Webster	53.33%
The Free Dictionary	86.67%
Simile Stack	71.67%

Table 28. Overall dictionary coverage of adjectival similes.

Three dictionaries do not reach the threshold of 50% coverage. Unsurprisingly, *English Idioms* has the lowest coverage (28.33%) due to its format and datedness. The second dictionary is *Oxford*, also exhibiting a fairly low coverage (36.67%) owing to its focus on the learner audience. The third dictionary is *Cambridge*, whose coverage also appears fairly low (48.33%), and it is challenging to find a plausible explanation. The dictionary includes American English and slang, but its data sources must be somehow limited. Perhaps *Cambridge's* specific quality criteria rule out source texts that would generate additional similes.

Merriam-Webster and *Collins* show a similar degree of coverage, 53.33% and 58.33%, respectively. Despite above 50%, the number is not much higher than *Cambridge*. This suggests that adjectival similes are perhaps too peripheral to merit meticulous coverage by renowned dictionaries.

Simile Stack, a website dedicated solely to similes, exhibits 71.67% coverage, which is significantly higher than the esteemed dictionaries. Nonetheless, it still translates to one out of four similes not being included. It follows that entries submitted by people cannot account for all the adjectival similes attested in corpora, despite the website's potential to include less frequent similes that might get overlooked in corpora due to insensitive cut-offs.

The Free Dictionary boasts the best coverage (86.67%). It includes similes that often do not appear in other major dictionaries, suggesting that it has access to extensive data or uses different policies and criteria regarding the inclusion of lexical items. It includes

various dictionaries of idioms among its sources, which possibly explains its high coverage of adjectival similes.

To conclude, it is evident that corpus frequencies do not necessarily correspond to better dictionary coverage. This goes against the premise that higher corpus frequencies correlate with dictionary coverage. Corpus evidence plays a role in dictionary inclusion, but other mechanisms and criteria may take precedence over corpus frequencies. The findings suggest that dictionaries mostly use corpora to check frequencies of predetermined adjectival simile types rather than to extract all attested examples. This is reflected in the overall dictionary coverage observed in the survey. It is also believed that including more online dictionaries would yield similar results.

9 Formal analysis of adjectival similes

This section presents a data-based classification of adjectival similes. The formal classifications are subdivided into general and specific.

9.1 General formal classification

The formal classifications presented in this section are designed to cluster similes with identical properties. General formal classification is presented in two different ways. The classification according to the syllable count of the mandatory constituting elements is purely formalistic and provides us with an overall picture of the length of typical similes, usually spanning from three to six syllables. Conversely, the classification according to rhythm may help us to understand why specific similes survive longer than others.

9.1.1 Syllable count

As was already mentioned earlier, the syllable count is a formal classification based on the length of the constituting elements. Both the initial *as* and the comparator *as* are monosyllabic; therefore, they are excluded from the classification as a stable element. This categorisation applies to all similes despite revealing little about their meaning or function.

9.1.1.1 Ground syllable count

The ground is arguably the most prominent constituent in adjectival similes. It is the initial member of the fixed frame and functions as the simile head, both syntactically and semantically. Lexical items representing the ground tend to be familiar words, easily recognised by most speakers. Consequently, the ground is typically a monosyllabic word. Figure 8 illustrates the ground syllable count in the 309 adjectival simile types collected.

GROUND SYLLABLE COUNT
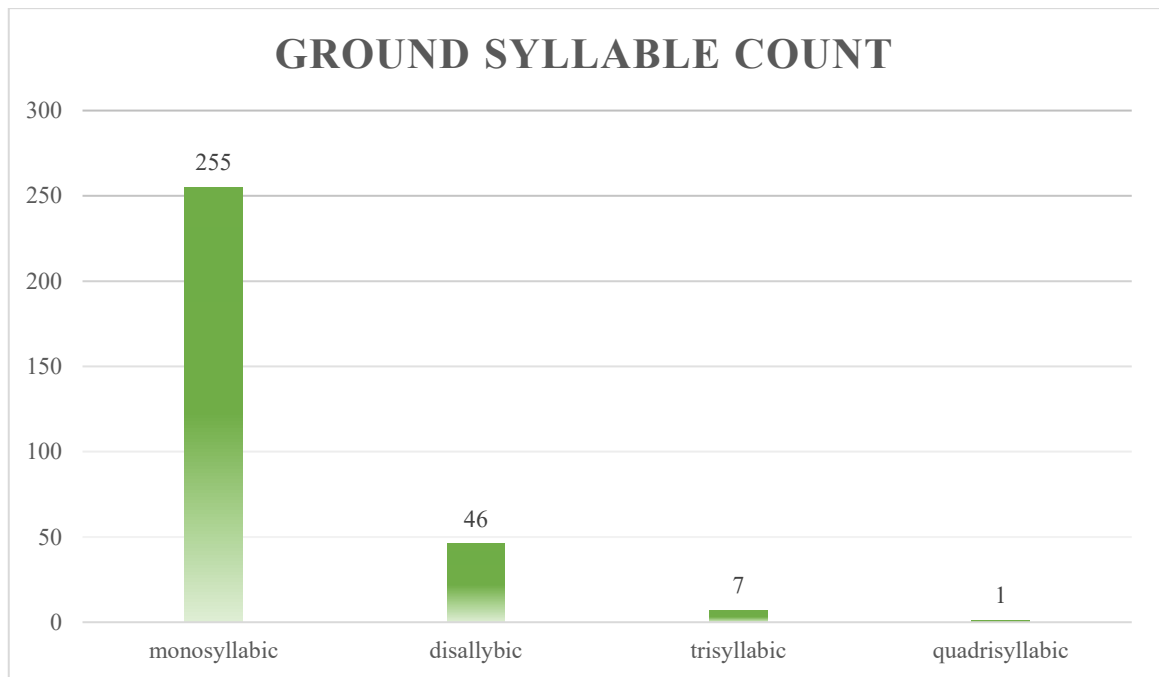
Figure 8. Ground syllable count overview.

In total, 255 simile types contain a monosyllabic ground, which equals 82.52% of all the types. The findings corroborate the assumption that monosyllabic grounds are the prevalent pattern. Examples of monosyllabic-ground similes are (AS) BLIND AS A BAT, (AS) HIGH AS A KITE or (AS) WHITE AS SNOW.

The second most frequent length of the ground is two syllables. A total of 46 similes (14.89%) have a disyllabic ground, including (AS) BUSY AS A BEE, (AS) HAPPY AS A CLAM or (AS) SIMPLE AS PIE.

The data suggest that trisyllabic-ground similes are relatively uncommon. Only seven simile types (2.27%) in the dataset contain a trisyllabic ground; for example, (AS) NATURAL AS BREATHING, (AS) REGULAR AS CLOCKWORK or (AS) SLIPPERY AS AN EEL. Similes with the ground *different* could also be considered to contain a trisyllabic ground; however, *different* is standardly pronounced as a disyllabic word in present-day English. The same argument might be used for *slippery*, but while dictionaries list both disyllabic and trisyllabic pronunciations, the recorded pronunciation is exclusively trisyllabic.

Tetrasyllabic grounds are extremely rare, as shown by the data. Only one simile type (0.32%) in the dataset undisputedly represents this group: (AS) AMERICAN AS APPLE PIE. The ground *miserable* might also be considered tetrasyllabic, but most speakers pronounce it as trisyllabic.

To summarise, a monosyllabic ground appears to be the prototypical form in adjectival similes. Determining the syllable count in the ground may be challenging due

to elision that tends to occur in trisyllabic and tetrasyllabic words with familiar suffixes. Naturally, elision is not easy to assess, as it usually depends on the speaker's preference, the communicative situation, and an overall feeling of appropriateness.

9.1.1.2 Source syllable count

Compared to the ground, the source syllable count is more versatile. The range of syllables mostly remains within one to four in conventional similes, but it may extend well beyond four syllables in novel or expanded similes.[97] Figure 9 shows the syllabic distribution in adjectival simile sources in the dataset.
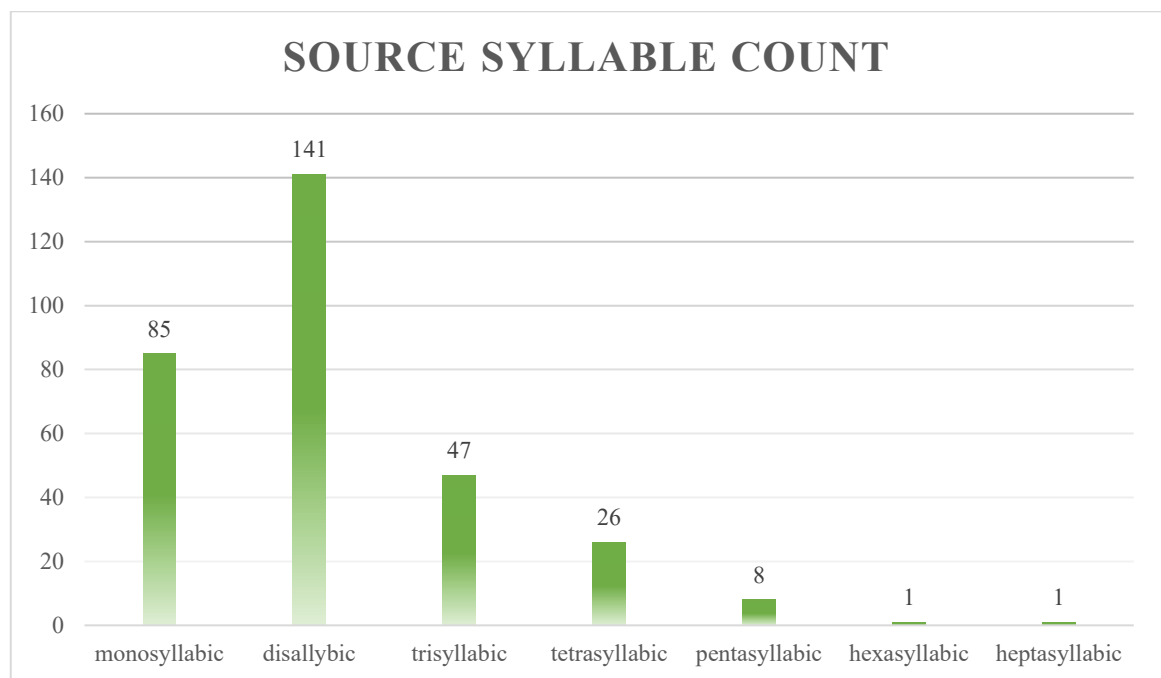
Figure 9. Source syllable count overview.

The source is primarily represented by a noun phrase, conventionally headed by a noun. The number category is an essential factor as it determines the occurrence of an article. When present, the article adds an extra syllable – an essential aspect in interpreting the source syllable count.

Among the adjectival simile types in the dataset, 85 (27.51%) have a monosyllabic source – the second most frequent syllable count. These include (AS) CLEAR AS DAY, (AS) HARD AS STEEL or (AS) TOUGH AS NAILS. The lack of an article typically means the head noun is uncountable or plural. However, instances of ungrammatical forms missing the article are not uncommon, especially in web crawler corpora.

---

[97] The term expanded similes refers to forms with additional (typically nonstandard) modifications and expansions of the ground.

The most frequent source length in the sample is disyllabic, occurring in 141 types (45.63%). This is mainly caused by an indefinite article accompanying a monosyllabic countable noun, for instance, (AS) CLEAR AS A BELL, (AS) HAPPY AS A CLAM or (AS) SHARP AS A KNIFE. However, examples of disyllabic-source similes with a definite article also exist, including (AS) BLUE AS THE SKY, (AS) CLEAR AS THE SUN or (AS) FREE AS THE WIND. Another group are similes with an uncountable disyllabic noun as a source, such as (AS) CLEAR AS CRYSTAL, (AS) QUICK AS LIGHTNING or (AS) SMOOTH AS BUTTER. Sometimes the source can be a disyllabic plural form, making the source disyllabic due to the lack of article, for example, (AS) BIG AS HOUSES, (AS) SHARP AS RAZORS or (AS) WIDE AS SAUCERS. Some plural forms are standard, while others are transformations of the original singular-form source. Lastly, the source can be a compound noun, as in (AS) BLACK AS MIDNIGHT, (AS) CLEAR AS DAYLIGHT or (AS) THICK AS PIGSHIT.

A trisyllabic source is also fairly frequent, occurring in 47 types (15.21%). Examples of a trisyllabic-source simile usually include the combination of an article or other determiner and a disyllabic head noun, such as (AS) CLEAN AS A WHISTLE, (AS) LIGHT AS A FEATHER or (AS) SICK AS YOUR SECRETS. Other examples are (AS) DIFFERENT AS CHALK AND CHEESE or (AS) THICK AS TWO SHORT PLANKS, with the former coordinating two monosyllabic head nouns and the latter combining a determiner and a premodifier of a monosyllabic head.

Simile types with a tetrasyllabic source occur in 26 instances (8.41%). The first group are trisyllabic head nouns with a determiner, such as (AS) BIG AS A FOOTBALL FIELD, (AS) COOL AS A CUCUMBER or (AS) OLD AS HUMANITY. The second group are similes with complex noun phrases as their source, for example, (AS) BLACK AS THE ACE OF SPADES, (AS) DUMB AS A BOX OF ROCKS or (AS) SMOOTH AS A BABY'S BUTT.

A pentasyllabic source is infrequent and appears in eight types (2.59%). Four out of those instances are synonymous similes: (AS) SMOOTH AS A BABY'S BEHIND, (AS) SMOOTH AS A BABY'S BOTTOM, (AS) SOFT AS A BABY'S BEHIND and (AS) SOFT AS A BABY'S BOTTOM. The remaining four types are (AS) HIGH AS AN ELEPHANT'S EYE, (AS) OLD AS CIVILIZATION, (AS) NAKED AS THE DAY (ONE) WAS BORN, and (AS) PLAIN AS THE NOSE ON (ONE'S) FACE, with the object typically represented by a monosyllabic pronoun, such as s*he* or *he,* in the latter two types.

Hexasyllabic and heptasyllabic sources occur only once each (0.32%), represented by (AS) BIG AS A FOOTBALL STADIUM and (AS) NAKED AS THE DAY ONE CAME TO THIS WORLD, respectively.

To summarise, no set limit exists for the length of the source, and novel or expanded similes may contain extensive sources that can be whole clauses. Nonetheless, the prevailing syllable count stays within the range of one to four syllables, as supported by the frequencies observed in the dataset.

9.1.2 Ground-source syllable count correlation

The syllable count correlation between the ground and the source helps to identify the average length of a simile. The minimum length is three syllables, comprising a monosyllabic ground, comparator *as*, and a monosyllabic source.[98] Similes with identical lengths of both primary constituents are labelled symmetric; those with varying lengths are asymmetric. Formal symmetry does not contribute to rhythmicality; in fact, it typically works against it (see Section 9.1.3).

9.1.2.1 Monosyllabic ground correlations

Figure 10 shows the distribution of sources for simile types (*n=255*) with a monosyllabic ground.
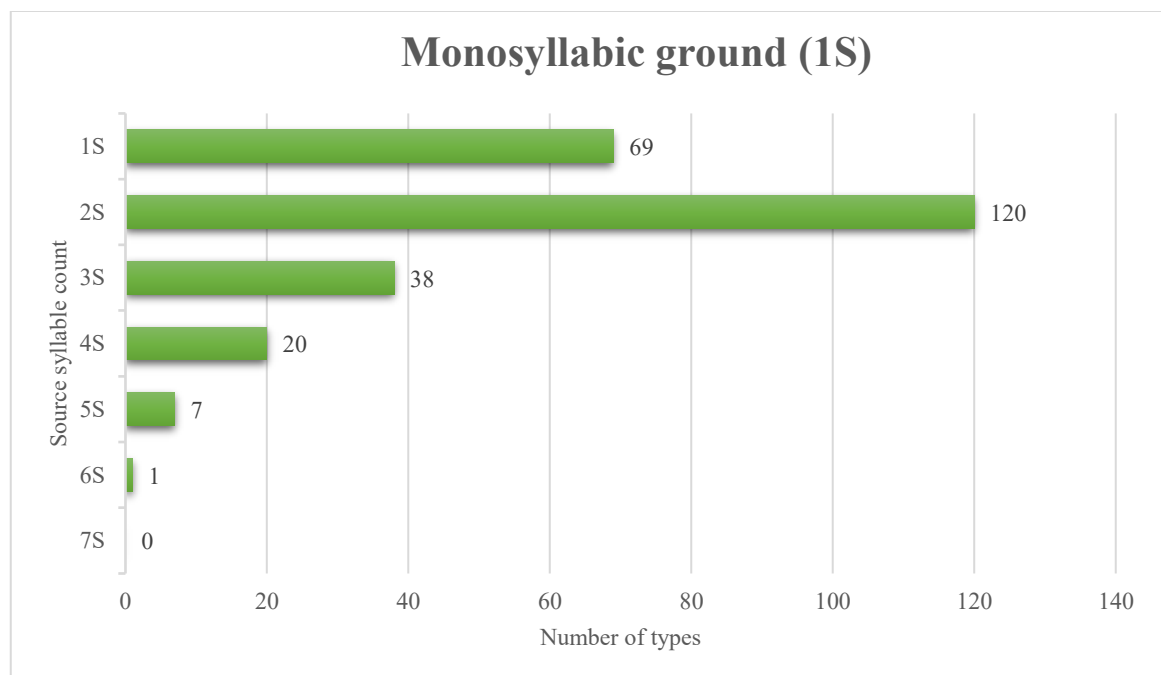
Figure 10. Monosyllabic ground correlations.

A total of 69 similes (27.06%) with a monosyllabic ground are symmetric, with the '1S *as* 1S' pattern. The most frequent combination is '1S *as* 2S', represented by 120 types

---

[98] The initial *as* is excluded from the count due to being optional.

(47.06%). The pattern '1S *as* 3S' occurs in 38 similes (14.90%), followed by '1S *as* 4S' in 20 types (7.84%), '1S *as* 5S' in 7 types (2.75%), and only a single instance (0.39%) of the '1S *as* 6S' pattern.

9.1.2.2 Disyllabic ground correlations

Figure 11 presents an overview of disyllabic ground correlations.
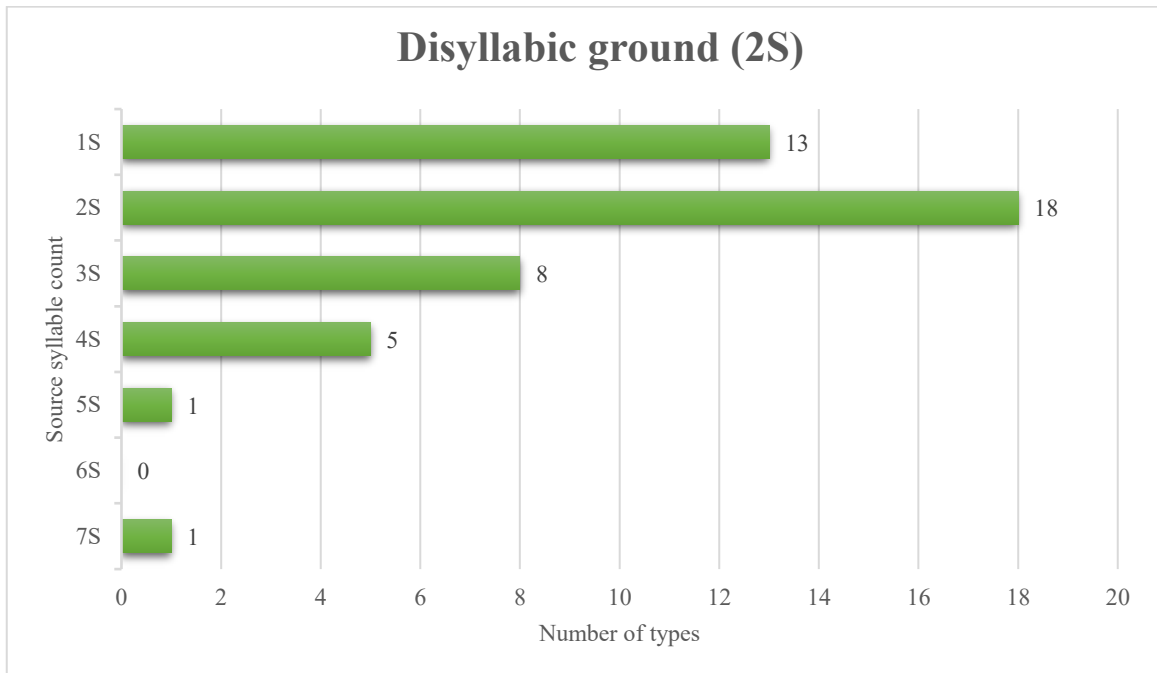
Figure 11. Disyllabic ground correlations

Of the 46 similes with a disyllabic ground, 18 (39.13%) are symmetric with the '2S *as* 2S' pattern. The '2S *as* 1S' pattern occurs 13 times (28.26%), '2S *as* 3S' appears in 8 types (17.39%), followed by '2S *as* 4S' in 5 similes (10.87%), and a single instance of '2S *as* 5S' and '2S *as* 7S' (2.17%).

9.1.2.3 Trisyllabic and tetrasyllabic ground correlations

Similes with tri- and tetrasyllabic grounds are relatively rare (8 types in the dataset), and their source lengths do not exceed three syllables. Of the seven trisyllabic similes, three have the '3S *as* 1S' pattern, three have the '3S *as* 2S' pattern, and only one is symmetrical ('3S *as* 3S'). The only simile with a tetrasyllabic ground has the '4S *as* 3S' pattern. The infrequency of these patterns corroborates the common observation that similes typically comprise morphologically simple components, which also applies to the source.

9.1.3 Rhythm

As another general classification, the categorisation according to rhythm applies to all similes. Moreover, it can help us understand why particular similes occur in poetry or songs, as their rhythm affects their functionality within various texts. Prosodic features are believed to be vital in language acquisition and reproduction. Therefore, rhythmicality should be accounted for in any phraseological expression.

This work adopts the stress-timed rhythm theory (e.g., Roach, 2009: 107). The idea of using poetic rhythm terminology was dismissed, as it is somewhat unreliable. The boundaries between possible combinations are fuzzy. For example, (AS) SLOW AS MOLASSES might be interpreted as either a trochee followed by an amphibrach or a dactyl followed by a trochee. Additionally, using the poetic rhythm approach would result in too many subgroups, compromising the categorisation's usefulness.

Most standard similes are dimetric (two feet) or trimetric (three feet) – the few extensive similes in the dataset are marginal examples. The word boundary is irrelevant to English rhythm; therefore, the following analysis focuses on stressed syllables as the beginnings of feet. This also allows us to ignore the optional initial *as* because it is typically unstressed and either belongs to a preceding foot or functions as a pre-head of the following foot at the beginning of a sentence.

9.1.3.1 Dimetric similes

Most adjectival similes comprise two metrical feet, one headed by the ground and the other by the source. The dimetric structure is represented by 265 simile types (85.76% of all the types collected in this research). Figure 12 illustrates the possible patterns along with their frequency.
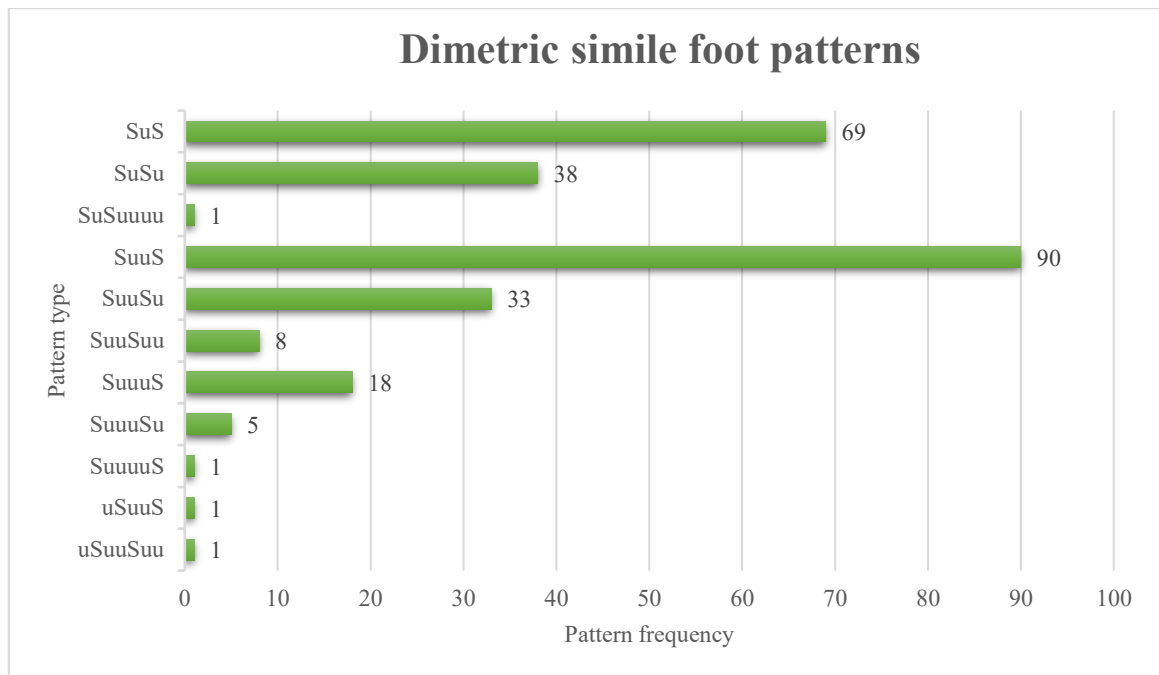
Figure 12. Dimetric simile foot patterns overview

The length of dimetric similes in the dataset ranges from three to seven syllables. The comparator *as* makes at least one unstressed syllable obligatory, as represented by the minimal pattern 'SuS', occurring 69 times (26.04%). Examples of the pattern are similes with monosyllabic grounds and sources, such as (AS) BLACK AS PITCH, (AS) GOOD AS SEX or (AS) PURE AS GOLD. Another unstressed syllable is frequently provided by an article, resulting in the 'SuuS' pattern, which is the most frequent of all the dimetric patterns, with 90 instances (33.96%), as in (AS) CLEAR AS A BELL, (AS) QUICK AS A FLASH or (AS) WHITE AS A SHEET.

The third most frequent dimetric pattern is 'SuSu', occurring in 38 simile types (14.34%). It is unanimously represented by similes with a monosyllabic ground and a disyllabic uncountable or plural source, for example, (AS) CLEAR AS CRYSTAL, (AS) SMOOTH AS BUTTER or (AS) WIDE AS SAUCERS. The pattern 'SuuSu' is the fourth most frequent, predominantly consisting of a monosyllabic ground, the comparator *as*, and a trisyllabic source comprising a determiner and a disyllabic noun phrase, such as (AS) BOLD AS A LION, (AS) CLEAN AS A WHISTLE or (AS) STRAIGHT AS AN ARROW. The only exception is (AS) SLOW AS MOLASSES, with the trisyllabic noun phrase stressed penultimately. The fifth pattern is 'SuuuS', occurring 18 times (6.79%). Unanimously, the pattern comprises an initially stressed disyllabic ground and monosyllabic noun phrase preceded by an article, as in (AS) BUSY AS A BEE, (AS) LONELY AS A CLOUD or (AS) SILENT AS THE GRAVE.

The remaining patterns are considered marginal, as suggested by their frequencies. The pattern 'SuuSuu' occurs eight times (3.02%), represented by similes with a monosyllabic ground and a tetrasyllabic source, such as (AS) COOL AS A CUCUMBER or (AS) OLD AS METHUSELAH. 'SuuuSu' is represented by five types (1.89%), four of which have either a disyllabic or trisyllabic ground, such as (AS) PRETTY AS A PICTURE or (AS) NATURAL AS BREATHING, with only one instance comprising a monosyllabic ground and a tetrasyllabic source ((AS) OLD AS THE REPUBLIC). The patterns 'SuSuuu' ((AS) WHITE AS ALABASTER), 'SuuuuS' ((AS) SLIPPERY AS AN EEL), 'uSuuS' ((AS) TRANSPARENT AS GLASS) and 'uSuuSuu' ((AS) UNIQUE AS A FINGERPRINT) are unique and occur only once each. The first pattern carries a pentasyllabic foot, caused by the combination of a trisyllabic[99] ground and an indefinite article in the source. The second and third patterns employ grounds stressed on the second syllable, which is rare in adjectival similes.

9.1.3.2 Trimetric similes

Trimetric adjectival similes contain complex sources. Naturally, this makes them relatively uncommon, as complexity often works against memorability and replicability. Figure 13 gives an overview of the trimetric patterns.
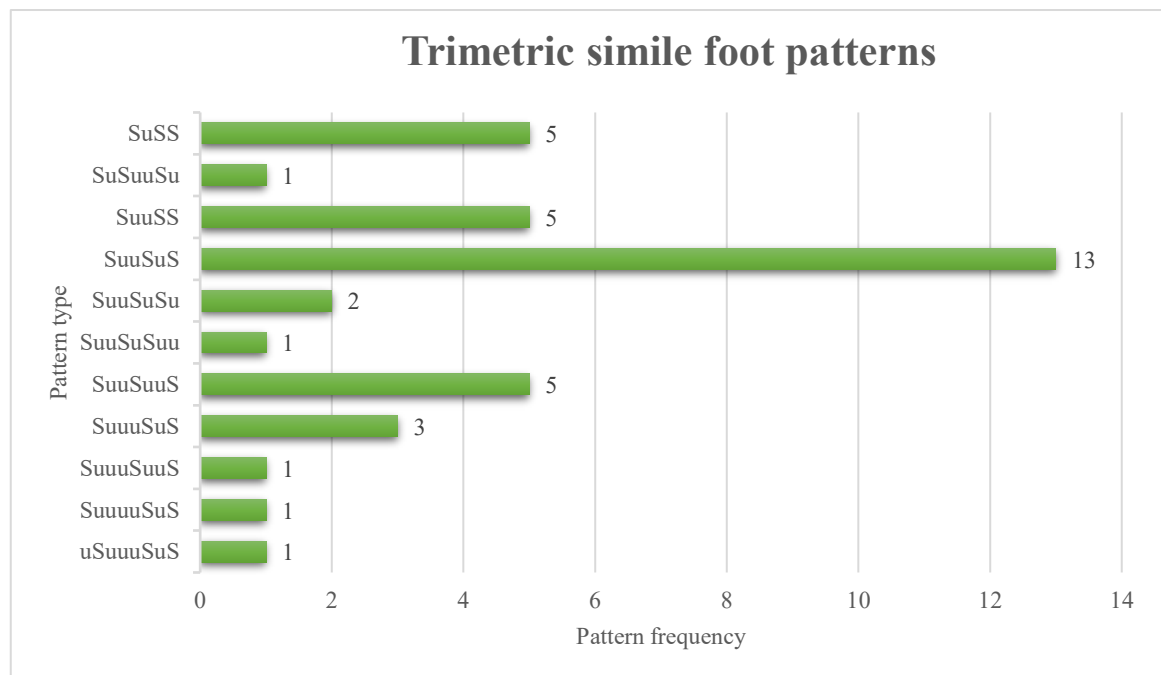
Figure 13. Trimetric simile foot patterns overview

---

[99] The ground *slippery* might be realised as disyllabic in rapid speech.

The overall frequency of trimetric simile types is 39, which is 12.62% of all the types in the dataset. Consequently, no pattern is considered frequent, as trimetric similes generally appear marginal.

The pattern 'SuuSuS' occurs in 13 trimetric types (33.33%) and comprises either a monosyllabic ground with a tetrasyllabic source, such as (AS) PURE AS THE DRIVEN SNOW or (AS) BLACK AS THE ACE OF SPADES, or a disyllabic ground with a trisyllabic source, for example, (AS) DIFFERENT AS DAY AND NIGHT.

Occurring in six simile types (15.38%), the pattern 'SuuSS' consists of either a monosyllabic source with a trisyllabic ground ((AS) POOR AS A CHURCH MOUSE), or a disyllabic source with a disyllabic ground ((AS) DAPPY AS PIG SHIT). The source is typically a compound.

Two patterns occur with the same frequency of five instances (12.82%). 'SuSS' comprises a monosyllabic ground with an uncountable or plural disyllabic compound as a source, as in (AS) OLD AS MANKIND or (AS) TOUGH AS OLD BOOTS. 'SuuSuuS' comprises a monosyllabic ground with a pentasyllabic source, usually including a possessive premodifier, for example, (AS) HIGH AS AN ELEPHANT'S EYE or (AS) SMOOTH AS A BABY'S BEHIND.

The pattern 'SuuuSuS' occurs in three synonymous trimetric types, (AS) HAPPY AS A PIG IN MUCK, (AS) HAPPY AS A PIG IN MUD and (AS) HAPPY AS A PIG IN SHIT. The pattern 'SuuSuSu' occurs twice and is, again, represented by two synonymous similes: (AS) SMOOTH AS BABY'S BOTTOM and (AS) SOFT AS BABY'S BOTTOM. These trimetric patterns suggest that speakers might remember the simile's meaning, but the lexical composition is not fixed as long as the original rhythmical pattern is preserved.

The remaining patterns, 'SuSuuSu' ((AS) OLD AS CIVILIZATION), 'SuuuSuSuu' ((AS) BIG AS A FOOTBALL STADIUM), 'SuuuSuuS' ((AS) NAKED AS THE DAY (ONE) WAS BORN), 'SuuuuSuS' ((AS) SERIOUS AS A HEART ATTACK) and 'uSuuuSuS' ((AS) AMERICAN AS APPLE PIE), appear only once each, making them exclusive to individual similes.

9.1.3.3 Tetrametric similes

Based on the collected data, tetrametric similes are rare. Only five simile types (1.62%) in all four corpora carry a tetrametric pattern. Three of those patterns exhibit an unusual concentration of three feet with no unstressed syllables 'SuSSS' appears in (AS) THICK AS TWO SHORT PLANKS, 'SuuSSS' occurs twice in synonymous types (AS) EASY AS ABC and (AS) SIMPLE AS ABC, and 'SuuuSSS' is represented by (AS) CROOKED AS A

DOG'S HIND LEG. The last pattern, 'SuuuSuSuuS', appears in (AS) NAKED AS THE DAY (ONE) CAME TO THIS WORLD. Arguably, this simile's length and rhythmical design are beyond the threshold of a memorable pattern.

9.1.3.4 Rhythm summary

To summarise this subsection, rhythm appears to be an essential factor in similes. The most frequent rhythmical patterns are dimetric, four of which are significantly recurrent. These four patterns lend themselves conveniently to the poetic rhythm interpretation, as presented in Figure 14.

(as) **good** as **gold**   (SuS - cretic)

(as) **sweet** as **su**gar   (SuSu - ditrochee)

(as) **smart** as a **whip**   (SuuS - choriambus)

(as) **flat** as a **pan**cake   (SuuSu - dactyl + trochee)

Figure 14. Recurring rhythmical patterns and their poetic interpretation.

The issue potentially arises with the initial *as* considered a part of the rhythmical pattern. (AS) GOLD AS GOLD could then be interpreted as a di-iambus and (AS) SMART AS A WHIP as an iambus followed by an anapaest. The realisation of initial *as* may well be dependent on the immediate context and its prosody, but such an assumption would require extensive testing.

Overall, the foot distribution observed in adjectival similes suggests that rhythm is crucial in standard similes. However, if we look at the most frequent similes in the dataset, the pattern 'SuS' appears prevalent (see Table 29).

Rank	Simile	Pattern
1	**CLEAR AS DAY**	SuS
2	**WHITE AS SNOW**	SuS
3	**TOUGH AS NAILS**	SuS
4	**EASY AS PIE**	SuS
5	**HARD AS A ROCK**	SuuS
6	**GOOD AS GOLD**	SuS
7	**LIGHT AS A FEATHER**	SuuS
8	**COLD AS ICE**	SuS

| 9 | **PLAIN AS DAY** | SuS |
| 10 | **OLD AS TIME** | SuS |

Table 29. The ten most frequent simile types and their patterns.

The most frequently occurring similes suggest that rhythmical simplicity is preferred over a complex design in basic communication. Consequently, language users might opt for trimetric and tetrametric similes in carefully conceived texts rather than ad hoc interactions. This postulate would require thorough research into the types of texts in which the similes occur.

9.2 Specific formal classification

While the general formal classifications apply to all adjectival similes, the specific ones only concern the similes containing a non-obligatory phenomenon, making them marked. Generally, similes can be considered prosodically unmarked when they do not possess any of the three features: alliteration, assonance, or rhyme. The data reveal that 247 simile types (79.94%) are unmarked.[100] Only two similes from the ten most frequent types are marked – the alliteration in (AS) GOOD AS GOLD and the assonance in (AS) PLAIN AS DAY. Despite undisputedly making the simile more memorable, prosodic markedness does not appear to be the main factor affecting a simile's frequency.

9.2.1 Alliterating similes

Alliteration is the most frequent prosodic feature in similes. In this work, only the initial phoneme repetition is considered. Examples such as (AS) WISE AS SERPENTS could also be mentioned as containing alliterating sibilant consonants /s/ and /z/, but these instances are not considered here.

The dataset provides us with a list of 38 simile types (12.30%) that are alliterating, for example, (AS) CLEAR AS CRYSTAL, (AS) GREEN AS GRASS or (AS) STILL AS A STATUE. All the alliterating types in the dataset have alliterating consonants; no vowel alliteration occurs. Table 30 illustrates the distribution of alliterating consonants in similes.

Consonant	Example simile	Frequency
/b/	BLIND AS A BAT	7
/s/	SOFT AS SILK	6
/d/	DEAD AS A DOORNAIL	5

---

[100] Moon (2008) discovered that 23% of the *as*-similes in Bank of English "either rhyme or have alliterating initial consonants." The corpora researched in this work contain 20.06% marked types, including assonance.

/p/	PRETTY AS A PICTURE	5
/h/	HIGH AS A HOUSE	3
/k/	COOL AS A CUCUMBER	2
/g/	GOOD AS GOLD	2
/l/	LARGE AS LIFE	1
/f/	FIT AS A FIDDLE	1
/m/	MAD AS A MARCH HARE	1
/r/	RIGHT AS RAIN	1
/t/	TALL AS A TREE	1
/tʃ/	CHEAP AS CHIPS	1
/θ/	THICK AS THIEVES	1
/w/	WEAK AS WATER	1

Table 30. Alliterating consonants in adjectival similes.

While the data do not provide conclusive evidence, plosive consonants seem to alliterate more often than others. Overall, 15 unique consonants alliterate, seven of which appear only once. The most frequent alliterating consonant is the plosive /b/, occurring in 7 simile types, followed by the sibilant /s/ in 6 types, and the plosives /d/ and /p/ in 5 types each. One example of a highly alliterating simile is (AS) DEAD AS A DODO, where the plosive /d/ occurs four times. However, such prominent examples of alliteration are marginal.

9.2.2 Assonantal similes

Assonance is arguably less striking than alliteration or rhythm. Nevertheless, it remains a factor contributing to a simile's memorability. The ground and the source in many similes are monosyllabic, but the prominence of assonance diminishes in longer similes as other vowels distort the sound pattern. Additionally, diphthong repetition is more prominent due to increased quantity. In a certain way, assonance can be considered an imperfect rhyme, as the syllable nucleus (vowel) repeats, but the coda does not match. Rhyming similes are not included in the list of assonantal similes, although they technically contain assonance.

The dataset yields 21 instances of assonance, resulting in 6.80% simile types marked with this prosodic feature. Table 32 shows the vowels occurring in assonantal similes along with their frequencies.

Vowel	Example simile	Frequency
/ɪ/	THICK AS PIG SHIT	6
/æ/	HAPPY AS A CLAM	5
/ɒ/	SOFT AS COTTON	4
/aɪ/	HIGH AS A KITE	2
/i/	EASY AS ABC	1
/eɪ/	PLAIN AS DAY	1
/əʊ/	COLD AS STONE	1
/uː/	COOL AS A CUCUMBER	1

Table 31. Vowels in assonantal similes.

The data do not reveal any prevalent pattern apart from short vowels being more frequent than long vowels and diphthongs. Six assonantal similes contain the vowel /ɪ/, five /æ/, and four /ɒ/. The remainder of the list are long vowels or diphthongs save for /i/, albeit with lower frequencies. A prominent example of assonance is (AS) EASY AS ABC, with /i/ occurring four times. The quantity of /i/ is disregarded, as the vowel quality is considered the primary factor determining assonance.

Lastly, two similes in the dataset combine alliteration and assonance: (AS) COOL AS A CUCUMBER and (AS) FIT AS A FIDDLE. The former's onset does not match (/kuː/ vs /kjuː/), making its prosodic form somewhat less striking. The latter exhibits the onset and nucleus repetition (/fɪ/), resulting in a memorable form.

9.2.3 Rhyming similes

Rhyme is the most prominent prosodic feature in adjectival similes. The operational definition of rhyme here is an exact repetition of the content following the onset of the stressed syllable in the ground. The specificity of rhyme makes it a rare feature, as evidenced by only five rhyming similes (1.62%) in the whole dataset.

Four of the rhyming similes have a monosyllabic ground and a disyllabic source comprising an article and a monosyllabic noun: (AS) DRUNK AS A SKUNK, (AS) HIGH AS THE SKY, (AS) SNUG AS A BUG and (AS) THICK AS A BRICK. Moreover, the poetic rhythm of these similes is a choriambus. The fifth rhyming simile is (AS) HIGH AS AN ELEPHANT'S EYE, whose extensive pentasyllabic source makes it an unusual example. However, the primarily stressed syllable in the source is the monosyllabic *eye*, which rhymes with the ground *high*.

10 Content analysis of adjectival similes

This section is dedicated to a content analysis of the adjectival simile sample. The semantic classifications are based on the focal constituent, making them either ground-based or source-based.

10.1 Ground-centred semantic classification

Semantic classifications of phraseological units are often problematic. Prototypical representatives can give the impression of seemingly fitting categories, but the applicability of those categories comes into question with less straightforward examples. The categories presented here are based on clustering adjectival similes sharing a generalised semantic feature expressed by the ground, such as 'colour'. The categories do not represent an exhaustive list of all possible semantic themes but rather a summary of the most prominent features observed in the dataset. It should also be noted that some similes can belong to multiple categories. In such cases, the similes are only assigned to a single group according to its most prominent feature.[101] The frequencies listed for the similes are the number of formal types.

10.1.1 Colour or light

One frequently recurring theme in adjectival similes is colour, as a digital description of colour is somewhat problematic. Most speakers learn colours at an early age. The process usually involves naming the colour of its prototypical bearer, for example, 'blue-sky', 'green-grass' or 'yellow-sun'. Most speakers cannot describe colours in terms of wavelengths, as such knowledge is beyond functional. Therefore, we resort to analogical descriptions of colours, using prototypical bearers as sources.

Within the dataset, five basic colours occur in 30 unique adjectival similes: *black, blue, green, red,* and *white.* Table 32 illustrates the 'colour' grounds along with their sources.

Ground	Source
black	*the ace of spades, coal, a crow, ink, midnight, (the) night, pitch, a raven('s wing), soot, thunder*
blue	*the sky*

---

[101] Ground-centred and source-centred classifications are considered separately. Therefore, each simile can occur once in each of the two general classifications.

green	*grass*
red	*a beet, a beetroot, blood, a tomato*
white	*alabaster, bone, chalk, death, a ghost, marble, milk, paper, (a) sheet(s), (the) snow*

Table 32. 'Colour' grounds and their sources.

The colours *black* and *white* occur in 12 types each, suggesting that *black* and *white* remain the two prototypical colours in terms of human cognition.[102] Additional colours or sources for the presented colours appear in the CAS transformations, such as JET(-)BLACK, RUBY(-)RED or SEA(-)BLUE.

Another related group of similes contains grounds referring to light and its intensity (dark, bright) or the absence of natural colour (*pale*). Table 33 presents these grounds with their sources.

Ground	Source
bright	*day, the Sun*
dark	*(the) night, the night sky*
pale	*death, a ghost, the moon*

Table 33. 'Light' grounds and their sources.

As shown by their sources, the grounds *dark* and *pale* are closely related to *black* and *white*, respectively.

Overall, colour is a prominent semantic theme (cf. Norrick, 1986; Moon, 2008). It is readily noticeable, but the data do not suggest any significant prevalence.

10.1.2 Character or behaviour

Another commonly recurring theme is a character or behaviour description. These two concepts are intertwined as one's behaviour typically mirrors one's personality traits. Furthermore, it is sometimes difficult to distinguish between a single occurrence and a repeating pattern of a particular behavioural feature. Moon (2008: 8) uses the term 'propensity' and includes descriptions of inanimate targets. The category of character or behaviour presented here is restricted to animate targets, which are predominantly people.

As most similes describe people, this category includes many similes. Common attributes described by these similes are intelligence, attitude or various facets of behaviour, such as courage, honesty or reliability. These similes are further subdivided

---

[102] The primary colours taught to children are typically *red, blue* and *yellow.* In physics, the three primary colours are *red, blue* and *green.* However, additional primary colour modules can be found in other fields.

according to their affective polarity into inherently positive (*n*=25) and inherently negative (*n*=32). Six types remain neutral, or their polarity is determined extrinsically by the context.[103] Table 34 lists the positive 'character/behaviour' grounds with their possible sources.

Ground	Source
bold	*a lion*
brave	*hope, a lion*
bright	*a button*
cool	*a cucumber*
free	*a bird, the wind*
gentle	*doves, a lamb*
good	*gold*
harmless	*doves*
innocent	*doves*
keen	*mustard*
pure	*the driven snow, gold, snow*
regular	*clockwork*
sharp	*a razor, a tack*
smart	*a whip*
steady	*a rock*
straight	*an arrow, a die*
sweet	*candy, honey, pie, sugar*

Table 34. Positive 'character/behaviour' grounds and their sources.

Positive polarity is usually inherited from the adjective. However, in adjectives without strictly delineated polarity, such as *bold*, the source ultimately determines the simile's polarity (compare positive (AS) BOLD AS A LION and negative (AS) BOLD AS BRASS). The 'character/behaviour' grounds in positive types strongly favour specific sources. These adjectival similes are used to praise, compliment, or describe a person either favourably or approvingly.

Ground	Source
bold	*brass*
crazy	*a loon*
daft	*a brush*
dappy	*pig shit*

---

[103] Pragmatic modifications, such as sarcasm, irony, or mockery, are not considered.

deaf	*(a) post(s)*
drunk	*a lord, a skunk*
dull	*dishwater*
dumb	*a box of hair, (a) (box of) rock(s), bricks, dirt*
mad	*a hatter, a march hare*
proud	*a peacock, Punch*
slippery	*an eel*
smooth	*butter, silk*
stiff	*a poker*
stubborn	*a mule*
thick	*a brick, pig( )shit, two short planks*
weak	*a kitten*
wise	*(a) serpent(s)*

Table 35. Negative 'character/behaviour' grounds and their sources.

Negative 'character/behaviour' similes (Table 35) also inherit negative polarity from the adjective. Moreover, it is often reinforced by the incongruence of the source and the ground, as in (AS) DAFT AS A BRUSH or (AS) THICK AS TWO SHORT PLANKS. A notable source variation occurs in the similes negatively describing a person's intelligence (*dumb, thick*).

The last group of similes are considered neutral, as they either include adjectives without clearly determined inherent polarity or require context for its interpretation. Table 36 provides an overview of neutral 'character/behaviour' simile types.

Ground	Source
busy	*(a) bee(s)*
hard	*nails*
thick	*thieves*
tough	*nails, old boots*

Table 36. Neutral 'character/behaviour' grounds and their sources.

Busy is inherently neutral, and the simile (AS) BUSY AS A BEE can have positive, neutral and negative connotations, depending on the amount of work and its perceived desirability. The simile can be used approvingly when said of a third party but also as an expression of dismay or frustration at the amount of work one needs to undertake. Naturally, it may simply mean 'very busy' without any affectivity. The polarity in similes describing one's mental toughness, such as (AS) TOUGH AS NAILS, varies with context. Such a quality may be seen as desirable as a display of endurance or inappropriate as a

146

form of reservation or hostility. (AS) THICK AS THIEVES is positive when referring to closeness and friendship but may take on negative connotations due to *thieves*' inherent negative polarity.

Lastly, some 'colour' adjectival similes might also be considered part of the 'character/behaviour' category, such as (AS) GREEN AS GRASS. These are not included.

10.1.3 Appearance or physical description

This theme clusters similes primarily describing the human body. Same as character or behaviour descriptions, they can be subdivided into positive (*n*=14), negative (*n*=13) and neutral (*n*=16) according to their affective polarity. Table 37 presents positive 'appearance' grounds together with their sources.

Ground	Source
clean	*a whistle*
fit	*a fiddle*
neat	*a pin*
pretty	*a picture*
strong	*a bear, a bull, Hercules, a horse, iron, a lion, an oak, an ox, steel*
tight	*a drum*

Table 37. Positive 'appearance' grounds with their sources.

Positive 'appearance' similes refer to physical and aesthetic features. The similes with *fit* and *strong* describe good constitution and physical prosperity. *Strong* is the only 'appearance' ground occurring in multiple similes whose meaning and function can overlap.[104] *Clean, neat* and *pretty* refer to visually appealing features related to neatness (of appearance) and physical beauty. The simile (AS) TIGHT AS A DRUM is rather peculiar, as it often refers to female buttocks when describing the human body.

Ground	Source
blind	*a bat*
fat	*a pig*
flat	*a board, a pancake*
skinny	*a rail, a rake*
thin	*a rail, a rake, paper*
ugly	*sin*
stiff	*a board, a poker*

---

[104] This is determined by the source's connotations. Naturally, *bull, horse, lion* and *ox* might have similar connotations, while *death* will undoubtedly trigger a different facet of *strong*.

weak	*a kitten*

Table 38. Negative 'appearance' grounds with their sources.

Negative 'appearance' grounds (Table 38) refer to visually unattractive features or otherwise unfavourable physical conditions. The features described by these similes include blindness, thinness, or posture stiffness. These similes are often derogatory.

Ground	Source
big	*a barn, a basketball, a football field, a football stadium, (a) house(s), a mountain, saucers, Texas, a whale*
naked	*a jaybird, the day one came to this world, the day one was born*
tall	*a house, a tree*
wide	*saucers*

Table 39. Neutral 'appearance' grounds with their sources.

The last group are neutral 'appearance' similes (Table 39). Inherently, *big, naked*, *tall* and *wide* are neither positive nor negative. However, contextual modulation can commit these similes to the negative pole.[105] For example, *he was tall as a tree and intimidating as ever* is likely to receive a negative interpretation.

Lastly, some 'colour' adjectival similes could be included in the 'appearance' category, for instance, (AS) WHITE AS A GHOST. These are not included.

10.1.4 Sensation or emotion

This group clusters adjectival similes (*n*=45) expressing various sensations or emotions. The aspect of affective polarity is not relevant in these similes, despite some of them committing to either pole, as determined by the desirability and the result of that particular sensation or emotion. Table 40 presents 'sensation' grounds with their sources.

Ground	Source
cold	*ice, stone*
cool	*ice*
fresh	*a daisy*
happy	*(a) clam(s), a lark, Larry, a pig (in muck/mud/shit)*
healthy	*a horse*
high	*a kite*
hot	*balls, fire, hell, an oven, the Sun*

---

[105] While the positive pole cannot be ruled out, *naked* can trigger associations with exposure and vulnerability, *big* and *tall* with abnormality and excess growth, and (AS) WIDE AS SAUCERS is exclusive to *eyes*, typically describing shock or surprise.

hungry	*a wolf*
lonely	*a cloud*
mad	*a hatter, a hornet, a march hare*
miserable	*sin*
nervous	*a cat*
pissed	*a fart, a nook*
pleased	Punch
quiet	*a/the grave, a church (mouse), a mouse, a tomb*
sick	*a dog, a god, a parrot*
silent	*the grave*
sweet	*candy, honey, pie, sugar*
warm	*toast*

Table 40. 'Sensation' grounds with their sources.

'Sensation or emotion' similes describe sense perceptions (*cold* or *quiet*), feelings (*lonely* or *nervous*), and emotions (*happy* or *mad*). *Happy*, *hot, quiet, mad* and *sweet* combine with various sources, each possibly triggering different connotations.

10.1.5 Material quality or texture

The last ground-centred semantic group of similes (*n*=35) describes a material or its texture, including assessments of quality or condition. Table 41 lists 'material quality or texture' grounds with their respective sources.

Ground	Source
hard	*concrete, diamond(s), granite, iron, (a) rock(s), steel, stone*
precious	*gold*
right	*rain*
sharp	*a knife, knives, (a) razor(s)*
smooth	*a baby('s ass/behind/bottom/bum/butt), butter, glass, silk, velvet*
soft	*a baby's (behind/bottom/butt), butter, cotton, silk, velvet*
solid	*(a) rock*

Table 41. 'Material quality or texture' grounds with their sources.

The ground *hard* combines with many sources, but the meaning typically remains 'very hard', although the sources' connotations might differ in some situations. *Smooth* and *soft* show significant semantic overlap, as illustrated by their preferred sources. However, *soft* often includes the sense of 'tender', which disqualifies the ground *glass*. *Precious* and *right* refer to fineness and quality.

10.2 Source-centred classification

As with the ground-centred classification, the source-centred ones cluster adjectival similes sharing a generalised semantic feature expressed by the source. The semantic feature can be a concrete concept, such as 'animal', or an abstract one, such as 'history'. It is not easy to design specific categorisations for abstract concepts, as they are primarily products of human conceptualisation of the world and aspects within it. Therefore, abstract sources are excluded from the categorisation presented here.

The source-centred classification is less fragmented than the ground-centred one, suggesting that the source slot is more limiting regarding its potential representatives. This aligns with the premise that the source is usually a familiar entity. After all, similes utilise analogies, and using unfamiliar sources would be counterproductive. Of the 309 adjectival simile types collected, 246 (79.61%) fit into one of the three categories presented in this subsection.

10.2.1 Animals

The category of animals is perhaps the most striking semantic theme shared by many similes. In the dataset, 54 (17.48%) simile types include an animal source. Thirty-four unique animals or their body parts appear in the similes. Interestingly, only 14 of those are potentially domesticated animals, where we can expect greater familiarity with the animal.

Varying physical and behavioural features are typically ascribed to person targets, with animals functioning as good sources due to animacy and general familiarity. It should be mentioned that the ascribed feature need not be a generally salient feature of the source based on our contemporary knowledge of the animal kingdom. Table 42 shows the list of 'domesticated animal' sources and their associated grounds.

Source	Ground
a bull	*strong*
a cat	*nervous*
a dog('s hind leg)	*crooked, sick*
an elephant's eye	*high*
hen's teeth	*rare*
a horse	*healthy, strong*
a kitten	*weak*
a lamb	*gentle*

a mule	*stubborn*
mutton	*dead*
an ox	*strong*
a parrot	*sick*
a peacock	*proud*
(a) pig (in muck/mud/shit)	*fat, happy, thick*

Table 42. 'Domestic animal' sources with their grounds.

The list of 'domesticated animal' sources suggests that each animal is usually associated with a single prominent feature and does not occur in other simile types. However, some animals may share a single prominent feature, such as *bull, horse* or *ox* share the feature *strong*.

Source	Ground
a bat	*blind*
a bear	*strong*
(a) bee(s)	*busy*
a bird	*free*
a bug('s ear)	*cute, snug*
(a) clam(s)	*happy*
a crow	*black*
dodo	*dead*
doves	*gentle, harmless, innocent*
an eel	*slippery*
a hornet	*mad*
a jaybird	*naked*
a lark	*happy*
a lion	*brave, bold, strong*
a March hare	*mad*
a rat	*quick*
a raven('s wing)	*black*
(a) serpent(s)	*wise*
a skunk	*drunk*
a whale	*big*
a wolf	*hungry*

Table 43. 'Wild animal' sources with their grounds.

The list of 'wild animal' sources (Table 43) shows a similar trend of individual animals being associated with a single feature. *Doves* and *lion* occur in multiple unique simile types, but their grounds are at least near-synonymous.

10.2.2 Natural entities or material

The group of adjectival similes labelled 'a natural entity or material' is designed for sources represented by things and phenomena occurring in nature without human contribution. Since the category is very general, it clusters 94 similes (30.42%) from the dataset and includes multiple possible subcategories, some of which are presented in Table 44.

Subcategory	Example simile
element	HOT AS FIRE, FREE AS THE WIND, CLEAR AS WATER
geographical entity	BIG AS A MOUNTAIN, HIGH AS THE SKY, OLD AS THE HILLS
material	BLACK AS COAL, CLEAR AS CRYSTAL, PURE AS GOLD
plant	FRESH AS A DAISY, RED AS A BEET, TALL AS A TREE
weather aspect	QUICK AS LIGTHNING, RIGHT AS RAIN, WHITE AS SNOW

Table 44. 'Natural entities or materials' subcategories with examples.

These subcategories are by no means definitive, and many other combinations are possible. The subcategory of 'element' is inspired by the Japanese philosophy Godai, where the four basic elements (earth, fire, water and wind) represent the possible manifestations of matter. The subcategory of 'material' can be further divided into minerals (*marble*), rocks (*stone*) or metals (*iron*). In 'plants', we could distinguish between fruit-bearing (*tomato*) and general (*tree*).

The purpose of the category 'natural entity or material' is not scientific accuracy but rather the grouping of various sources within a generally understood theme, regardless of its scientific validity.

10.2.3 Human-made objects

The 'human-made object' category consists of 98 (31.72%) adjectival similes from the dataset whose source results from human interference or contribution. This category is likewise extensive and can be subdivided into more specific groups (Table 45).

Subcategory	Example simile
construction/arrangement	BIG AS A BARN, HIGH AS A HOUSE, QUIET AS THE GRAVE
food	EASY AS PIE, FLAT AS A PANCAKE, WARM AS TOAST

material	CLEAR AS GLASS, DUMB AS BRICKS, STRONG AS STEEL
product	CLEAR AS A BELL, SKINNY AS A RAIL, STRAIGHT AS A DIE
tool	CLEAN AS A WHISTLE, NEAT AS A PIN, SHARP AS A KNIFE

Table 45. 'Human-made objects' subcategories with examples.

These subcategories illustrate a form of specification without going too deep into the technical levels of differentiation. The category 'product' could include the categories 'food' and 'tool' if we applied the broad definition 'produced by people'. The category 'construction or arrangement' could be split into two, and the category of 'food' could be further subdivided into meal (*pie*), ingredient (*butter*), and drink (*wine*).

11 Comparison of corpus-extracted English and Czech adjectival similes

This section presents a semantic comparison of English adjectival similes extracted in this research with a list of Czech adjectival similes. The comparison is by no means exhaustive and represents just a brief survey of similarities and differences observed in English and Czech adjectival similes. The major semantic categories are compared regarding frequencies, followed by descriptions of the most frequent equivalent similes from the sets.

11.1 The compared lists

The list of Czech similes (Appendix 2) was kindly made available by M. Hnátková and V. Petkevič. It was compiled for the purposes of the Czech electronic database of MWEs, LEMUR, using data from Čermák et al. (1983–2009), FRANTALEX (a list of phrasemes and collocations compiled by M. Hnátková) and two Czech SYN corpora (Křen et al. 2015, 2019). After revision, the list comprises 886 adjectival simile types, which is considerably more than the list of English similes (309 types).

One general observation can be made regarding the transformational variability in the source slot in Czech similes. Many grounds in the list appear with a great number of sources. This variability does not appear in the English similes collected from the researched corpora. However, the list of Czech similes is a combined product of corpus data (the SYN corpora) and dictionary data. It is important to note that dictionaries typically inherit existing word lists and update them by adding new items. Many items become obsolete over time and may no longer be attested in present-day corpora, but they are rarely removed from the dictionaries.[106] Consequently, the list of Czech similes naturally contains more types as it represents the typical 'dictionary bloating'.

Lastly, the Czech set does not list any frequencies for individual similes, so their actual usage remains unknown. The English language contains many other adjectival similes not occurring in the 309-item dataset, and complementing the corpus data with, for instance, Sommer's Similes Dictionary (2013) would result in a vast list of adjectival simile types. However, such lists provide little information about simile usage in present-day languages.

---

[106] Common practice is explicitly describing them 'dated' or 'obsolete'.

It follows from these considerations that drawing conclusions regarding the colourfulness of similes based on this comparison has limited validity. Nevertheless, the comparison can hint at similarities and differences in the features described by adjectival similes in English and Czech and provide grounds for further research.

11.2 Grounds

The survey shows that the ground categories presented in Section 10.1 apply to Czech similes as well. Table 46 presents an overview of ground-centred categories along with the type frequencies for the English and Czech simile types in this research.[107]

Category	English	Czech
Colour or light	30 (9.709)	95 (10.722)
Character or behaviour	63 (20.388)	197 (22.235)
Appearance or physical description	43 (13.916)	246 (27.765)
Sensation or emotion	45 (14.563)	102 (11.512)
Material quality or texture	35 (11.327)	109 (12.302)
Intoxication	3 (0.971)	28 (3.160)

Table 46. English and Czech ground-centred category distribution.

As the numbers show, 99 English and 108 Czech adjectival similes are not assigned any of the designed ground-centred categories due to the categories either not fitting the simile properly or the simile's semantic or functional peculiarity. This is not considered an issue since the categorisation is mostly for illustrative purposes. Apart from 'appearance or physical description' and 'intoxication', the categories appear to be represented evenly in English and Czech, suggesting a great cultural and cognitive overlap.

One additional category appeared prominent in the Czech set: 'intoxication'. It could arguably be treated as a type of sensation, but its recurrence and peculiarity deserve a separate category. Whereas the English set provides only three expressions describing a form of intoxication, (AS) DRUNK AS A LORD, (AS) DRUNK AS A SKUNK and (AS) HIGH AS A KITE, the Czech list contains 28 types with various grounds and sources, for instance, NALITÝ JAK(O) SUD [poured as barrel], OŽRALÝ JAK(O) ČUNĚ [drunk as pig] or ZMALOVANÝ JAK(O) INDIÁN [war-painted as Amerindian].

---

[107] The brackets show relative frequencies of simile types per one hundred (t.p.h.) to illustrate how frequent the individual categories are within the samples.

11.3 *Black/černý* and *white/bílý*

In Czech similes, the category of 'colour' (*n=95*; 10.970 t.p.h.) shows a similar representation as in English similes (*n=30*; 9.709 t.p.h.). In both languages, the dominant colours are *black* and *white*. The Czech set also contains 13 simile types with *červený* [red]. Other colours remain marginal in both languages, but the Czech set shows more variability (16 unique grounds) compared to English (8 unique grounds). Table 47 presents a comparison of the grounds *black* and *white* with their possible sources in both languages.

Ground	Source
black	*the ace of spades, coal, a crow, ink, midnight, (the) night, pitch, a raven('s wing), soot, thunder*
***černý* [black]**	*antracit, ašant, bota, cikán(ka), čert, ďábel, eben, havran, havraní křídla, hřích, inkoust, kolomaz, mouřenín, noc, saze, smola, smůla, uhel, vrána*
white	*alabaster, bone, chalk, death, a ghost, marble, milk, paper, (a) sheet(s), (the) snow*
***bílý* [white]**	*(z) alabastr(u), křída, lilie, mléko, mramor, mrtvola, padlý sníh, papír, sejra, slonová kost, smrt, sníh, stěna, tvaroh, vápno, zeď*

Table 47. The comparison of black/černý *and* white/bílý *grounds.*

The ground *black* appears in 12 unique types in English and 20 in Czech. The source identity occurs in nine examples: *a raven/havran*[108], *a raven's wing/havraní křídla, ink/inkoust, (the) night/noc, soot/saze, pitch/smola, pitch/smůla, coal/uhel* and *crow/vrána*. The remaining sources do not have attested lexical equivalents in the sets.

The ground *white* offers a similar picture. The English set contains 12 unique types, and 17 occur in the Czech set. The source identity is observed in seven cases: *alabaster/(z) alabastr(u), chalk/křída, milk/mléko, marble/mramor, paper/papír, death/smrt* and *(the) snow/sníh*. The remaining sources do not have matching lexical equivalents in the sets.

The languages show a significant degree of overlap while maintaining a number of sources unique to either of them. The connotations of lexical equivalents are expected to match, but this would require access to the Czech similes' concordances to verify.

---

[108] *Raven* is typically translated into Czech as *havran*, despite the accurate translation being *krkavec*. Therefore, it is still considered an identity.

11.4 Strong/silný

Adjectival similes with the grounds *strong/silný* primarily describe physical strength and fall in the category of 'appearance or physical description'. This category appears significantly more frequently in the Czech similes (*n=246*; 27.765 t.p.h.) than the English ones (*n=43*; 13.916 t.p.h.).

Adjectival similes with *strong/silný* are a fairly large group, clustering 16 Czech and 10 English types. Most of the similes belong in this semantic category with the exception of (AS) STRONG AS DEATH and its Czech equivalent SILNÝ JAK(O) SMRT. There belong in the category 'sensation or emotion', as they describe either the taste (typically of coffee) or the intensity of a feeling.

Ground	Source
strong	*a bear, a bull, death, Hercules, a horse, iron, a lion, an oak, an ox, steel*
silný	*Bivoj, býk, dub, golem, Herkules, hrom, kobyla, kůň, lev, lvi, medvěd, noha, Samson, smrt, tur, z oceli*

Table 48. The comparison of strong/silný *grounds.*

Most of the similes in Table 48 refer to raw strength, but their conceptualisation of strength differs according to the source. Three major groups of conceptualisations could be described as bulkiness (*oak*), toughness (*steel*) and muscularity (*horse*).

Eight of the similes exhibit source identity: *a bear/medvěd, a bull/býk, death/smrt, Hercules/Herkules, a horse/kůň, a lion/lev, an oak/dub, steel/z oceli*. 'Animal' sources are frequent in both languages as bearers of physical strength, complemented by mythical figures (*Bivoj, Hercules, Samson*). The other groups are materials (*iron, steel*) and other unique phenomena (*oak* or *hrom* [thunder]).

Again, the degree of overlap is relatively high for similes with *strong*. Some lexical variation occurs, but the general categories remain mostly identical.

11.5 Big/velký

Adjectival similes with the grounds *big/velký* are another large group primarily from the category 'appearance or physical description'. The data suggest that compared to the rest of the types, *big*-similes are more prevalent in English (*n=16*; t.p.h. 5.179) than in Czech (*n=17*; 1.919 t.p.h.). The English similes (AS) BIG AS LIFE and (AS) LARGE AS LIFE

are the only ones not describing physical size, as they are used to describe one's surprise at seeing someone in person.

Ground	Source
big	*a barn, a basketball, a field, a football (field/stadium), (a) house(s), a mountain, a whale, life, marbles, saucers, Texas, the sky*
large	*life*
velký	*almara, Brno, holubí vejce, hrachy, kobyla, kolo od vozu, kráva, medvěd, náklaďák, pěst, pětník, skřivánek, tele, vejce*
veliký	*pěst, slon, stodola*

Table 49. The comparison of big/velký *and related grounds.*

Table 49 provides an overview of the grounds *big/velký*, complemented by *large* and *veliký* [big] since they are closely related. One immediate observation is that there is almost no overlap between the English and the Czech sources. Only a single source is identical: *a barn/stodola*. This is perhaps due to *big* being semantically underspecified and, thus, contextually dependent. Sources such as *pětník* [penny or nickle] or *skřivánek* [lark] are likely to be humorous, as they are used rather ironically. The remaining sources represent entities of varying size (compare *basketball* with *mountain*) due to different target preferences.

Similes with the ground *big* have limited dictionary coverage in English, perhaps due to theoretical issues. The criteria used for the simile diagnosis will affect whether many *big*-similes qualify as similes or just literal comparisons. The recurrence of these patterns in corpus data suggests that these are standard similes, further supported by the concordances. This conclusion can also be applied to Czech similes, but it would require thorough data analysis to confirm.

11.6 *Smooth/hladký*

Adjectival similes with the grounds *smooth/hladký* belong to the category 'material quality or texture'. They are presented together with the similes containing the English ground *soft* and the Czech grounds *hebký* [smooth or soft] and *měkký* [soft or tender]. The reason for including these grounds is a significant semantic overlap, as illustrated by the preferred sources in Table 50. This English cluster contains 17 types (5.502 t.p.h.), while 26 types (2.935 t.p.h.) occur in the Czech group.

Ground	Source
smooth	*a baby('s ass/behind/bottom/bum/butt), butter, glass, silk, velvet*
soft	*a baby's (behind/bottom/butt), butter, cotton, silk, velvet*
hladký	*alabastr, dětská prdelka, had, hedvábí, led, mejdlo, mramor, mýdlo, oblázek, olej, samet, sklo, stůl, úhoř, zrcadlo*
hebký	*hedvábí, samet*
měkký	*hedvábí, houba, koberec, kobereček, máslo, mech, samet, tvaroh, vosk*

Table 50. The comparison of smooth/hladký *and related grounds.*

While it could be argued that *smooth* and *soft* have different meanings, the similes appear to prefer identical grounds. Only *glass* and *cotton* occur with *smooth* and *soft*, respectively, making them semantically different from the rest.

The five overlapping sources are *a baby's bottom/dětská prdelka, butter/máslo, glass/sklo, silk/hedvábí* and *velvet/samet*. English has similes with many source variants equivalent to the single Czech simile HLADKÝ JAK(O) DĚTSKÁ PRDELKA. With the ground *smooth*, the variants are *a baby, a baby's ass, a baby's behind, a baby's bottom, a baby's bum* and *a baby's butt*. For *soft*, the attested sources are *a baby's, a baby's behind, a baby's bottom* and *a baby's butt*. The data suggest that the source variation in these similes is not conditioned by the ground. Czech has many variants for the word *prdelka*, but none of them occurs with *hladký* in a simile. This shows that the diminutive *prdelka* is a fixed preference in the Czech simile and typically does not allow other variants.

While the Czech similes contain sources not attested in the English data, the fabrics *silk/hedvábí* and *velvet/samet* are perhaps the best sources to illustrate the undisputed connection among these similes. The English sources occur with both *smooth* and *soft*, and the Czech are attested with all three grounds, *hladký, hebký* and *měkký*.

Lastly, the Czech simile HLADKÝ JAK(O) ÚHOŘ is somewhat peculiar. Its English equivalent is (AS) SLIPPERY AS AN EEL. Both similes share the meaning 'deceptive' or 'treacherous' along with (AS) SMOOTH AS SILK. The Czech set contains other semantically synonymous similes: SLIZKÝ JAK(O) HAD, ÚLISNÝ JAK(O) HAD, ÚSKOČNÝ JAK(O) HAD, ZRÁDNÝ JAK(O) HAD. Another related English simile is (AS) WISE AS SERPENTS. These similes belong to the category 'character or behaviour'.

11.7 *Hard/tvrdý*

The last grounds analysed here are *hard/tvrdý*, complemented by *tough*. Most belong to the ground-centred category 'material quality or texture'. However, the similes (AS)

HARD AS NAILS, (AS) TOUGH AS NAILS and TVRDÝ JAK(O) HŘEBÍKY[109] rather belong to the category 'character or behaviour'.

Ground	Source
hard	*concrete, diamond(s), granite, iron, nails, (a) rock(s), steel, stone*
tough	*nails, old boots*
tvrdý	*beton, brok, hřebíky, kámen, kost, křemen, mlat, oblázek, ocel, podrážka, roh, skála, suk, z křemene, z oceli, železo, žula*

Table 51. The comparison of hard/tvrdý *grounds*

Table 51 present the 13 (4.207 t.p.h.) English and 17 (1.919 t.p.h.) Czech adjectival similes with the grounds *hard, tough* and *tvrdý* [hard]. The overlapping sources (*n*=8) are *concrete/beton, granite/žula, iron/železo, nails/hřebíky, a rock/kámen, rock/skála, steel/ocel* and *stone/kámen*.

One striking semantic field is minerals, represented by the sources *diamond, granite/žula, stone/kámen, křemen* [quartz] and *oblázek* [pebble]. Another group are alloys (*iron/železo[110], steel/ocel*) and other sturdy materials or products (*concrete/beton, nails/hřebíky, brok* [pellet]). Additionally, the Czech similes contain *kost* [bone]*, roh* [horn], *mlat* [threshing floor] and *suk* [knot]. The Czech source *podrážka* [sole] is related to the English source *old boots*.

11.8 Animal sources

The last compared category is adjectival similes comprising animal sources. This category is convenient as 'animal' similes are striking and easily diagnosed. The English list yields 54 instances (17.476 t.p.h.) of similes with an animal source, and the Czech list contains 264 unique types (29.797 t.p.h.). The relative frequency suggests that 'animal' adjectival similes are more prevalent in Czech.

Rank	English	Source types	Czech	Source types
1	pig	6	pes	10
2	dove	3	prase	10
3	lion	3	had	8
4	bug	2	opice	8
5	horse	2	kotě	5

---

[109] A Google Search does not find any examples of the simile. Establishing its actual meaning without the source text is impossible, but it is likely a translation of the standard English simile (AS) HARD AS NAILS. Consequently, its meaning is assumed to be 'physically or mentally tough'.

[110] In (AS) HARD AS IRON, *iron* typically refers to alloys comprising iron rather than the actual element.

Table 52. The most versatile animal sources in English and Czech.

Table 52 shows an overview of the top five animal sources in both lists. The Czech list could also be complemented by examples of diminutives, as they are often directly related to non-diminutive forms. However, the connotations of diminutives tend to be different from the non-diminutive forms since they are usually positive.

11.8.1   English animal sources

Pig/prase is the only animal appearing in both top five lists. In English, *pig* occurs in six simile types with three different grounds: *fat, happy* and *thick*. In Czech, the source *prase* is more complicated due to its occurrence in the intensifying post-modification (Emmer, 2020).[111] Ten unique similes comprise the source *prase*, each having a unique ground: *červený* [red], *líný* [lazy], *ožralý* [drunk], *špinavý* [dirty], *těžký* [heavy], *tlustý* [fat], *velký* [big], *vypasený* [chubby], *vyžraný* [porky] and *zlitý* [drunk]. The grounds illustrate that *prase* has exclusively negative connotations in Czech. The English grounds *fat* and *thick* are also inherently negative, but *happy* is primarily positive. The overlap is observed only in *fat/tlustý*. Czech has several grounds describing larger body mass (*tlustý, vypasený, vyžraný*), but *prase* in Czech adjectival similes is not associated with intelligence, as in the English simile (AS) THICK AS PIG( )SHIT.

The second animal source from the English list is *dove.* It appears with the grounds *gentle, harmless* and *innocent*, and it is always in the plural. The Czech list contains only two simile types with dove that are directly related and contain the same ground *mírná* [placid], MÍRNÁ JAK(O) HOLUBICE and MÍRNÁ JAK(O) HOLUBIČKA.

The third English animal source, *lion*, appears with the grounds *bold, brave* and *strong*. The Czech set comprises only two similes with this animal, HLADOVÝ JAK(O) LEV [hungry as lion] and SILNÝ JAKO LEV [strong as lion]. This shows that *lion* in English is associated with positive attributes, whereas in Czech, it can also be rather negative.

Animal number four in English is *bug*. It occurs with two unique grounds, *cute* and *snug*. The Czech list contains no simile with *bug*. Only a single simile in the Czech set contains a type of bug, OTRÁVENÝ JAK(O) ŠVÁB, but its meaning and connotations are completely unrelated to the English examples.

The last English animal source is *horse*. In English, it occurs with two unique grounds: *healthy* and *strong*. In Czech, three grounds are associated with horses in similes,

---

[111] Several items with *prase* had to be excluded from the Czech adjectival simile list since they are rather examples of IP (intensifying post-modification), for instance, *hořký jako prase* or *drahý jako prase*.

exhibiting partial overlap: *silný* [strong], *uštvaný* [worn out] and *utahaný* [worn out]. Same as with *lion*, English perception of *horse* is positive, whereas Czech also utilises the horses' exertion in farming.

11.8.2   Czech animal sources

Due to the larger size of the dataset, some Czech animal sources appear with more unique grounds. However, this trend should not be overstated, as many animal sources still remain exclusive to single similes.

One of the two most versatile sources in the Czech set is *dog* [pes]. It occurs with ten unique grounds: *hladový* [hungry], *hubený* [thin], *křivý* [crooked], *opuštěný* [abandoned], *platný* [useful], *utahaný* [worn out], *uvázaný* [tied], *věrný* [loyal], *vychrtlý* [skinny], *vyzáblý* [scraggy], *vzteklý* [furious] and *zlý* [evil]. In three of the instances, *dog* occurs as part of a complex source, for instance, PLATNÝ JAK(O) PES V KOSTELE [useful as dog in church]. In the English set, *dog* only occurs with two grounds: *crooked* and *sick*. Apart from *věrný*, *dog* appears to be associated with negative grounds in both languages. The other source occurring with ten unique grounds in Czech is *prase*, which was already described before.

Another Czech source *had* [snake], appears with eight unique grounds: *hladký* [smooth], *lstivý* [cunning], *mrštný* [agile], *slizký* [sleazy], *studený* [cold], *úlisný* [slimy], *úskočný* [deceitful] and *zrádný* [treacherous]. In the English set, only the simile (AS) WISE AS SERPENTS contains an equivalent of *had*. It is perhaps for the biblical depiction of snakes that they primarily combine with negative grounds.

Opice [monkey] is another source occurring with several unique grounds: *drzý* [cheeky], *chlupatý* [hairy], *chytrý* [clever], *mrštný* [agile], *ošklivý* [ugly], *škaredý* [ugly], *učenlivý* [quick to learn] and *zvědavý* [curious]. The source *monkey* in Czech is associated with positive and negative grounds. The English set does not contain any simile with the source *monkey*, likely owing to its racially offensive metaphorical use. The racial aspect is not prominent in present-day Czech.

The last Czech animal source described here is *kotě* [kitten]. It appears with five unique grounds: *hravý* [playful], *mazlivý* [cuddly], *ospalý* [sleepy], *slepý* [blind] and *utahaný* [worn out]. The English set contains a single simile with the source *kitten*, (AS) WEAK AS A KITTEN, which describes a different aspect of baby cats compared to the Czech similes.

12 Discussion of findings

The last section of this work is dedicated to discussing the results, reviewing the hypotheses and considering the implications of this research. The hypotheses in this work and their testing were intended as complementary goals to the overall survey of adjectival similes in the selected corpora. The primary focus was to map adjectival similes in corpora, their annotation, and scrutinise the mining process.

Using the data mined from four English corpora, I collected a reliable sample of adjectival similes objectively reflecting the current state of their use in English, which was the primary goal of this thesis. By imposing a frequency limit to this representative sample, I then compiled and described what could be termed a 'similes minimum' (by analogy to the so-called paremiological minimum), in other words, a set of similes that can be regarded as the essential minimal knowledge of similes displayed by native speakers which learners of English should at least know, or ideally, be able to use.

Next, the data analysis provides a description of English adjectival similes from both a formal and a semantic point of view. I believe such a corpus-based description has not yet been put forward. It shows that despite being formally fixed, adjectival similes are sometimes difficult to find in corpus data by using generalised queries. One of the hypotheses, *the recall-precision ratio of adjectival similes is effective when using CQL pattern-queries*, reflected the premise that the formalised frame should simplify the adjectival simile retrieval. This hypothesis proved to be false, as shown in Section 6. The existing studies on similes rarely present the queries used for their retrieval. Constructing CQL queries is generally considered a rudimentary knowledge of corpus research, but the process is often more complex than initially anticipated.

Idiomatic multi-word units are sometimes challenging to extract due to their non-compositional nature. Unless the corpus contains phraseological tagging[112], we must work with the tagging of individual constituting elements. Lexical decomposition of idioms is generally unfeasible, but it is necessary for a corpus-based investigation. Consequently, annotation accuracy becomes a major factor affecting the return's recall and precision.

General queries return many irrelevant items, making the precision drop sometimes below 10%, and making the queries more restrictive does little to improve the precision.

---

[112] Phraseological tagging remains relatively rare, as it is fairly inaccurate when done automatically and extremely time consuming to do manually.

This resulted in general queries being used for retrieving initial simile lists. Subsequently, individual similes needed to be searched using lemma queries to bypass erroneous annotation. This often yielded additional instances of many similes. However, this requires an existing list of adjectival similes without resorting to guesswork. Such simile collection might prove difficult to obtain, as was already shown by dictionaries often providing us with incomplete lists.

To summarise, designing general queries for adjectival simile extraction is a complex task, and there is no ideal solution. General-query returns exhibit an extremely low precision, and highly restrictive queries result in poor recall. This makes mining novel similes challenging. The implications are that retrieving adjectival similes from corpora requires a meticulous multi-layered mining approach. Otherwise, the data will lack many instances.

The diagnosis of adjectival similes is not without problems, either. This work shows that traditional dictionaries do not recognise many frequent adjectival similes. The general hypothesis that *higher corpus frequencies of adjectival similes tend to result in their better dictionary coverage* assumed that corpus evidence would serve as the primary source of dictionaries. Therefore, the most frequent adjectival similes were expected to appear in the online dictionaries without exception. This was tested in Section 8 by reviewing the coverage of adjectival similes in several dictionaries, some of which are considered highly authoritative. Surprisingly, the results indicate that the hypothesis is false. Corpus frequencies do not appear to be a prominent factor in determining whether adjectival similes appear in a dictionary.

The reviewed dictionaries likely use their own corpora for data extraction, but the corpora researched in this work are believed to be representative of general trends related to adjectival similes. The BNC contains data up to 40 years old, while the COCA contains some data no older than four years. Consequently, the dataset can be described as containing adjectival similes used in English over the last 40 years. Naturally, other similes not extracted in this work are expected to be used in the language, albeit not very frequently. Those extracted with higher frequencies represent similes that are generally used, making their absence from some dictionaries peculiar. We can consider many reasons for the absence of adjectival similes in dictionaries and speculate about their importance, but one of them is, undoubtedly, that some adjectival similes are erroneously dismissed as literal comparisons.

The formal analysis of adjectival similes shows a strong preference for monosyllabic constituting elements. The majority of adjectival similes are dimetric and do not exceed four syllables. Longer similes exist but are significantly less frequent. The hypothesis that *prosodically marked adjectival similes do not occur more frequently than prosodically unmarked ones* proved true. The data show that every fifth simile is prosodically marked. However, these similes do not appear to be particularly frequent compared to the rest. It is impossible to assess how much prosodic features affect a simile's frequency individually, but the overall picture strongly suggests that prosodic features, generally, are not a factor determining a simile's frequency.

The content analysis of the sample illustrates several general semantic categories within which the similes can be clustered. One of the more prominent categories is the 'animal' similes. The popularity of 'animal' sources is not surprising. Animals lend themselves to comparisons with people, as they are living beings. Many are associated with a set of characteristic features which most speakers readily recognise. The features relate to character, behaviour, appearance or emotional states, and analogous descriptions of said features are often more apt than extensive digital explanations.

Based on existing evidence, the hypothesis was formed that *adjectival similes with 'animal' sources constitute a large group within the dataset*. In the collected sample, 'animal' similes represent 17.48%, which is significantly less than Norrick's 38% (1986: 40). However, the hypothesis can be considered valid. Unlike the other source-centred categories, the 'animal' one is very specific, and once we compare the 'animal' category to the other less inclusive subcategories, the former becomes the largest group by far.

Finally, the comparison of English and Czech similes, although based on samples of unequal length, suggests a significant overlap between the two languages as far as the ground and the source of the simile are concerned. The ground and source categories apply to both languages, and the lexical representation is often identical.

The overlap, as illustrated in Section 11, may have several explanations. Both languages inherited adjectival similes from the same sources, for instance, the Bible (e.g. (AS) OLD AS METHUSELAH/STARÝ JAK(O) METUZALÉM) or Ancient Greek mythology (e.g. (AS) STRONG AS HERCULES/SILNÝ JAK(O) HERKULES). Speakers of the two languages originally come from Europe. Therefore, they share much of the geographic (e.g. (AS) HARD AS GRANITE/TVRDÝ JAK(O) ŽULA), weather (e.g. (AS) FAST AS LIGHTNING) and other types of experience tied to the continent. This is related to one of the Conceptual Metaphor Theory's pillars – similes, too, are grounded in experience.

Additionally, adjectival similes, again much like metaphors, are cognitive phenomena. Human cognition determines our perception of similarities between things. This extralinguistic dimension is shared by people regardless of their mother tongue. This is not to argue that language or national history do not play a prominent role (e.g. (AS) COMMON AS MUCK or VELKÝ JAK(O) BRNO), but some comparisons simply result from prominent extralinguistic realities and are shared across many languages (e.g. (AS) BLACK AS NIGHT/ČERNÝ JAK(O) NOC).

The hypothesis that *frequent English adjectival similes have lexical equivalents in Czech* proved only partly true. There is significant lexical overlap in both the ground and the source slots, but some frequent English similes do not have a lexical equivalent in Czech (e.g. (AS) RIGHT AS RAIN or (AS) EASY AS ABC) and vice versa. These are usually products of national heritage, for instance, films, songs, anecdotes, and popular commercials, but perhaps also social or political issues.

In light of the findings presented in this work, it is understandable that adjectival similes attract much less attention than, for example, metaphors. They are considered infrequent, their mining is complex, and their dictionary coverage is lacking. Theoretically, they are often dismissed as explicit metaphors, fancy literary expressions, or mere literal comparisons.

This work is by no means exhaustive, and many aspects of adjectival similes require further research. For instance, the common conception that similes are primarily aesthetic literary devices requires extensive data analysis. The similes extracted in this work suggest that adjectival similes are frequently used in spoken English, and some appear almost exclusively colloquial ((AS) COMMON AS MUCK or (AS) HARD AS A ROCK).

Another aspect is the exclusiveness of adjectival similes regarding regional variants. This work occasionally comments on regional preferences of certain similes, but it offers no complex account of regional variation. The dataset hints at many similes being popular in specific English-speaking countries.[113] However, more data would need to be analysed to obtain conclusive evidence.

Adjectival simile transformations are likewise an interesting area, especially compound adjectival similes (CAS). Many adjectival similes have CAS forms, whose frequencies are much higher than those of standard simile forms. CAS demonstrate the conceptual similes' ability to adapt syntactically. The empirical evidence does not suggest

---

[113] This primarily concerns the United States of America and the United Kingdom.

that standard similes are cognitively prominent lexical units. This is where an investigation of CAS could help us explain what helps conceptual similes survive in the language. Moon (2008: 34) argues that *as*-similes "likely (…) belong to receptive vocabularies" and considers them "negligible in terms of frequency" in corpora. Although standard adjectival similes undisputedly are infrequent compared to other single- and multi-word units, their CAS transformations offer a significantly different picture.

V. CONCLUSIONS

This work's primary purpose was the investigation of adjectival similes in corpus data. Four different corpora were chosen for the simile extraction to ensure data heterogeneity, allowing the collection of similes from various types of discourse. This offered insight into how various corpora handle the annotation of multi-word units and how restrictive the corpora are both technically and functionally.

The process of data mining was a tasking endeavour, but it revealed a lot about adjectival similes, their various alternative forms, and the individual corpora. The assumption that adjectival similes would be easy to extract from the corpora proved to be false despite the conventional formal frame. The precision of general CQL queries is overwhelmingly insufficient, and individually searching every simile is extremely time-consuming and limits the data we can obtain. The formal frame appears frequent in English, albeit only marginally used by adjectival similes.

Standard dictionaries often contain a limited simile set, which typically does not reflect the empirical evidence. Furthermore, the majority of simile entries are described as either 'fixed expressions' or 'idioms'. Consequently, the dictionaries generally do not provide us with any 'simile lists' to use as a point of departure. Using specialised idiom or simile dictionaries does not help either. Firstly, they are exceedingly rare since online dictionaries typically incorporate idioms entries or whole sections. Secondly, dictionaries such as Sommer's *Similes Dictionary* (2013) contain many creative similes that are not attested in corpus data and are often clearly author-specific expressions. These dictionaries contribute to the literary image of similes and distort their overall perception.

Naturally, several limitations ought to be mentioned. Any corpus-based research is limited by the chosen corpora. The data never represent a complete picture and only provide a fraction of the language reality. Although conclusions based on corpus investigations cannot be viewed as ultimate truths, they can be indicative of general tendencies in the language with great reliability.

Furthermore, corpus-based studies are restricted by our theoretical preconceptions. This can perfectly be demonstrated by adjectival similes. On many occasions throughout this research, I considered including various transformations to complement the standardised *as*-forms in the overall frequencies. This was partly due to the standard similes being relatively infrequent but also because, fundamentally, the transformations

represent the comparison concept equally. Ultimately, the statistics only include the standard forms due to space constraints. However, the transformations are believed to belong in the adjectival simile research.

To conclude, this work sheds light on many aspects of adjectival similes. Conversely, it also gives rise to many questions. The goals of this enquiry were achieved, but more research is necessary to further our understanding of adjectival similes and their various forms.

REFERENCES

Barnden, J. (2009). Metaphor and Context: A Perspective from Artificial Intelligence. In Musolff, A., Zinken, J. (eds) *Metaphor and Discourse*, 79-94. Palgrave Macmillan, London. https://doi.org/10.1057/9780230594647_6

Barnden, J.A. (2016). Metaphor and simile: Categorizing and comparing categorization and comparison. In Gola, E. & Ervas, F. (Eds), *Metaphor and Communication*, 25-46. John Benjamins.

Bartlett, J. R. (1848). *Dictionary of Americanisms: a glossary of words and phrases, usually regarded as peculiar to the United States*. New York: Bartlett and Welford, No. 7 Astor House.

Beardsley, M. (1958). *Aesthetics: Problems in the Philosophy of Criticism*. New York: Harcourt, Brace & World.

Black, M. (1962) Metaphor. In Black, M., *Models and Metaphors*. Ithaca, New York: Cornell University Press.

Bush, S.J. (2020). Ambivalence, Avoidance, and Appeal: Alliterative Aspects of Anglo Anthroponyms. *Names – A Journal of Onomastics, 68*:3, 141-155. DOI: 10.1080/00277738.2020.1775471

Čermák, F. (2007). *Frazeologie a idiomatika česká a obecná / Czech and General Phraseology*. Praha: Univerzita Karlova v Praze – Nakladatelství Karolinum.

Čermák, F. et al. (1983-2009): *Slovník české frazeologie a idiomatiky (SČFI)*, vol. 1-4. Praha: Academia / Leda.

Coard, R. L. (1959). Wide-Ranging Alliteration. *Peabody Journal of Education, 37* (1), 29-35. DOI: 10.1080/01619565909536881

Cresswell, J. (2021). *Oxford Dictionary of Word Origins.* Oxford University Press.

Cruse, A. (2006). *A glossary of semantics and pragmatics*. Edinburgh: Edinburgh University Press.

Cruse, D. A. (2011): *Meaning in Language. An introduction to Semantics and Pragmatics.* (3rd ed.). Oxford University Press.

Davidson, D. (1978). What Metaphors Mean. *Critical Inquiry, 5*(1), 31-47. Retrieved from http://www.jstor.org/stable/1342976

Dienhart, J. M. (1999). A linguistic look at riddles. *Journal of Pragmatics, 31*(1), 95-125.

Emmer, J. (2020). Boring as hell: a corpus study of intensifying post-modification of predicative adjectives in the ADJ as NOUN frame. *Linguistica Pragensia, 2020/2*(30), 127-137. DOI: 18059635.2020.2.2

Fishelov, D. (1993). *Metaphors of genre: the role of analogies in genre theory*. University Park (Pa.): Pennsylvania state university press.

Fogelin, R. (1988). *Figuratively Speaking*. New Haven, CT: Yale University Press.

Gargani, A. (2014). *Poetic Comparisons: How Similes Are Understood* (Doctoral thesis). University of Salford.

Gargani, A. (2016). Similes as poetic comparisons. *Lingua, Vol. 175-176*, 54-68.

Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford: Oxford University Press.

Gilquin, G. (2010). Corpus, cognition and causative constructions. *Studies in Corpus Linguistics*, *39*. Amsterdam/Philadelphia: John Benjamins.

Glucksberg, S., Haught, C. (2006). On the Relation Between Metaphor and Simile: When Comparison Fails. *Mind & Language, Vol. 21* No. 3, 360–378.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.

Gray, B., Biber, D. (2015). Phraseology. In D. Biber & R. Reppen (Authors), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press.

Huddleston, R.D. (2002). Comparative constructions. In Huddleston, R.D., Pullum, G.K. (Authors). *The Cambridge Grammar of the English Language. Cambridge*. UK: Cambridge University Press.

Israel, M., Harding, J., Tobin, V., Achard, M. & Kemmer, S. (2004). *On simile*. Stanford, California: CSLI Publications. Cognitive Linguistics Bibliography (CogBib).

Lakoff, G., Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Leech, G. (1981). *Semantics: The study of meanings*. Penguin Books.

Lindquist, H. (2009). *Corpus Linguistics and the Description of English*. Edinburgh, Edinburgh University Press.

Lipka, L. (1992). *An Outline of English lexicology: Lexical structure, word semantics and word-formation*. Tübingen: Niemeyer.

Lycan, W. G. (2001) *Philosophy of Language. A Contemporary Introduction*. London/New York: Routledge.

Makkai, A. (1972). *Idiom Structure in English*. Berlin, Boston: De Gruyter Mouton.

Makkai, A. (2013). *Handbook of Commonly Used American Idioms*. Barrons Educational Series Inc.

McEnery, T., Hardie, A. (2012) *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press.

Moon, R. (2008). Conventionalized as-similes in English: a problem case. *International Journal of Corpus Linguistics*, *13*(1), 3-37. https://doi.org/10.1075/ijcl.13.1.03moo

Moon, R. (2011). Simile and dissimilarity. *Journal of Literary Semantics, 40*(2), 133-157.

Niculae, V., Danescu-Niculescu-Mizil, C. (2014). Brighter than Gold: Figurative Language in User Generated Comparisons. *Proceedings of the 2014 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), 2008-2018. doi:10.3115/v1/d14-1215

Norrick, N. R. (1986). Stock Similes. *Journal of Literary Semantics XV/1*, 39-52.

Norrick, N. R. (1987). Semantic aspects of comparative noun-adjective compounds. In Asbach-Schnitker, B., Roggenhofer, J. (eds.) *Neuere Forschungen zur Wortbildung und Historiographie der Linguistik*, 145–154. Gunther Narr Verlag, Tübingen.

Novoselec, Z., Parizoska, J. (2012). A corpus-based study of similes and cognate adjectival forms in English, Swedish and Croatian. In Pamies, A., Pazos Bretaña, J.M., Luque Nadal, L. (eds.) *Phraseology and Discourse: Cross Linguistics and Corpus-based Approaches*, 101-110. Schneider Verlag, Baltmannsweiler.

O'Donoghue, J. (2009): Is a Metaphor (Like) a Simile? Differences in Meaning, Effects and Processing. *UCL Working Papers in Linguistics, 21* (2009), 125-149.

Oleniak, M. (2018). Semantic representation of similes (based on the Ukrainian, English and Polish languages). *Topics in Linguistics, 19*(1), 18-32. doi: https://doi.org/10.2478/topling-2018-0002

Ortony, A. (1979). Beyond literal similarity. In *Psychological Review 86*/3, pp. 161-180.

Palmer, F. R. (1976). *Semantics: A new outline.* Cambridge: Cambridge University Press.

Parizoska, J., & Filipović Petrović, I. (2017). Variation of Adjectival Slots in kao ('as') Similes in Croatian: A Cognitive Linguistic Account. In R. Mitkov (Ed.), *Computational and Corpus-Based Phraseology.* EUROPHRAS 2017. Lecture in Computer Science, 10596, 348-362.

Plag, I. (2003). *Word-formation in English.* Cambridge, UK: Cambridge University Press.

Qadir A., Riloff, E., & Walker, M.A. (2015). Learning to recognize affective polarity in similes. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 190-200, Lisbon, Portugal, September. Association for Computational Linguistics.

Qadir, A., Riloff, E., & Walker, M.A. (2016). Automatically Inferring Implicit Properties in Similes. *Proceedings of the 2016 Conference North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Roach, P. (2009). *English Phonetics and Phonology. A practical course.* (4th ed.). Cambridge University Press.

Roper, J., ed. (2011). *Alliteration in Culture*. Palgrave MacMillan London.

Searle, J. (1979). Metaphor. In Searle, J. (Ed.), *Expression and meaning: Studies in the theory of speech acts*, pp. 76-116. Cambridge: Cambridge University Press.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Berlin: Language Science Press.

Tartakovsky, R., & Shen, Y. (2019). Meek as milk and large as logic: A corpus study of the non-standard poetic simile. *Language and Literature, 28*(3), 203-220. https://doi.org/10.1177/0963947019865445

Veale, T. (2013). Humorous similes. In Ford, T. (Ed.), *Humor: International Journal of Humor Research, 26*(1), 3-22.

Wikberg, K. (2008). 8. phrasal similes in the BNC. *Phraseology*, 127-142. https://doi.org/10.1075/z.139.14wik

Corpora

Davies, Mark. (2008-) *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/ .

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. *7th International Corpus Linguistics Conference CL*, 125-127.

Křen, M.. Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P. Et Zasina, A. (2015). *SYN2015: reprezentativní korpus psané češtiny*. Praha: Ústav Českého národního korpusu FF UK. Available at: http://www.korpus.cz

Křen, M.. Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P. et Zasina, A. (2019). *Korpus SYN, verze 8 z 12. 12. 2019*. Praha: Ústav Českého národního korpusu FF UK. Available at: http://www.korpus.cz

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3): 319-344. DOI: 10.1075/ijcl.22.3.02lov

The British National Corpus (BNC) – Hoffmann, S., & Evert, S. (1996). BNCweb (CQP-edition). Accessed from https://cqpweb.lancs.ac.uk/ (last accessed on June 25, 2023).

Sources

Ammer, C. (2013a). *The American Heritage Dictionary of Idioms* (2nd ed.). Houghton Mifflin Harcourt.

Ammer, C. (2013b). *The Dictionary of Clichés*. Skyhorse.

Cambridge dictionary. Cambridge University Press. Accessed from https://dictionary.cambridge.org/ (Last accessed on June 25, 2023)

Collins dictionary (online). Collins. Accessed from https://www.collinsdictionary.com/ (last accessed on June 25, 2023)

Farlex International (2017). *The Farlex Idioms and Slang Dictionary*. Createspace Independent Publishing Platform.

Merriam-Webster dictionary. Accessed from https://www.merriam-webster.com (Last accessed on June 25, 2023)

Milton, J. (2003). *Paradise lost* (J. Leonard, Ed.). Penguin Books.

NET Bible free Bible software Bible studies. (n.d.). The Gospel of Matthew. NET bible: The Biblical Studies Foundation. Retrieved March 2, 2023 from https://bible.org/netbible/

Online Etymology Dictionary. Accessed from https://www.etymonline.com/ (Last accessed on June 25, 2023)

Oxford Learner's Dictionaries. Oxford University Press. Accessed from https://www.oxfordlearnersdictionaries.com/ (Last accessed on June 25, 2023)

Seidl, J., McMordie, W. (1988). *English idioms and How to Use Them* (5th ed.). Ofxord University Press.

Seidl, J., McMordie, W. (1978). *English Idioms and How to Use Them*. The Compact Edition of the Oxford English Dictionary 1971, Oxford University Press.

Simile Stack. Accessed from https://www.similestack.com/ (Last accessed on June 25, 2023)

Sommer, E. (2013). *Similes Dictionary* (2nd ed.). Visible Ink Press.

The Free Dictionary. Farlex, Inc. Accessed from https://www.thefreedictionary.com/ (Last accessed on June 25, 2023)

APPENDIX 1 – THE ENGLISH SIMILE SAMPLE (309 items)

American as apple pie	brave as hope
big as a barn	bright as a button
big as a basketball	bright as day
big as a field	bright as the Sun
big as a football	busy as a bee
big as a football field	busy as bees
big as a football stadium	clean as a whistle
big as a house	clear as a bell
big as a mountain	clear as crystal
big as a whale	clear as day
big as houses	clear as daylight
big as life	clear as glass
big as marbles	clear as mud
big as saucers	clear as the sun
big as Texas	clear as water
big as the sky	cold as ice
black as a crow	cold as stone
black as a raven	common as dirt
black as a raven's wing	common as muck
black as coal	cool as a cucumber
black as ink	cool as ice
black as midnight	crazy as a loon
black as night	crooked as a dog's hind leg
black as pitch	cute as a bug
black as soot	cute as a bug's ear
black as the ace of spades	cute as a button
black as the night	daft as a brush
black as thunder	dappy as pig shit
blind as a bat	dark as night
blue as the sky	dark as the night
bold as a lion	dark as the night sky
bold as brass	dead as a dodo
brave as a lion	dead as a doornail

dead as mutton	good as her word
deaf as a post	good as his word
deaf as posts	good as sex
different as day and night	green as grass
different as chalk and cheese	guilty as sin
different as chalk from cheese	happy as a clam
different as night and day	happy as a lark
drunk as a lord	happy as a pig
drunk as a skunk	happy as a pig in muck
dry as a bone	happy as a pig in mud
dry as a desert	happy as a pig in shit
dry as dust	happy as clams
dull as dishwater	happy as Larry
dumb as a box of hair	hard as a rock
dumb as a box of rocks	hard as concrete
dumb as a post	hard as diamond
dumb as a rock	hard as diamonds
dumb as bricks	hard as granite
dumb as dirt	hard as iron
dumb as rocks	hard as nails
easy as ABC	hard as rock
easy as pie	hard as rocks
fast as lightning	hard as steel
fast as the wind	hard as stone
fat as a pig	harmless as doves
fit as a fiddle	healthy as a horse
flat as a board	heavy as lead
flat as a pancake	high as a house
free as a bird	high as a kite
free as the wind	high as an elephant's eye
fresh as a daisy	high as the sky
gentle as a lamb	hot as an oven
gentle as doves	hot as balls
good as gold	hot as fire

hot as hell	old as the world
hot as the sun	old as time
hungry as a wolf	old as war
cheap as chips	pale as a ghost
innocent as doves	pale as death
keen as mustard	pale as the moon
large as life	pissed as a fart
light as a feather	pissed as a nook
light as air	plain as a pikestaff
lonely as a cloud	plain as day
mad as a hatter	plain as the nose on (one's) face
mad as a hornet	pleased as punch
mad as a march hare	poor as a church mouse
miserable as sin	poor as dirt
naked as a jaybird	poor as church mice
naked as the day (one) came to this world	precious as gold
naked as the day (one) was born	pretty as a picture
natural as breathing	proud as a peacock
neat as a pin	proud as punch
nervous as a cat	pure as gold
nice as pie	pure as snow
old as America	pure as the driven snow
old as civilization	quick as a flash
old as dirt	quick as a rat
old as history	quick as a wink
old as humanity	quick as lightning
old as mankind	quiet as a grave
old as Methuselah	quiet as a church
old as the Bible	quiet as a church mouse
old as the earth	quiet as a mouse
old as the hills	quiet as a tomb
old as the history of sth	quiet as the grave
old as the republic	rare as hen's teeth
old as the universe	red as a beet

red as a beetroot	smooth as velvet
red as a tomato	snug as a bug
red as blood	soft as a baby's
regular as clockwork	soft as a baby's behind
right as rain	soft as a baby's bottom
rich as Croesus	soft as a baby's butt
rough as arseholes	soft as butter
safe as houses	soft as cotton
serious as a heart attack	soft as silk
sharp as a knife	soft as velvet
sharp as a razor	solid as a rock
sharp as a tack	solid as rock
sharp as knives	steady as a rock
sharp as razors	stiff as a log
sick as a dog	stiff as a board
sick as a parrot	stiff as a poker
sick as your secrets	still as a statue
silent as the grave	still as a stone
simple as ABC	straight as a die
simple as pie	straight as a ramrod
skinny as a rail	straight as an arrow
skinny as a rake	strong as a bear
slippery as an eel	strong as a bull
slow as molasses	strong as a horse
smart as a whip	strong as a lion
smooth as a baby	strong as an oak
smooth as a baby's ass	strong as an ox
smooth as a baby's behind	strong as death
smooth as a baby's bottom	strong as iron
smooth as a baby's bum	strong as steel
smooth as a baby's butt	stubborn as a mule
smooth as butter	sure as night
smooth as glass	sure as the sun
smooth as silk	sweet as candy

sweet as honey	white as snow
sweet as pie	white as the snow
sweet as sugar	wide as saucers
sweet as wine	wise as a serpent
tall as a house	wise as serpents
tall as a tree	
thick as a brick	
thick as pig shit	
thick as pigshit	
thick as thieves	
thick as two short planks	
thin as a rail	
thin as a rake	
thin as paper	
tight as a drum	
tough as nails	
tough as old boots	
transparent as glass	
true as steel	
ugly as sin	
unique as a fingerprint	
warm as toast	
weak as a kitten	
weak as water	
white as a ghost	
white as a sheet	
white as alabaster	
white as bone	
white as death	
white as chalk	
white as marble	
white as milk	
white as paper	
white as sheets	

APPENDIX 2 – THE CZECH SIMILE LIST (886 items)

bázlivý jako králík	blbý jako štandlík
bázlivý jako zajíc	blbý jako tágo
bdělý jako Argus	blbý jako troky
bezbranný jako dítě	blbý jako vantrok
bezmocný jako dítě	blbý jako vantroky
bezradný jako dítě	bledý jako alabastr
bílý jako alabastr	bledý jako duch
bílý jako křída	bledý jako křída
bílý jako lilie	bledý jako měsíc
bílý jako mléko	bledý jako mramor
bílý jako mramor	bledý jako mrtvola
bílý jako mrtvola	bledý jako papír
bílý jako padlý sníh	bledý jako sedma
bílý jako papír	bledý jako slonová kost
bílý jako sejra	bledý jako smrt
bílý jako slonová kost	bledý jako stěna
bílý jako smrt	bledý jako tvaroh
bílý jako sníh	bledý jako vosk
bílý jako stěna	bledý jako z vosku
bílý jako tvaroh	bledý jako zeď
bílý jako vápno	bledý jako zelenka
bílý jako z alabastru	bledý jako zelinka
bílý jako zeď	bohatý jako Argentinec
blbý jako bedna kytu	bohatý jako koza rohatá
blbý jako bota	bohatý jako Krésus
blbý jako kopyto	bohatý jako Rothschild
blbý jako motyka	bojovný jako kohout
blbý jako pařez	bradatý jako Kristus
blbý jako patník	bujný jako hříbě
blbý jako poleno	bystrý jako sokol
blbý jako pučtok	černý jako antracit
blbý jako rádio	černý jako ašant
blbý jako škopek	černý jako bota

černý jako cikán	čilý jako pometlo
černý jako cikánka	čilý jako rtuť
černý jako čert	čilý jako ryba
černý jako ďábel	čilý jako rybička
černý jako eben	čilý jako veverka
černý jako havran	čiperný jako veverka
černý jako havraní křídla	čistý jako alabastr
černý jako hřích	čistý jako klícka
černý jako inkoust	čistý jako křišťál
černý jako kolomaz	čistý jako lilie
černý jako mouřenín	čistý jako lilium
černý jako noc	čistý jako padlý sníh
černý jako saze	čistý jako panic
černý jako smola	čistý jako sklo
černý jako smůla	čistý jako slovo boží
černý jako uhel	čistý jako studánka
černý jako vrána	čistý jako z alabastru
červený jako cihla	čistý jako z bavlnky
červený jako jablíčko	čistý jako z cukru
červený jako jahoda	čistý jako ze škatulky
červený jako kohout	čistý jako zrcadlo
červený jako králík	děravý jako cedník
červený jako malina	děravý jako ementál
červený jako panenka	děravý jako řešeto
červený jako paprika	děravý jako síto
červený jako prase	děravý jako sýr
červený jako rak	divoký jako ďábel
červený jako růže	dlouhý jako chmelová tyčka
červený jako růžička	dlouhý jako jitrnice
červený jako vlčí mák	dlouhý jako Lovosice
čilá jako koroptev	dlouhý jako skládací metr
čilý jako čečetka	dlouhý jako slonbidlo
čilý jako koroptvička	dlouhý jako štangle
čilý jako mník	dlouhý jako tasemnice

dlouhý jako tyčka	hladký jako mramor
dlouhý jako týden	hladký jako mýdlo
dobrý jako anděl	hladký jako oblázek
dotěrný jako moucha	hladký jako olej
dotěrný jako štěnice	hladký jako samet
drsný jako smirkový papír	hladký jako sklo
drsný jako struhadlo	hladký jako stůl
drzý jako opice	hladký jako úhoř
drzý jako stádo opic	hladký jako zrcadlo
drzý jako štěnice	hladový jako herec
dutý jako bambus	hladový jako lev
falešný jako Jidáš	hladový jako pes
falešný jako kočka	hladový jako vlk
falešný jako pětník	hloupý jako bota
fousatý jako Krakonoš	hloupý jako dlabaný troky
fousatý jako Kristus	hloupý jako dřevo
hbitý jako jelen	hloupý jako husa
hbitý jako ještěrka	hloupý jako kopyto
hbitý jako lasice	hloupý jako osel
hbitý jako lasička	hloupý jako pařez
hbitý jako ohař	hloupý jako patník
hebký jako hedvábí	hloupý jako tágo
hebký jako samet	hloupý jako tele
hezký jako anděl	hloupý jako troky
hezký jako andělíček	hluchý jako pařez
hezký jako andílek	hluchý jako poleno
hezký jako obrázek	hluchý jako špalek
hezký jako panenka	hluchý jako tetřev
hladký jako alabastr	hnědý jako hovno
hladký jako dětská prdelka	hodný jako anděl
hladký jako had	holý jako koleno
hladký jako hedvábí	horký jako oheň
hladký jako led	horký jako turecká fangle
hladký jako mejdlo	hořký jako blín

hořký jako pelyněk	chladný jako mramor
hořký jako utrejch	chladný jako ocel
hořký jako zeměžluč	chladný jako psí čumák
hořký jako žluč	chladný jako smrt
hravý jako kotě	chladný jako sníh
hrubý jako smirkový papír	chlupatý jako Ezau
hubatý jako čert	chlupatý jako medvěd
hubený jako bič	chlupatý jako opice
hubený jako bidlo	chudý jako kostelní myš
hubený jako drožkářská kobyla	chudý jako myš
hubený jako chroust	chytrý jako advokát
hubený jako chrt	chytrý jako čert
hubený jako komár	chytrý jako četník
hubený jako koza	chytrý jako liška
hubený jako louč	chytrý jako opice
hubený jako luňák	chytrý jako opička
hubený jako lunt	chytrý jako policajt
hubený jako nit	chytrý jako rádio
hubený jako nitka	chytrý jako stádo opic
hubený jako nudle	jasný jako Brno
hubený jako pes	jasný jako den
hubený jako sirka	jasný jako facka
hubený jako šindel	jasný jako křišťál
hubený jako špejle	jasný jako pár facek
hubený jako tříska	jednoduchý jako facka
hubený jako tyčka	jednoduchý jako násobilka
hubený jako žížala	jednoduchý jako pár facek
huňatý jako medvěd	jemný jako hedvábí
hustý jako hrachová polévka	jemný jako pavučinka
hustý jako mlha	jemný jako peří
hustý jako smetana	jemný jako peříčko
chladný jako hrobka	klidný jako beránek
chladný jako kámen	klidný jako smrt
chladný jako led	kluzký jako mejdlo

kluzký jako mýdlo	lehký jako peří
kluzký jako úhoř	lehký jako peříčko
krásná jako bohyně	lehký jako pírko
krásná jako víla	lehoučký jako pápěří
krásný jako Adónis	lehounký jako pápěří
krásný jako anděl	lehounký jako peří
krásný jako andělíček	lehounký jako peříčko
krásný jako andílek	lehounký jako pírko
krásný jako antický bůh	lesklý jako drahokam
krásný jako Apollón	lesklý jako smůla
krásný jako bůh	lesklý jako zrcadlo
krásný jako madona	levý jako šavle
krásný jako obrázek	levý jako ta šavle
krásný jako růže	líný jako kanec
krásný jako sen	líný jako kapr
krásný jako Venuše	líný jako prase
krásný jako z pohádky	líný jako svině
krásný jako zrcadlo	líný jako štěnice
krotký jako beránek	líný jako valach
krutý jako Nero	líný jako vepř
křehký jako pápěří	líný jako veš
křivý jako když kanec chčije	lstivý jako had
křivý jako když pes chčije	lstivý jako liška
křivý jako turecká šavle	lstivý jako Odysseus
kudrnatý jako beránek	malý jako cvoček
kudrnatý jako ovce	malý jako cvrček
kyselý jako ocet	malý jako klícka
kyselý jako šťovík	malý jako náprstek
lakomý jako čert	malý jako špendlíková hlavička
lakomý jako chrt	mazaný jako advokát
ledový jako rampouch	mazaný jako čert
lehký jako facka	mazaný jako liška
lehký jako pápěrka	mazlivý jako kotě
lehký jako pár facek	měkký jako hedvábí

měkký jako houba	mrštný jako lasička
měkký jako koberec	mrštný jako opice
měkký jako kobereček	mrštný jako úhoř
měkký jako máslo	mrštný jako veverka
měkký jako mech	mrzutý jako dudy
měkký jako samet	mrzutý jako staré dudy
měkký jako tvaroh	nabalený jako cibule
měkký jako vosk	nabalený jako pumpa
milý jako trn v patě	nacucaný jako houba
mírná jako holubice	nacucaný jako žok
mírná jako holubička	nadýchaný jako obláček
mírný jako beránek	nadýchaný jako peříčko
mírný jako jehně	nadýchaný jako peřinka
mírný jako ovce	nahá jako Eva
mírný jako ovečka	nahý jako Adam
mlsný jako čert	nahý jako ašant
mlsný jako kocour	nalitý jako dělo
mlsný jako kočka	nalitý jako dobytek
mlsný jako koza	nalitý jako pupen
modrý jako nebe	nalitý jako slíva
modrý jako noc	nalitý jako snop
modrý jako pomněnky	nalitý jako sud
modrý jako šmolka	nalitý jako žok
mokrý jako hastrman	naložený jako mezek
mokrý jako hnůj	naložený jako soumar
mokrý jako houba	namačkaní jako herynci
mokrý jako myš	namačkaní jako sardinky
moudrý jako Šalamoun	namačkaní jako slanečci
moudrý jako Šalomoun	napařený jako Dán
mrštný jako had	napínavý jako kšandy
mrštný jako hádě	napitý jako houba
mrštný jako ještěrka	napjatý jako luk
mrštný jako kočka	napjatý jako provazy
mrštný jako lasice	napjatý jako struna

napjatý jako strunky	opilý jako štěně
napjatý jako špagát	opilý jako zvíře
napjatý jako tětiva	opilý jako žok
napnutý jako kšandy	opuštěný jako hruška v poli
napnutý jako provazy	opuštěný jako hruška v širém poli
napnutý jako struna	opuštěný jako pes
napnutý jako špagát	ospalý jako kotě
napružený jako péro	ospalý jako sysel
nasraný jako brigadýr	ospalý jako štěně
nasraný jako kanonýr	ostrý jako břitva
nasraný jako kýbl	ostrý jako dýka
natažený jako struna	ostrý jako jed
navlečený jako cibule	ostrý jako jehla
navlečený jako na severní točnu	ostrý jako meč
němý jako kapr	ostrý jako nůž
němý jako ryba	ostrý jako šídlo
němý jako rybička	ostrý jako žiletky
neoblomný jako skála	ostříhaný jako trestanec
neomylný jako papež	ošklivá jako čarodějnice
nešťastný jako šafářův dvoreček	ošklivá jako ropucha
neústupný jako pařez	ošklivá jako stará čarodějnice
nevinný jako anděl	ošklivý jako ďábel
nevinný jako andílek	ošklivý jako noc
nevinný jako lilie	ošklivý jako opice
nevinný jako lilium	oškubaný jako lípa
oblečený jako hastroš	oteklý jako bakule
oblečený jako na severní točnu	otrávený jako šváb
ohavný jako noc	otrhaný jako cikán
ohyzdný jako noc	otrhaný jako lípa
opálený jako ašant	ožralý jako čuně
opálený jako Indián	ožralý jako Dán
opilý jako čuně	ožralý jako dělo
opilý jako Dán	ožralý jako prase
opilý jako mraky	ožralý jako slíva

ožralý jako svině	pracovitý jako mraveneček
ožralý jako štěně	pracovitý jako pejsek
pádný jako pěst	pracovitý jako včela
paličatý jako mezek	pracovitý jako včelička
paličatý jako osel	pracovitý jako včelka
pěkný jako panáček	prohnutý jako luk
pevný jako ocel	proměnlivý jako počasí
pevný jako skála	prostřílený jako řešeto
pevný jako židovská víra	prostřílený jako síto
pevný jako žula	protivný jako činže
pichlavý jako ježek	provrtaný jako řešeto
pilný jako mravenec	pružný jako proutek
pilný jako mraveneček	pružný jako vrbový proutek
pilný jako včela	přelétavý jako motýl
pilný jako včelička	přesný jako hodinky
pilný jako včelka	přesný jako hodiny
pitomý jako ovce	přesný jako švýcarské hodinky
pitomý jako pučtok	pyšný jako páv
pitomý jako tágo	pyšný jako pávice
pitomý jako vantrok	rezavý jako liška
pitomý jako vantroky	rovný jako když kanec chčije
plachý jako laň	rovný jako když střelí
plachý jako laňka	rovný jako mlat
plachý jako srna	rovný jako pravítko
platný jako pes v kostele	rovný jako stůl
plavý jako lvice	rovný jako svíce
plochý jako stůl	rovný jako šíp
podobný jako vejce vejci	rovný jako šňůra
pomalý jako šnek	rovný jako turecká šavle
pomalý jako želva	rozbitý jako turecká fangle
poslušný jako pejsek	rozcuchaná jako čarodějnice
potrhlý jako švec	rozcuchaná jako stará čarodějnice
pověrčivý jako stará baba	rozpálený jako cihlička
pověrčivý jako stará bába	rozpálený jako dvířka od kamen

rozpálený jako kamna	silný jako dub
rozpálený jako oheň	silný jako golem
rozpálený jako pec	silný jako Herkules
rozpálený jako trajfus	silný jako hrom
rozpálený jako žehlička	silný jako kobyla
rozstřílený jako cedník	silný jako kůň
roztažený jako žába	silný jako lev
roztrhaný jako turecká fangle	silný jako lvi
rozvrzaný jako staré schody	silný jako medvěd
rozzuřený jako bejk	silný jako noha
rozzuřený jako býk	silný jako Samson
rudý jako krocan	silný jako smrt
rudý jako mák	silný jako tur
rudý jako pivoňka	silný jako z oceli
rudý jako rak	skromná jako fialka
růžový jako prasátko	skromný jako fialinka
růžový jako selátko	skromný jako chudobka
rychlý jako blesk	slabý jako čaj
rychlý jako ďábel	slabý jako dech
rychlý jako chrt	slabý jako duch
rychlý jako jelen	slabý jako hnilička
rychlý jako ohař	slabý jako komár
rychlý jako raketa	slabý jako moucha
rychlý jako sokol	slabý jako pápěrka
rychlý jako střela	slabý jako pápěří
rychlý jako šíp	slabý jako třtina
rychlý jako šipka	slabý jako z kozy duch
rychlý jako vítr	sladký jako cecek
scvrklý jako křížala	sladký jako cukr
sdílný jako led	sladký jako cumel
shrbený jako stařec	sladký jako med
silný jako bejk	sladký jako vánek
silný jako Bivoj	slepý jako kotě
silný jako býk	slepý jako krtek

slepý jako netopýr	svobodný jako Amerika
slepý jako patrona	šedivý jako popel
slizký jako had	šeredný jako ďábel
smrdutý jako zkažený zub	šeredný jako noc
smutný jako želva	šikovný jako hrom do police
snědý jako cikán	šikovný jako motovidlo
snědý jako cikánka	široký jako almara
sprostý jako dlaždič	škaredá jako ropucha
sprostý jako erární deka	škaredý jako noc
sprostý jako kanálník	škaredý jako opice
starý jako Abrahám	špičatý jako jehla
starý jako lidstvo	špičatý jako šídlo
starý jako lidstvo samo	špinavý jako bota
starý jako Metuzalém	špinavý jako cikán
starý jako svět	špinavý jako kanec
stejný jako každý druhý	špinavý jako onuce
stejný jako každý jiný	špinavý jako prase
stísněný jako klícka	špinavý jako špek
studený jako had	šťastný jako blecha
studený jako kámen	šťastný jako děcko
studený jako kus ledu	šťastný jako malé dítě
studený jako led	štíhlá v pase jako vosa
studený jako mramor	štíhlý jako bříza
studený jako psí čumák	štíhlý jako gazela
studený jako psí nos	štíhlý jako jedle
studený jako rampouch	štíhlý jako laň
studený jako ryba	štíhlý jako laňka
suchý jako drn	štíhlý jako proutek
suchý jako louč	štíhlý jako svíce
suchý jako šindel	štíhlý jako svíčka
suchý jako trn	švorcový jako eso pikový
suchý jako troud	švorcový jako pikové eso
svěží jako broskev	temný jako noc
svěží jako orosená broskev	tenký jako hůlky

tenký jako lupínek	tlustý jako koule
tenký jako nit	tlustý jako měch
tenký jako nitka	tlustý jako noha
tenký jako nudle	tlustý jako prase
tenký jako papír	tlustý jako štandlík
tenký jako pavučinka	tlustý jako valach
tenký jako sirka	tlustý jako vepř
tenký jako struna	tmavý jako noc
tenký jako strunka	trpělivý jako beránek
tenký jako tříska	trpělivý jako Job
tenký jako vlas	trpělivý jako ovečka
tenký jako vlásek	tuhý jako houžev
teplý jako chcanky	tuhý jako podrážka
teplý jako kafe	tupý jako motyka
teplý jako polívka	tupý jako pařez
těžký jako cent	tupý jako poleno
těžký jako hrom	tupý jako prdel
těžký jako kámen	tvárný jako vosk
těžký jako kráva	tvrdohlavý jako beran
těžký jako olovo	tvrdohlavý jako kozel
těžký jako prase	tvrdohlavý jako mezek
tichý jako myš	tvrdohlavý jako osel
tichý jako myška	tvrdý jako beton
tichý jako oukropeček	tvrdý jako brok
tichý jako pěna	tvrdý jako hřebíky
tichý jako smrt	tvrdý jako kámen
tlustý jako bagoun	tvrdý jako kost
tlustý jako basa	tvrdý jako křemen
tlustý jako bečka	tvrdý jako mlat
tlustý jako bejk	tvrdý jako oblázek
tlustý jako buřt	tvrdý jako ocel
tlustý jako býk	tvrdý jako podrážka
tlustý jako čuně	tvrdý jako roh
tlustý jako hroch	tvrdý jako skála

tvrdý jako suk	velký jako kolo od vozu
tvrdý jako z křemene	velký jako kráva
tvrdý jako z oceli	velký jako medvěd
tvrdý jako železo	velký jako náklaďák
tvrdý jako žula	velký jako pěst
učenlivý jako opice	velký jako pětník
ukecaný jako stará bába	velký jako skřivánek
úlisný jako had	velký jako tele
umazaný jako uhlíř	velký jako vejce
umíněný jako beran	věrný jako pes
umíněný jako mezek	volný jako motýl
umouněný jako cikáně	volný jako pták
urostlý jako bříza	vyčůraný jako mraky
urostlý jako topol	vychrtlý jako lusk
úskočný jako had	vychrtlý jako pes
uštvaný jako kůň	vyjukaný jako poledne
utahaný jako alík	vylekaný jako králík
utahaný jako čokl	vylekaný jako zajíc
utahaný jako kočka	vymačkaný jako citrón
utahaný jako kotě	vymóděný jako hastroš
utahaný jako kůň	vypasený jako čuně
utahaný jako pes	vypasený jako louka
utahaný jako štěně	vypasený jako prase
útlý jako proutek	vypasený jako vepř
uvázaný jako pes u boudy	vyplašený jako králík
úzký jako jitrnice	vyplašený jako poledne
veliký jako pěst	vyplašený jako zajíc
veliký jako slon	vyprahlý jako poušť
veliký jako stodola	vyprahlý jako Sahara
velký jako almara	vyprahlý jako troud
velký jako Brno	vyschlý jako došek
velký jako holubí vejce	vyschlý jako louč
velký jako hrachy	vyschlý jako treska
velký jako kobyla	vyschlý jako trn

vyschlý jako troud	zelený jako sedma
vysoký jako slonbidlo	zkroucený jako paragraf
vysoký jako věž	zkroucený jako preclík
vystrojený jako hastroš	zlitý jako Dán
vysušený jako treska	zlitý jako prase
vyzáblý jako pes	zlý jako pes
vyžraný jako prase	zmačkaný jako cumel
vyžraný jako zub	zmačkaný jako onuce
vzácný jako šafrán	zmalovaný jako Indián
vzteklý jako čert	zmatený jako Goro před Tokiem
vzteklý jako ďábel	zmrzlý jako drozd
vzteklý jako křeček	zmrzlý jako hovno
vzteklý jako pes	zmrzlý jako kost
záhadný jako smrt	zmrzlý jako preclík
zahrabaný jako krtek	zmrzlý jako rampouch
zalezlý jako krtek	známý jako falešný pětník
zamilovaný jako student	zpitý jako dobytek
zamračený jako noc	zpocený jako myš
zarostlý jako Ezau	zrádný jako had
zarostlý jako Tarzan	zrudlý jako krocan
zavřený jako v kleci	zrzavý jako liška
zazobaný jako sysel	ztuhlý jako hovno
zdravý jako buk	ztuhlý jako rampouch
zdravý jako dub	ztuhlý jako sloup
zdravý jako hřib	ztuhlý jako socha
zdravý jako křen	zvědavý jako Eva
zdravý jako lípa	zvědavý jako opice
zdravý jako pařízek	zvědavý jako opička
zdravý jako ryba	zvědavý jako stará baba
zdravý jako rybička	zvědavý jako stará bába
zdravý jako rys	zvědavý jako straka
zdravý jako řípa	žárlivý jako Othello
zdravý jako tuřín	živý jako rtuť
zelený jako brčál	žíznivý jako poušť

žlutý jako citrón

žlutý jako kanárek

žlutý jako šafrán

žravý jako kobylka