‘For the fifty-eleventh time’:
Examining cross-linguistic properties of hyperbolic numerals and quasi-numeral expressions through parallel text extraction

Amanda Kann

Department of Linguistics
Master’s Thesis, 30 ECTS credits
Typology and Linguistic Diversity – Master’s Thesis, LIM050
Master’s Programme in Language Science with a Specialisation
in Typology and Linguistic Diversity, 120 ECTS credits
Spring semester 2022
Supervisors: Ljuba Veselinova, Robert Östling
Swedish title: ‘För femtioelfte gången’: En tvärspråklig
parallellkorpusstudie av räkneord och kvasinumeriska
uttryck i hyperbolisk användning

Stockholm
University

‘For the fifty-eleventh time’:
Examining cross-linguistic properties of hyperbolic numerals and quasi-numeral expressions through parallel text extraction

Abstract

In some languages, vague and exaggerated quantities can be represented using certain conventionalised numeral expressions with cross-linguistically varying values, such as Danish *hundredesytten* ‘117’. Hyperbolic quantities can also be expressed using other quantifier expressions (such as English *zillion*) which, while they do not denote a specific numerical value, have both structural and functional similarities with exact numerals. These hyperbolic numerals and quasi-numerals are common in a variety of primarily informal contexts, but have yet to be subject to a systematic cross-linguistic investigation. In this study, hyperbolic numerals and quasi-numeral expressions are extracted from a massively parallel corpus of film and TV subtitles in a convenience sample of 50 languages, using automatic word alignments and seed expressions in 5 languages. Additional expressions are also obtained through elicitation. The collected expressions are subsequently analyzed and categorized to investigate patterns in distribution, value, morphology, function and usage. Findings include a cross-linguistic preference toward certain powers of the numeral base for hyperbolic numerals, and consistent patterns of construction with varying degrees of cross-linguistic prevalence for quasi-numerals.

Keywords

hyperbole, hyperbolic quantification, numeral typology, parallel texts, quasi-numerals

Sammanfattning

Obestämda och överdrivna mängder kan i vissa språk uttryckas genom specifika räkneord med tvärspråkligt varierande, exakta numeriska värden, t ex danska *hundredesytten* ‘117’. Överdrivna mängder kan även uttryckas med andra kvantifieraruttryck (t ex engelska *zillion*) som liknar räkneord i form och funktion, trots att de inte har något eget numeriskt värde. Dessa hyperboliska räkneord och kvasinumeriska uttryck är vanligt förekommande i informellt språk, men har ännu inte varit föremål för en systematisk tvärspråklig undersökning. I denna studie extraheras hyperboliska räkneord och kvasinumeriska uttryck från en massivt parallell korpus av film- och TV-undertexter i ett bekvämlighetsurval av 50 språk, genom automatisk ordlänkning och seed-uttryck på 5 språk. Ytterligare uttryck samlas in genom elicitering. De identifierade uttrycken analyseras och kategoriseras för att undersöka tvärspråkliga mönster i förekomst, numeriskt värde, morfologi, funktion och användning. Hyperboliska räkneord tenderar tvärspråkligt att föredra vissa potenser av talbasen som värden, och konsekventa mönster (med varierande tvärspråklig frekvens) för konstruktion av kvasinumeriska uttryck identifieras.

Nyckelord

hyperbol, hyperbolisk kvantifiering, kvasinumeriska uttryck, parallelltexter, räkneordstypologi

Contents

1 Introduction

Vague and hyperbolic expressions are ubiquitous in informal language, and are used for a wide range of pragmatic functions.

A commonly encountered type of hyperbole is hyperbolic quantification, in which the magnitude of a given quantity is exaggerated, typically for emphatic or emotive function. Hyperbolic quantification can be expressed using a variety of strategies, commonly involving a 'round' numeral, as in example (1):

(1) Hyperbolic numeral usage in English [eng] (McCarthy and Carter 2004, p. 167, ex. 8)
I live in Nottingham now cos I came here to study at the university. Been here for ***about a thousand years****.*

While the use of 'round' numeral expressions in hyperbole seems to be cross-linguistically common (Lavric 2010, pp. 136–137), in line with the frequent general approximative usage of these numerals (Krifka 2009, pp. 2–3), certain exact non-'round' numerals also occur as hyperbolic quantifiers in some languages, despite not being used in non-hyperbolic approximation. Numerals with this property include French *trente-six* '36' (Lavric 2010, p. 136), as well as Danish *hundredesytten* '117', as demonstrated in example (2). The cross-linguistic prevalence of this type of hyperbolic numeral is not well studied, and no tendencies or patterns in numerical value have been suggested.

(2) Hyperbolic usage of *hundredesytten* in Danish [dan] (from daTenTen20, Jakubíček et al. 2013)

Lad	*os,*	*for*	***hundredesytten-de***	***gang***	*rekapitulere*	*problematikk-en.*
let.IMP	1PL	for	**117-th**	**time**	recapitulate.INF	problem-DEF

'Let us restate the problem for **the umpteenth time**.'

Another category of exclusively hyperbolic quantifiers, including English *umpteen* and Italian *fantastilione*, are morphosyntactically and functionally similar to numeral expressions yet have no definite numerical value of their own. These hyperbolic *quasi-numeral* expressions also include constructions such as Czech *x-krát* '*x*-times', where a numeral is replaced by an algebraic variable such as *x* or *n*. A number of hyperbolic quasi-numerals in English are analyzed by Chrisomalis (2016), but the category has not yet been subject to systematic cross-linguistic investigation.

For hyperbolic numerals, prior cross-linguistic comparisons of their properties have likewise been focused exclusively on English and a few other Indo-European languages, leaving vast amounts of potential variation outside this limited set of languages unexplored (Veselinova 2020).

This study aims to broaden the cross-linguistic understanding of hyperbolic numerals and quasi-numeral expressions by collecting, analyzing and typologizing hyperbolic numerals and quasi-numerals across a language sample with greater areal and genealogical coverage than prior work.

As phenomena predominantly observed in informal, spoken language are rarely well documented in traditional language descriptions, hyperbolic numerals and quasi-numeral expressions are instead extracted from the massively parallel *OpenSubtitles2018* corpus of film and TV subtitles using word alignments and seed expressions in 5 languages.

2 Background

Section 2.1 provides a cross-linguistic account of the general functional and structural properties of numerals in language, with a particular focus on numeral 'roundness', which is a relevant factor in both hyperbolic and non-hyperbolic approximation, and basic derivations of cardinal numerals which appear relevant to the construction of certain hyperbolic quantifiers. In section 2.2, the use of numerals for vague and hyperbolic quantification is explored, and prior work on the hyperbolic usage of both numerals with exact values and indefinite numeral-derived (quasi-numeral) expressions is discussed. Finally, aims and individual research questions are presented in section 2.3.

2.1 Numerals

2.1.1 Basic distinctions and series within numeral systems

Numerals are prototypically used to quantify, to enumerate and to describe value and other numerical properties. Although discrete numerals do not seem to be present in all languages without exception (see Hammarström 2010, pp. 17–21), the vast majority of the world's languages have numeral expressions that form one or several distinct and often productive numeral systems (Comrie 2013).

Numeral expressions in productive numeral systems are fundamentally divided into simple (monomorphemic) and complex (polymorphemic) numerals. Complex numerals are constructed using a *numeral base*, the value of which is specific to the numeral system (Comrie 2013). While a range of numeral bases can be identified among the world's numeral systems, including the relatively rare bases 4, 6 and 12 (Hammarström 2010, pp. 24–31), the cross-linguistically most common types seem to be decimal (base 10) and vigesimal (base 20) systems (Comrie 2013). These two types are not entirely discrete – there is also a continuum of hybrid decimal-vigesimal systems, where different numeral bases are used for numerals within specific ranges. For instance, Basque systematically uses base 20 for numerals between 20 and 99, but base 10 to form numerals above 100 (Comrie 2013). Similar systems are found in Danish and certain varieties of French, where a more restricted subset of numerals below 100 are constructed vigesimally. Base variation can also be context-dependent, as in some Oceanic languages where base 4 is used to count only certain objects (Hammarström 2010, p. 27).

Within a system, numerals can generally be divided further into separate classes or series based on their function. The most fundamental of these is the cardinal numeral series, which is most commonly used for counting and quantifying entities. Cardinal numerals are also typically the base from which other numeral series are derived, although this is also not without exception (Hammarström 2010, pp. 34–35).

Ordinal numerals seem to be the cross-linguistically most common numeral series following cardinal numerals (Veselinova 2004), and have also received more cross-linguistic attention than other numeral derivatives (Veselinova 2020). The primary function of ordinal numerals is to denote the order or position of an entity in a sequence (Stolz and Veselinova 2013), as demonstrated in example (2).

Other numerals are used to describe the frequency of an action or occurrence. These frequentative or multiplicative numerals can either function similarly to cardinals, indicating the number of repetitions as in (3a), or to ordinals, indicating the position of a particular action in a sequence of repetition as in (3b). Some frequentative numerals can also occur in attributive position, as in (3c).

(3) Examples of frequentative numerals in Swedish [swe] (own data)

a.	*Hon*	*ring-de*	*fem*	*gång-er.*
	3SG	call-PST	5	time-PL

'She called five times.'

b.	*Hon*	*ring-de*	*för*	*fem-te*	*gång-en.*
	3SG	call-PST	for	5-th	time-DEF

'She called for the fifth time.'

c.	*Hon*	*är*	*fem-faldig*	*mästare.*
	3SG	be.PRS	5-fold	champion

'She is a five-time champion.'

Ordinals and frequentatives seem to be cross-linguistically common, with the notable exception of languages with restricted numeral systems (Veselinova 2004).

2.1.2 Formation of complex cardinals, ordinals and frequentatives

In languages with extensive and productive numeral systems, a considerable majority of numerals are by necessity complex. In English, for instance, all cardinal numerals are composed of multiple numeral components except the monomorphemic numerals *zero* through *nine*, *eleven* and *twelve*, the numeral base *ten* and a restricted set of powers of the base (10^2 *hundred*, 10^3 *thousand*, 10^6 *million* and 10^{100} *googol*) (Stump 2010, p. 229). English numerals belonging to other numeral series, such as ordinals, are generally derived from cardinal numerals and are therefore polymorphemic by definition: with the exception of the suppletive ordinals *first*, *second* and *third*, all ordinals are either derived using the suffix *-th* or compounds involving a suppletive ordinal.

Derivation of ordinal numerals from cardinals through affixation seems to be the most common strategy cross-linguistically as well (Veselinova 2020). In systems where suppletion does occur in ordinals, only the lowest-value ordinals are suppletive, and the number of consecutive suppletive ordinals varies between languages (Stolz and Veselinova 2013).

Frequentative numerals are also commonly derived from cardinals through affixation, but periphrastic frequentative constructions (as in (3a) and (3b)) occur as well. The frequentative marker is typically multifunctional, and often has the meaning 'times' (Veselinova 2004).

Finally, the principles for formation of complex cardinal numerals (and thus, by extension, complex ordinals and frequentatives) seem to be subject to both cross-linguistic consistency and variation. Comrie (2013) provides a generic representation $xn + y$ for complex numerals, where x and y are numerals (simple or complex) and n is the numeral base (as defined in 2.1.1). This representation is in many cases a significant simplification, as there is great cross-linguistic structural variation in the strategies used by different languages to represent a given value (see Stump 2010, p. 211).

Even within individual languages, there is often a multitude of complex numeral constructions that can all be used to represent the same value, such as the English expressions *thirty-six*, *forty minus four* and *six squared* all denoting the value 36. Despite this potential structural variety, however, these expressions are generally not used interchangeably or with equal frequency – *forty minus four* and *six squared* are both significantly more marked representations of 36 than *thirty-six*, and are used in more restricted contexts.

Similarly, *forty-five* is the preferred representation of the value 45 rather than **thirty-fifteen* or **twenty-twenty-five*, despite all three following the same pattern for complex numeral construction. Though certainly not without exception, the preferred representation is typically the one which has as large of a numerical difference as possible between its constituents. This constraint is termed the Packing Strategy by Hurford (1975, p. 67).

These observations suggest a more general cross-linguistic tendency to prefer one representation of a given value over others. Hurford (1987, pp. 269–273) attributes this tendency to the communicative advantages of using a single standardized representation, suggesting that it follows from Grice's (1975, p. 46) maxim of Manner – if no additional information of relevance is communicated by using a particular, less common representation, there is no pragmatic reason to use anything but the standardized expression.

2.1.3 Numeral frequency and 'round' numerals

As mentioned in section 2.1.2, the set of monomorphemic cardinal numerals in English is limited to the numerals 0 through 12 and certain powers of the numeral base 10. These also seem to be among the most frequently occurring English numerals. As shown in several quantitative studies using various English corpora (Francis and Kučera 1982; Johansson 1980), the numerals *one* through *ten* occur considerably more frequently than higher numerals, and there is a general tendency for frequency to decrease as numerical value increases. This tendency holds for most English cardinal numerals, with the notable exceptions of *twelve*, *fifteen* and multiples of 10 which generally occur more frequently than their neighboring numerals.

This pattern does not seem to be unique to English, as demonstrated by Dehaene and Mehler (1992) in a cross-linguistic quantitative comparison of numeral expression frequencies. All 7 languages included in the study exhibit similar patterns of frequency among cardinal numerals: frequency of occurence decreases as numerical value increases, with exceptions for sharp local increases in frequency for the numerals 10, 12, 15, 20, 50 and 100, as well as less dramatic increases for the remaining multiples of 10 (up to 90) and the powers of 10 10^3, 10^6 and 10^9.

Dehaene and Mehler (1992, pp. 13–15) assert that the regularity of this pattern cannot be fully explained by non-linguistic (mathematical or environmental) factors or by the influence of sampling artifacts, and suggest an explanation grounded in two factors: a cognitive preference for smaller numerals, and an extensive approximative use of multiples and powers of 10 (referred to as 'round' by Dehaene and Mehler (1992, p. 13)). This hypothesis is in line with findings from cognitive studies (such as Rosch 1975) suggesting that 'round' English numerals, e.g. multiples of 10, function as cognitive reference points within the cardinal numeral series.

It is important to note that Dehaene and Mehler's (1992) analysis is limited to a sparse and non-representative language sample, only containing languages from one macroarea (Eurasia) and three language families (five Indo-European, one Dravidian and one Japonic). Perhaps most crucially, only one numeral base – decimal – is represented, leading to a potentially problematic cross-linguistic conflation of 'round' numerals with multiples of the decimal base 10. Krifka (2009) broadens this perspective somewhat by comparing numeral frequencies in online text across three languages with different numeral bases: a decimal system (Norwegian), a vigesimal system (Basque) and a hybrid decimal-vigesimal system (Danish). While Norwegian numerals denoting multiples of 10 generally follow the pattern of frequency observed by Dehaene and Mehler (1992), with a local maximum at *femti* '50', this maximum does not appear in Danish or Basque, where the less morphologically complex numeral expressions denoting 40 and 60 occur significantly more frequently than those denoting 50. However, *berrogei ta hamar* '50' is still considerably more frequent in Basque text than its adjacent, more morphologically complex numerals.

Krifka (2009, p. 15) suggests that these uneven frequency distributions are the result of a bias toward two types of simplicity: morphologically simple expressions, and simple representations on specific numerical scales. Multiples (and powers) of the numeral base constitute a particularly important and coarse-grained scale, hence the differences in frequency of the numeral 50 between languages with decimal and vigesimal systems. However, other numeral scales also have equidistant reference points – these can be subdivisions of the aforementioned numeral base multiples (such as multiples of 5 in decimal or vigesimal systems), or domain-specific scales (such as the use of multiples of 12 when counting months or hours). These additional reference point numerals are visible in Krifka's (2009, p. 14) numeral frequency data for Basque, in which local frequency maxima are observed at 48 (a multiple of 12) as well as 45 and 55 (both multiples of 5).

The 'roundness' of a given numeral in a given language thus seems to depend on several interconnected factors: morphological complexity, the language's numeral base and the specific context in which the numeral is used. While some numerals (such as the aforementioned multiples of 5 and 12) would be considered 'rounder' than their neighbors in certain languages and contexts, the most fundamental or general 'round' numerals in a given language following Dehaene and Mehler (1992) and Krifka (2009) seem to be the least morphologically complex multiples and powers of a given language's numeral base.

2.2 Numerals in approximation and hyperbole

2.2.1 Vague quantification

While the prototypical function of numeral expressions is to convey an exact value or quantity, certain numerals are also used in various contexts to approximate and exaggerate, as in examples (4) and (5).

The number of bats born in Bracken Cave each summer would realistically vary from year to year, and it is extremely unlikely that the exact, 'round' numeral *ten million* in (4) would ever be precisely accurate. Similarly, although the numeral expression *tusen gånger* 'a thousand times' in (5) has a specific, exact numerical value of 1000, this is a clearly unrealistic number of times for the speaker to have repeated themselves, both in amount and 'roundness'. Despite this, neither sentence would generally be considered by a hearer or reader to be misleading or convey false information, because the numeral expressions are interpreted as approximate (and, in the case of (5), hyperbolic) in spite of their exact numerical reference.

(4) Example of approximative numeral usage in English [eng] (BNC Consortium 2007; BNC F9F 642)
Ten million bats are born each summer in Bracken Cave, Texas.

(5) Example of hyperbolic numeral usage in Swedish [swe] (own data)

Nu	*ha-r*	*jag*	*sagt*	*det*	*tusen*	*gång-er.*
now	have-PRS	1SG	say.PTCP	3SG	1000	time-PL

'I have said it a thousand times now.'

The approximation in (4) and hyperbole in (5) are both instances of vagueness of quantity. Vague expressions are widely and frequently occurring across different languages, modalities and registers. While vagueness is generally associated with spoken, informal and colloquial registers (Tárnyiková 2010, p. 73), expressions denoting vagueness of quantity are not exclusive to these domains.

Channell (1994, pp. 173–195) describes a variety of situations in which vague language, including vagueness of quantity, is commonly used. These broadly cluster into three categories:

1. situations where precision is not possible, such as when an exact quantity is not known or cannot be expressed;

2. situations where precision is not relevant, such as when a needlessly exact figure (in a particular context) can be approximated to save time or reduce cognitive load; and

3. face motivated vagueness – to communicate uncertainty, politeness or deference, or to avoid the positive face threat of having an inaccurate statement or figure corrected.

As Channell (1994, pp. 173, 192) points out, the pragmatic motivations of vagueness as described above align well with Grice's (1975, pp. 45–46) conversational maxims – both of Quality, in cases of vagueness when precise expressions might be inaccurate, and of Quantity, in cases of vagueness when precise expressions might be extraneous.

A variety of different strategies can be used to refer to quantities vaguely. These range from approximation through numeral expressions with specific numerical values (as in (4)) to the use of non-numerical quantifiers like *a bunch of*, and represent diverse, often contextually defined approximate quantities. Channell (1994, pp. 95–96) groups non-numerical vague quantifiers into the broad categories (+ for quantity), (neutral) and (– for quantity) based on the relative sizes of the quantities they represent – (+ for quantity) expressions like *lots of* refer to larger, sometimes exaggerated quantities, while (– for quantity) expressions such as *a bit of* refer to smaller and minimized quantities.

Using this taxonomy, Tárnyiková (2010) investigates which nouns commonly co-occur with various non-numerical vague quantifiers in English and Czech, and finds significant variation in noun collocates both across and within Channell's categories of quantity. Tárnyiková (2010, pp. 82–83) also finds similarities in noun collocates between specific vague quantifiers in English and Czech, suggesting some cross-linguistic commonalities in the strategies used for vague quantification of particular nouns.

There is also considerable variation in the ways quantities can be approximated using numeral expressions. Most fundamentally, certain numerical values are used more commonly than others in approximative contexts. 'Round' numerals, as discussed in section 2.1.3, are particularly prone to approximative usage since they function as cognitive reference points, and are thus more likely to be interpreted as approximations than other numerals – the less morphologically complex a numeral is and the coarser the numerical scale on which it exists, the likelier the approximate interpretation of it becomes (Krifka 2009, pp. 8–10).

Plural numeral expressions such as *thousands* and *millions*, which inherently express a range rather than an exact numerical value, are also frequently used in approximation, although Channell (1994, p. 91) notes that only specific numerals – typically higher-value monomorphemic numerals – are used in this way.

Finally, approximators such as the English expressions *like* or *about* are commonly used in conjunction with a numeral to unambiguously indicate approximation.

2.2.2 Hyperbolic quantification

Hyperbole is a rhetorical device involving deliberate exaggeration. In contrast to non-hyperbolic vague language, which can in some contexts be a *more* accurate or efficient way of conveying information than precise language, hyperbole by necessity involves a conscious misrepresentation of a fact. While this would seem to violate the Gricean maxim of Quality, McCarthy and Carter (2004, pp. 152–153) suggest that the overtness of the falsehood in hyperbolic utterances distinguishes them from other types of untruthful statements – they are clearly not intended to be interpreted as factual. Nevertheless, this property of exaggeration means that hyperbole is somewhat more domain specific to informal (and typically spoken) contexts than other types of vagueness (McCarthy and Carter 2004, p. 150).

As with other expressions of vagueness, however, hyperbole is used in a variety of informal settings and contexts. Although hyperbole is typically associated with irony, it also occurs in non-ironic narrative, descriptive and evaluative contexts, particularly when emphasizing a contrast (McCarthy and Carter 2004, p. 158).

Within the domain of vague quantification, hyperbole is used to exaggerate the size of a quantity through either maximizing (auxesis) or minimizing (meiosis) (McCarthy and Carter 2004, p. 151). [1] Although hyperbolic and approximative quantification are naturally closely related concepts, there are instances of both non-hyperbolic approximative quantification (such as in (4), where there is no deliberate exaggeration) and non-approximative hyperbolic quantification (such as the hyperbolic quantifier *zillions*, which has no explicit or precise referent to which a quantity is approximated).

Both numeral expressions and non-numerical quantifiers can be used in hyperbolic contexts: Tárnyiková (2010, p. 82) mentions exaggeration and boasting as particularly common uses of non-numerical quantifiers in the (+ for quantity) category, both in English and in Czech, and a prototypical example of hyperbolic usage of numerals is given in (5).

While the approximating properties of approximators such as the English expressions *like* and *around* are presumably lost when they are used in hyperbole, approximators still occur in hyperbolic quantifier constructions (Lavric 2010, p. 144). This holds true for both hyperbolic numerals, as exemplified in (1), and non-numerical quantifiers.

McCarthy and Carter (2004) investigate hyperbolic uses of various English numerical and non-numerical quantifiers in a corpus of informal spoken language, and find significant variation in both the absolute frequency and proportion of hyperbolic usage between individual quantifiers. Although 'round' numerals which are frequently used in other approximative contexts (as discussed in section 2.2.1) seem to be the most common hyperbolic numerals in general, there is significant intra-numeral variation – *dozen* and *million* are used hyperbolically in 30–32 percent of analyzed occurrences whereas *thousand* and *hundred* are only hyperbolic in 3–5 percent of occurrences. As with approximative numeral quantifiers, plural forms are also found to be particularly hyperbole-prone, both in frequency and proportion – *dozens* and *millions* are nearly exclusively used hyperbolically.

Very little systematic work has been done on hyperbolic uses of numerals from a cross-linguistic perspective. Lavric (2010), the only cross-linguistic study of this phenomenon, investigates the values and contexts of hyperbolic numerals in French, English, Italian, Spanish and German, and finds a variety of numerals used in both minimizing and maximizing hyperbolic contexts, corresponding to Channell's (1994) categories of quantity for non-numerical quantifiers. Lavric (2010, p. 132) also finds that hyperbolic quantification occurs in a broad variety of contexts – while some types of nouns (such as frequentative markers, as exemplified in (3), and units of time measurement like *years*) are particularly frequent in hyperbolic quantification, nearly any quantifiable noun can be hyperbolically quantified.

In line with McCarthy and Carter (2004), 'round' hyperbolic numerals seem to be particularly common cross-linguistically, but Lavric also finds other cross-linguistically consistent patterns in value for hyperbolic numerals. These include adding 1 to or subtracting 1 from a given 'round' numeral, as exemplified by numerals like French *mille et une* '1001' and Italian *novantanove* '99'. Lavric suggests that these patterns are, if not domain specific, at least par-

[1] While there are differing views on whether the term hyperbole encompasses both auxesis and meiosis or only auxesis, this study uses the broader definition of hyperbole as used in McCarthy and Carter 2004 and Lavric 2010.

ticularly prominent in specific contexts – for instance, hyperbolic numerals of the $N-1$ type are used mainly in hyperbolic expressions of percentage or chance. Similar restricted usage is found in hyperbolic numerals of the type which Lavric calls 'more-than-totality' ($(N+1)/N$), as in French *vingt-cinq heures sur vingt-quatre* 'twenty-five hours out of twenty-four'. These numerals require the existence of a contextually relevant reference numeral N to which the hyperbolic numeral is related, such as *vingt-quatre* '24' denoting the number of hours in a day in the above example.

A particularly interesting hyperbolic numeral is *trente-six* '36' in French, which is idiomatically used to convey a (specifically maximizing) hyperbolic meaning in specific and limited contexts, often involving negation (as demonstrated in example (6)). Unlike all other maximizing hyperbolic numerals and patterns covered by Lavric (2010), *trente-six* is not self-evidently related to a particular reference point numeral.

(6) Example of hyperbolic usage of *trente-six* in French [fra] (Lavric 2010, p. 125, own gloss)

il	*n=y*	*a*	*pas*	*trente-six*	*façon-s*	*de*	*voir*	*la*	*chose*
3SG	NEG=LOC	have.PRS	NEG	36	way-PL	to	see.INF	the	thing

'There aren't many ways of seeing the thing.'

Lavric (2010) does not provide any extensively developed suggestions as to why this particular numeral (rather than, for instance, its neighbours *trente-cinq* or *trente-sept*) is used hyperbolically in a conventionalized manner, aside from highlighting that 36 is a multiple of 12. This property places 36 as a reference point numeral on a common and relatively coarse-grained numerical scale following Krifka's (2009, pp. 11–12) definition, but still does not explain why numerals denoting other multiples of 12, such as *vingt-quatre* '24' or *quarante-huit* '48' are not used hyperbolically in the same conventionalized manner.

A similarly unexplained exact numerical value used hyperbolically is the Danish numeral *hundredesytten* '117', which occurs frequently in a variety of hyperbolic contexts. Unlike *trente-six*, there is no clear coarse-grained numerical scale on which *hundredesytten* can function as a reference point numeral, and no other etymological explanations for the choice of value have been proposed.

Despite this lack of sufficient explanation, both *trente-six* and *hundredesytten* are clearly conventionalized in hyperbolic usage and function similarly to 'round' and perhaps more prototypical hyperbolic numerals, despite lacking their non-hyperbolic approximative usage.

2.2.3 Quasi-numerals

In their survey of English expressions commonly used in hyperbolic contexts, McCarthy and Carter (2004) briefly discuss the expression *zillions*, which they refer to as a 'colloquial formation', 'pragmatically specialised for hyperbole' (McCarthy and Carter 2004, p. 167). The core of this pragmatic specialisation, distinguishing *zillions* from other hyperbolic quantifiers, is that it lacks either an exact numerical value (like *million* '10^6' and other hyperbolic numeral expressions) or a literal reference (like *heaps of* and other hyperbolic non-numeral quantifiers). As a result, it occurs exclusively in hyperbolic contexts.

Lavric (2010, p. 139) also briefly mentions *zillions* as a hyperbolic expression distinguished from (yet related to) hyperbolic uses of exact numerals, comparing it to the similarly used English expression *umpteen* and various (mainly frequentative) expressions with a variable-marking letter in place of a numeral, such as Italian *per l'ennesima volta* 'for the *n*th time' and Spanish *equis veces* 'X times'.

While these expressions are all used exclusively as vague quantifiers in hyperbolic (maximizing) contexts and lack an exact numerical value or a literal referent, they also share both lexical and structural properties with numerals that have exact values. *Zillion* and *umpteen* are constructed along the same patterns as English numerals denoting higher powers of 10 (*million*, *billion*, *trillion*) and the values 13–19 (*thirteen* to *nineteen*) respectively, and *ennesima* '*n*th' follows the pattern of derivation of ordinal numerals in Italian. Based on these similarities, the expressions discussed above can be seen as members of a distinct category of hyperbolic *quasi-numerals* with shared structural and functional properties.

A subset of these quasi-numeral expressions are analyzed in greater depth by Chrisomalis (2016), who provides a diachronic account of what he terms IHNs – indefinite hyperbolic numerals – in English. Chrisomalis defines IHNs using similar criteria to those discussed above for quasi-numerals in general: while IHNs lack direct numerical referents, they resemble numerals in both morphological form and usage. IHNs are also distinguished from hyperbolic non-numerical quantifiers by their ability to be used as constituents in complex (quasi-)numeral constructions together with other numerals (such as *umpteen thousand*), which is generally not the case for other quantifiers (e.g. **heap thousand*). Which numerals can be used to form complex constructions also seems to vary between quasi-numerals – for instance, *umpteen* can only occur in constructions where other *-teen* numerals could also occur, such as *umpteen thousand* but not **six umpteen*.

With this property in mind, Chrisomalis (2016, p. 7) makes a further categorical distinction between 'major' IHNs (such as *zillion*) and 'minor' IHNs (such as *umpteen*). While the primary differences between major and minor IHNs relate to their morphological composition – major IHNs are constructed using the *-illion* pseudo-morpheme present in numerals denoting higher powers of 10, while minor IHNs 'pattern after the teens or the decades' (Chrisomalis 2016, p. 7) – Chrisomalis suggests that these morphological differences are reflected in the magnitude of hyperbole as well, with major IHNs representing larger indefinite quantities than minor IHNs. In addition, the hyperbolic quantities represented by some major IHNs like *zillion* and *jillion* can be magnified further by appending 'intensifier' prefixes like *ba-* and *ga-*, while no minor IHNs seem to be modifiable in this way.

The oldest attested English quasi-numeral expression Chrisomalis finds is *forty-leven*, which was first attested in print in 1839. This expression is particularly notable in that it is constructed purely from numerals with exact values, in contrast to expressions like *zillion* or *umpteen*. Although *forty-leven* could be analyzed as a regular complex cardinal numeral construction $40 + 11$, denoting the value 51, it would be a non-standard representation of this value and also violate the difference-maximizing constraints of Hurford's (1975, p. 67) Packing Strategy. There do not seem to be records of systematic use of *forty-leven* in anything but hyperbolic contexts, which supports its classification as a quasi-numeral rather than a numeral despite its theoretical value.

This complex-numeral-mirroring pattern of quasi-numeral construction seems relatively unique to *forty-leven*, which is described by Chrisomalis (2016, p. 9) as an 'idiomatic and not particularly productive' minor IHN. However, as he subsequently points out, *seventy-eleven* has also been attested, and *fifty-eleven* is present in contemporary song lyrics, as in example (7), suggesting some potential variety in at least which multiple of 10 is chosen.

(7) Example of hyperbolic usage of *fifty-eleven* in English [eng] (Chrisomalis 2016, p. 10)
See me up in the club with fifty-eleven girls / Posted in the back, diamond fangs in my grills

Notably, these observations (and by extension the subcategorization of IHNs into major and minor types) only concern quasi-numeral expressions in English. Although Chrisomalis asserts that ‘no other language has such an extensive lexicon of IHN as English’ (Chrisomalis 2016, p. 7), he does not provide any quantitative evidence for this claim, let alone the type of exhaustive and systematic cross-linguistic survey that would be necessary to credibly make this assertion. While the lexicon of quasi-numeral expressions in English has certainly been most thoroughly investigated, structurally and functionally equivalent expressions to English quasi-numerals like *umpteen* and *zillion* have been attested in other languages. These cross-linguistic observations include both prototypical major IHNs such as *fantastilione* in Italian (which is directly mentioned by Chrisomalis (2016, pp. 6–7)), and prototypical minor IHNs such as *ørten* in Norwegian (Språkrådet and University of Bergen 2022).

Similarly, *forty-leven* has a Swedish equivalent in *femtioelva* ‘fifty-eleven’, which is constructed in the same convention-breaking manner as *forty-leven* and occurs frequently in informal language. Tárnyiková (2010, p. 83) also identifies the Czech expression *x-krát* ‘*x*-times’ which clearly belongs in the same category of quasi-numeral as the variable constructions pointed out by Lavric (2010, p. 139).

It is thus clear that conventionalized quasi-numeral expressions with explicitly and exclusively hyperbolic uses exist in more languages than those accounted for in prior work. As evidenced by all hitherto discussed examples being from Indo-European languages, the broader cross-linguistic distribution of these expressions is even more unexplored.

2.3 Aims and research questions

The goal of this study is to conduct an exploratory cross-linguistic survey and investigation of the distribution, value, morphological structure and usage of hyperbolic numerals and quasi-numeral expressions. The survey will be performed across a diverse language sample, through extraction and analysis of these expressions in massively parallel texts.

This aim can be divided into five distinct research questions:

Research question 1: What is the cross-linguistic distribution of hyperbolic numerals and quasi-numeral expressions?

Research question 2: Are there cross-linguistic tendencies in the numerical value of conventionalized hyperbolic numerals?

Research question 3: Are there cross-linguistically consistent patterns of construction for hyperbolic quasi-numeral expressions?

Research question 4: Is there areal or genealogical cross-linguistic variation in which types of hyperbolic numerals and quasi-numeral expressions occur?

Research question 5: Is there variation in the contexts in which different hyperbolic numerals and quasi-numeral expressions occur?

3 Method

In this chapter, the data sources, language sample and methodology used in this study are presented. Section 3.1 presents and motivates the types of primary and secondary data sources used, and outlines the structure of the study's main data source. In section 3.2, the language sample is presented and briefly discussed. In sections 3.3 and 3.4, the methods used for extraction, elicitation and analysis of relevant expressions for each research question are explained. Finally, section 3.5 provides a brief summary of the procedure used in this study.

3.1 Data

The main primary data source for this study is parallel text data. Primary data is preferable as a main source when studying phenomena that are not yet well-defined or commonly accounted for in secondary sources such as grammars (Stolz 2007, pp. 101–102). In addition to parallel texts, there are several different types of primary data used in typological research, including original texts and various forms of questionnaires (Wälchli and Cysouw 2012, pp. 673–674). Of these, parallel text data provides the unique advantage of functionally parallel (and therefore easily comparable) examples across a greater number of languages than would be feasible with any other primary data source. This advantage is particularly clear given the recent prominence of what Cysouw and Wälchli (2007, p. 95) term 'massively parallel texts' – texts with aligned translations in many different languages, enabling large-scale cross-linguistic comparison of individual examples.

The parallel text data used in this study was gathered from the OpenSubtitles2018 parallel corpus, which is part of the OPUS collection of freely available parallel data (Lison and Tiedemann 2016; Tiedemann 2012). OpenSubtitles2018 is a massively parallel corpus of TV and film subtitles, containing large amounts of document- and sentence-aligned data in 62 languages. The corpus consists of bitexts in all possible language pairs for which subtitles for shared sources exist in the *OpenSubtitles* database [2]. The bitexts are aligned at sentence level through a multi-step approach which utilises time-slot overlap in the original subtitle files and language-pair specific dictionaries created from automatic word alignments (Tiedemann 2008). Cross-lingual links have also been generated at both document and sentence level to allow for simultaneous retrieval of specific sentences across more than two languages at once (Tiedemann 2016).

In addition to the full bitexts, several pre-processed resources compiled from OpenSubtitles data, including automatic word alignments and phrase translation tables, are freely available through the OPUS database. The phrase translation tables are generated using the Moses toolkit for statistical machine translation (Koehn, Zens et al. 2007), and contain a list of co-occurring n-grams across any two languages in the parallel corpus (for which at least one parallel text is available) along with their co-occurrence frequencies and word alignments between the respective n-grams.

As with any study where corpus data is used to investigate linguistic phenomena that are known (or presumed) to be domain-specific in some regard, care should be taken to ensure that the doculect of the texts used appropriately reflects the domain of interest. In the case of this study, as has been discussed in section 2.2, hyperbolic expressions are most commonly found in informal, spoken language (McCarthy and Carter 2004, p. 150). Most larger massively parallel corpora consist of written language, often in specialized registers such as parliamentary proceedings or Bible texts (Cysouw and Wälchli 2007, p. 97), and would therefore not be ideal

[2] http://www.opensubtitles.org/

for investigating this phenomenon. TV and film subtitles, meanwhile, represent a spoken, broad and typically informal language domain. This property makes the *OpenSubtitles2018* parallel corpus better suited for this study than other parallel corpora of comparable size and language diversity.

The use of massively parallel texts is advantageous in many ways, but there are also several issues that make relying solely on parallel text data seem unwise. For instance, the distribution of available languages (and amount of available parallel texts in each language) in many massively parallel corpora, including OpenSubtitles2018, is considerably biased towards European and high-resource languages. In addition, the use of translated data necessitates that source language inference be controlled for, to ensure the validity of cross-linguistic comparisons (Cysouw and Wälchli 2007, pp. 98–99). For these reasons, additional primary data sources were used to complement the OpenSubtitles data: a limited elicitation (described further in section 3.3.3) was conducted with the aim to improve the areal and genealogical coverage of the language sample, and monolingual corpus data was used to boost the validity of conclusions drawn from the parallel text data.

In the selection of monolingual corpora, size, comparability and appropriate register was prioritized over the presence of annotation. For these reasons, large web crawl corpora constructed using the TenTen and Web as Corpus (WaC) architectures were used for the vast majority of languages in the sample where such corpora were publicly available. These corpora are constructed using systematic and comparable web crawl methods, as described in Jakubíček et al. 2013 and Kilgarriff, Reddy et al. 2010. For languages without TenTen or WaC corpora where hyperbolic numerals or quasi-numerals requiring monolingual frequency analysis were identified, other large web crawl corpora were used. All monolingual corpora were accessed and queried through the ‘Sketch Engine’ web service (Kilgarriff, Baisa et al. 2014).

In addition to monolingual corpora, dictionaries and grammars were also used in the analysis of the construction and value of hyperbolic numerals/quasi-numeral expressions. A full overview of primary and secondary data sources used for each language included in the study can be found in Appendix A.

3.2 Sample

The language sample used in this study is a convenience sample based on the 62 subcorpora in 58 languages available in the OpenSubtitles2018 parallel corpus. As one of the aims of this study is to assemble a database of hyperbolic numerals and quasi-numeral expressions in as many languages as possible, no language in which sufficient OpenSubtitles data was available was excluded from the sample.

As will be described further in the following section, the procedure for extracting candidate hyperbolic numerals/quasi-numeral expressions from parallel text data in a given language relies on the availability of accurately aligned sentence pairs containing an already identified hyperbolic numeral/quasi-numeral expression in another language. As a result, 9 languages were excluded from the sample because of a lack of sufficient data. Following the elicitation procedure described in section 3.3.3, candidate hyperbolic numerals/quasi-numeral expressions were obtained in 1 additional language, bringing the number of languages included in the study to a total of 50. The ISO 639-3 codes for all languages in the sample are listed in Table 1, and a genealogical breakdown of the sample is provided in Appendix A. Figure 1 displays the areal distribution of languages in the sample, along with an overview of which types of data (*OpenSubtitles* and/or elicitation) were analyzed for each language.

As shown in the column *OpenSubtitles corpus size* in Appendix A, the amount of *Open-*

Language sample and data types

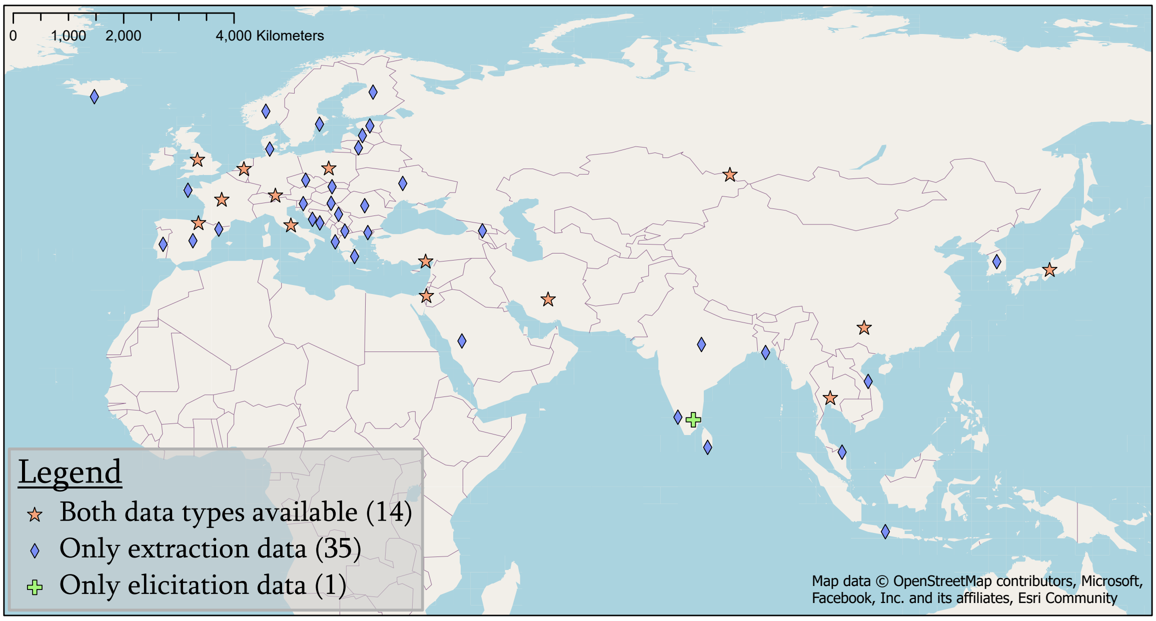

0 1,000 2,000 4,000 Kilometers
Legend
☆ Both data types available (14)
◊ Only extraction data (35)
✚ Only elicitation data (1)
Map data © OpenStreetMap contributors, Microsoft, Facebook, Inc. and its affiliates, Esri Community

Figure 1: Visualization of all languages in the sample, by the types of data available for analysis

Subtitles data available varies considerably between languages – the English subcorpus, for instance, contains nearly 1000 times as much data as the Bengali subcorpus. Of course, this impacts both the possible coverage of intra-lingual investigation in languages with smaller subcorpora and the validity of cross-linguistic comparisons. The language sample is also not stratified, resulting in a distinctly uneven genealogical and areal distribution of languages. Of the 50 languages included in the study (as displayed in Table 1), 33 are Indo-European, and all 50 are primarily spoken in the Eurasian macroarea. The consequences of these uneven distributions of languages and data are discussed further in section 5.6.

In order to make the investigation of genealogical variation meaningful despite the unbalanced sample, analysis is performed at genus level, following Dryer 1992, rather than at language family level. Genera are maximal groupings of related languages at a time depth no greater than 4000 years (Dryer 1992), and allow for somewhat more balanced genealogical comparisons. While there are a number of genus classification paradigms in use, the classification used in Table 1 and Appendix A follows that of WALS (Dryer and Haspelmath 2013).

Table 1: Language sample, organized by family and genus

Language family	Number of languages	Genus	Number of languages	Languages (ISO 639-3)
Indo-European	33	Slavic	11	bul, bos, ces, hrv, mkd, pol, rus, slk, slv, srp, ukr
		Germanic	7	dan, deu, eng, isl, nld, nob, swe
		Romance	6	cat, fra, ita, por, ron, spa
		Indic	3	ben, hin, sin
		Baltic	2	lav, lit
		Albanian	1	als
		Celtic	1	bre
		Greek	1	ell
		Iranian	1	pes
Uralic	3	Finnic	2	est, fin
		Ugric	1	hun
Afro-Asiatic	2	Semitic	2	arb, heb
Austronesian	2	Malayo-Sumbawan	2	ind, msa
Dravidian	2	Southern Dravidian	2	mly, tam
Sino-Tibetan	1	Chinese	1	cmn
Austroasiatic	1	Viet-Muong	1	vie
Basque (isolate)	1	Basque	1	eus
Japonic	1	Japanese	1	jpn
Kartvelian	1	Kartvelian	1	kat
Koreanic	1	Korean	1	kor
Tai-Kadai	1	Kam-Tai	1	tha
Turkic	1	Turkic	1	tur

3.3 Procedure

Given the exploratory nature of this study, the procedure for extraction of candidate hyperbolic numerals/quasi-numeral expressions was divided up into two parts: a pilot study, establishing common contexts and seed expressions in a limited set of languages, and a secondary procedure in which the extraction was extended to the full language sample. Both procedures were performed using the same parallel text data from the *OpenSubtitles2018* corpus.

3.3.1 Pilot study

Prior to the extraction of candidate expressions from the full set of parallel text data, a limited pilot study was carried out on OpenSubtitles data in a convenience sample of 6 languages: Danish, English, French, Japanese, Norwegian and Swedish. The pilot study had three goals: (i) to identify a number of hyperbolic numerals/quasi-numeral expressions for use as seed expressions in the main study's extraction procedure, (ii) to evaluate the criteria used for categorizing expressions as hyperbolic numerals or quasi-numeral expressions, and (iii) to identify preliminary cross-linguistic patterns in the contexts in which the identified expressions commonly occur.

Using the corpus query interface available through OPUS (Tiedemann 2012), regular expressions matching the hyperbolic quasi-numerals *umpteen* (English), *ørten* (Norwegian) and *femtioelva* (Swedish) as well as common variants and derivations were searched for in all OpenSubtitles parallel texts available in the languages of the searched-for expressions. Once a sentence containing one of the searched-for expressions was identified, the corresponding expressions were manually extracted from all available translations of the sentence. In instances where an extracted expression was used to quantify a noun, the form of the quantified noun itself was also recorded. The regular expression search terms used and their target matches are listed in Appendix C.

After all sentences returned by the corpus query were processed in this way, the extracted hyperbolic expressions in all 6 languages were categorized as either HN (hyperbolic numeral) or HqN (hyperbolic quasi-numeral) according to the following criteria: If an expression in a given language was found to denote an exact numerical value, it was categorized as HN. Otherwise, if numeral constituents or derivational patterns used in the language's numeral system were identified in analysis of the expression's morphological composition, and the expression was found to predominantly or exclusively occur in maximizing hyperbolic contexts, it was categorized as HqN. Expressions which did not meet any of these criteria were categorized as non-numerical quantifiers and were not retained for further analysis.

The number of occurrences of each expression and quantified noun was also recorded.

3.3.2 Full extraction of candidate expressions

Following the pilot study, another extraction of candidate hyperbolic numeral/quasi-numeral expressions was performed across all languages available in *OpenSubtitles2018*, using expressions identified in the pilot study as seed expressions. All hyperbolic quasi-numeral expressions identified in the pilot study were added as seed expressions. Ordinal and frequentative forms of the Danish hyperbolic numeral *hundredesytten* '117' were also included, since additional searches in OpenSubtitles data revealed that these were nearly exclusively used in hyperbolic contexts. Other hyperbolic numerals with exact numerical values were not used as seed expressions, as their potentially broad non-hyperbolic usage would require additional analysis of the context of each search result to determine whether or not they are used hyperbolically. As in the pilot study, regular expressions matching these seed numeral expressions were constructed and are listed in Appendix C.

To aid the extraction and analysis steps, pre-compiled phrase translation tables generated from OpenSubtitles data (as described in section 3.1) were queried instead of the full aligned bitexts used in the pilot study. Phrase translation tables for all available language pairs including at least one language used in the pilot study were obtained through the OPUS API (Aulamo et al. 2020). While most languages had available bitexts (and therefore also phrase translation tables) for all pilot languages, some language combinations did not exist in *OpenSubtitles2018* and could therefore not be included in the extraction. Phrase translation tables for pairs consisting of two languages included in the pilot were also included – as phrase translation tables generated using the Moses toolkit are bidirectionally symmetrical, the same table could be used for extraction in either language.

In total, phrase translation tables for 224 language pairs were obtained, corresponding to the number of language pairs in which hits for the searched-for seed expressions were found. Table 2 shows how many language pairs (and, by extension, phrase translation tables) with sufficient data for analysis were available for each of the seed expression languages.

Table 2: Number of languages in which searches for seed expression returned hits in *OpenSubtitles2018* data, for each language of seed expressions

Language of seed expressions	**Number of languages with sufficient data**
Danish	43
English	48
French	48
Norwegian	41
Swedish	44

Phrase pairs containing one of the seed expressions represented by the regular expressions in Appendix C were identified in each phrase translation table, and numeral/quasi-numeral expressions in the corresponding phrases were extracted using the automatic word alignments present in the table. A probability threshold of 0.0001 was applied to both the inverse and direct phrase translation probabilities in order to filter out phrases resulting from word alignment errors or misaligned subtitles. In cases where this threshold was insufficient to determine whether an extracted correspondence was spurious (typically for language pairs with too little parallel data to generate any phrase translation probabilities below the threshold), language resources listed in Appendix A were consulted.

For language pairs where no phrases containing seed expressions were found in the phrase translation tables, manual corpus queries were performed to ensure that no singleton matches had been ignored in the generation of the tables. If a match was found, the numeral/quasi-numeral expression in the corresponding sentence was manually extracted using dictionaries.

3.3.3 Elicitation

Parallel to the extraction of candidate hyperbolic numeral and quasi-numeral expressions from OpenSubtitles2018 data, an elicitation query was distributed through two prominent typology-centric mailing lists, LINGUISTList and Lingtyp. The goal of this query was to elicit hyperbolic numeral and quasi-numeral expressions in any language, corresponding to a set of examples in Swedish, French, Danish and English. The text of the elicitation query can be found in Appendix D.

In total, 35 responses were received, resulting in a total of 31 elicited expressions deemed relevant to the study across 15 languages. While 13 of these expressions were also present in the list of extractions from OpenSubtitles data, the remaining 18 expressions were not identified in the extraction procedure.

These expressions were added to the list of extracted candidate expressions and analyzed according to the same criteria. The languages in which elicited expressions were obtained are displayed in Figure 1, and listed along with language consultants' names and dates of correspondence in Appendix A.

3.4 Analysis

Once a list of candidate hyperbolic numeral and quasi-numeral expressions had been compiled from the extraction and elicitation procedures, all identified word forms of each candidate quasi-numeral expression were searched for in the monolingual web corpora listed in Appendix A in order to confirm their usage in original non-translated text. [3] Candidate expressions that did not occur at least once in their corresponding corpora were not retained for further analysis, and expressions occurring fewer than 5 times were subject to manual analysis of randomly sampled sentences to verify their occurrence in authentic contexts.

3.4.1 RQ1: Distribution of hyperbolic numerals and quasi-numerals

To analyze the cross-linguistic distribution of hyperbolic numerals and quasi-numeral expressions, the retained expressions were categorized as either HN or HqN according to the criteria used in 3.3.1, using dictionaries and other language resources as detailed in Appendix A.

The presence of HN and HqN expressions in each language in the sample was subsequently mapped and analyzed to identify areal, genealogical and categorical tendencies.

3.4.2 RQ2: Numerical value of hyperbolic numerals

To investigate tendencies in the value of hyperbolic numerals, the numerical values of all extracted expressions categorized as HN were analyzed using dictionaries and numeral databases, and commonly occurring values and potential patterns in value were identified.

Differing patterns of occurrence between round hyperbolic numerals (such as French *mille* '10^3') and non-round numerals (such as Danish *hundredesytten* '117') in the pilot study implied that it would be relevant to distinguish between these two categories in the full analysis of the values of extracted hyperbolic numerals as well. For reasons discussed in 2.1.3, it is difficult to establish a value-based, cross-linguistically consistent definition of numeral roundness which takes all involved factors into account. With this in mind, non-exhaustive base-derived roundness criteria were established following Dehaene and Mehler 1992 and Krifka 2009: multiples of a given language's numeral base n up to n^2 (above which morphological complexity generally increases) and all powers of n were categorized as round numerals, and all other identified numerals were categorized as non-round numerals.

[3] Only quasi-numerals were searched for in monolingual corpora, since they (unlike hyperbolic numerals) would not be expected to occur in any non-hyperbolic contexts.

3.4.3 RQ3: Construction of hyperbolic quasi-numeral expressions

To investigate the construction of hyperbolic quasi-numerals, the morphological composition of all extracted expressions categorized as HqN was analyzed using dictionaries and other language descriptions, and commonly occurring patterns of construction were identified.

3.4.4 RQ4: Distribution of types of HN/HqN expressions

Analysis of the cross-linguistic distribution of the identified types of hyperbolic numerals and quasi-numeral expressions was performed at genus level following Dryer 1992, given the uneven genealogical distribution of languages in the sample as described in 3.2.

For each category of HN/HqN established in the analysis of RQ2 and RQ3, the number of languages in which expressions of this category had been found was counted and compared across genera.

3.4.5 RQ5: Patterns in usage of HN/HqN expressions

To investigate whether different hyperbolic numerals and quasi-numeral expressions have different common contexts of occurrence, expression-to-expression translation probabilities were calculated between all expressions identified in the pilot study. Pairs of expressions with higher translation probabilities were used as translations for each other to a higher degree in the pilot study data, and were as such found to co-occur more frequently in identical contexts.

For a given pair of expressions t_A and t_B in languages A and B, translation probability was calculated as the mean of the direct translation probability $\varphi(t_A|t_B)$ and the reverse translation probability $\varphi(t_B|t_A)$.

3.5 Summary

In this study, hyperbolic numeral and quasi-numeral expressions in a convenience sample of 50 languages were extracted from *OpenSubtitles2018* parallel text data and elicited through mailing lists.

The extraction procedure was conducted in two steps. First, expressions were manually extracted from the parallel data in 6 pilot languages by searching for translated contexts matching the seed expressions *umpteen* (English), *ørten* (Norwegian) and *femtioelva* (Swedish). All hyperbolic quasi-numeral expressions identified in this way were subsequently used as seed expressions for the extraction of expressions across all 50 languages in the sample. In the second extraction, expressions were extracted using phrase translation tables and word alignments compiled from *OpenSubtitles* parallel data.

Extracted and elicited expressions were verified through searches in monolingual web corpora, and finally analyzed in relation to prior work and the study's research questions.

4 Results

In this chapter, the results of both the pilot study and the full extraction and subsequent analysis of hyperbolic numerals and quasi-numerals across the entire language sample are presented. The findings of the pilot study are presented in section 4.1, followed by a presentation of results for each research question.

A full table of extracted hyperbolic numeral and quasi-numeral expressions is provided in Appendix B.

4.1 Pilot study

Using the seed expressions *umpteen*, *ørten* and *femtioelva* as described in section 3.3.1, a total of 99 contexts with translations in at least 2 of the pilot languages were obtained for analysis. Each context consisted of a sentence or sentence fragment corresponding to a single subtitle in the source data. An example context is given in (8).

(8) Danish [dan] (*OpenSubtitles2018* movie 421947, sub 234307)
For 117. gang sagde jeg nej.
for 117.th time say.PST 1SG no

For the umpteenth time I said no.

As not all contexts in the *OpenSubtitles* parallel data have been translated into all pilot languages, the number of sentences obtained varied between languages, as shown in Table 3. For Japanese, equivalent sentences were only available in 4 of the 99 contexts, which meant that significantly fewer Japanese expressions could be extracted.

Table 3: Number of pilot contexts analyzed for each pilot language

Language	Number of contexts
French	78
English	73
Swedish	62
Danish	47
Norwegian	46
Japanese	4

A variety of quantifying functions and quantified nouns were identified in the pilot contexts. Table 4 shows the number of contexts containing quantified nouns in a number of distinct categories.

The overwhelmingly most common quantified nouns were iterative units (such as *gang* in example (8)), which are used as frequentative markers in periphrastic constructions of frequentative numerals in all pilot languages. Outside of this category, no individual noun occurred more than twice in the pilot data.

The second most common category of quantified noun, People, includes nouns denoting specific groups (such as *kids* or *exes*) as well as members of a lineage, as in example (9).

In three contexts, no quantified noun was present. This had two distinct causes: either a previously mentioned noun was omitted, or the (quasi-)numeral was not used in a quantifying function (as in the sentence ‘How many is umpteen?’).

Table 4: Frequency of categories of quantified nouns in the pilot contexts

Category of quantified noun	**Frequency**
Iteration of an event (e.g. *times*)	64
People (e.g. *inmates*)	6
Events (e.g. *politics lecture*)	5
Objects (e.g. *class A drugs*)	5
Units of time (e.g. *weeks*)	3
Money (e.g. *dollars*)	2
Other nouns	11
No quantified noun	3

(9) French [fra] (*OpenSubtitles2018* movie 1232790, sub 3483871)

Jean	*XXIII,*	*et*	*Clitoris*	*le*	*Én-ième,*	*ce*	*qui*	*suffi-t*	*à*	*vous*	*rendre*	*païen.*
Jean	23	and	Clitoris	the	*n*-th	3SG	which	suffice-3.PRS	to	2PL	render.INF	pagan

John XXIII and Clitoris the umpteenth, which is enough to turn anyone pagan.

The distribution of cardinal and ordinal functions of the quantifiers identified in the pilot contexts also varied significantly, as detailed in Table 5. In contexts where the quantified noun was a frequentative marker, ordinal functions (such as *117. gang* 'the umpteenth time' in (8)) occurred more than twice as frequently as cardinal functions (such as *umpteen times*). In all other contexts, however, cardinals occurred more frequently than ordinals.

Table 5: Frequency of cardinal and ordinal functions of quantifiers in the pilot contexts [4]

Function of quantifier	**Frequency**
Ordinal (frequentative)	44
Cardinal (non-frequentative)	21
Cardinal (frequentative)	20
Ordinal (non-frequentative)	14

Finally, a variety of hyperbolic quantifiers were extracted from the pilot contexts. Table 6 lists the extracted word forms of each quantifier expression present in the pilot contexts, with expressions which occurred more than once listed in bold. Although hyperbolic numerals were found in all pilot languages, and quasi-numerals in all languages except Japanese, the number of identified expressions of each type varied greatly between languages.

A number of non-numerical hyperbolic quantifiers in all 6 languages were also extracted. While the hyperbolic function of these expressions corresponds to that of the hyperbolic quasi-numeral seed expressions, they differ from the observed quasi-numerals in both semantic and structural properties, either by not resembling numerals or by being usable in non-hyperbolic literal or approximative contexts.

[4] Although quantifier function was cross-linguistically consistent for each pilot context (i.e., a single context did not contain different quantifier functions in different languages), the number of translated contexts differed between languages (as displayed in Table 3), meaning that the total distribution as displayed in this table may not reflect the distribution for each individual language.

Table 6: Extracted hyperbolic quantifier expressions in the pilot study, by category

Category	Language	Expressions
Hyperbolic numerals	dan	*dusiner*, *syttende*, ***50***, *to og halvtreds*, *100*, ***117***, ***hundrede og syttende***, ***hundredesyttende***, *400.*, *10.000*, *mange millioner*, *bilion*
	eng	*dozen*, *fifteenth*, *17th*, *46*, ***50***, *two and fifty*, *fifty-seventh*, ***hundred***, *157th*, *400th*, *thousand*, ***million***, *80 million*
	fra	***10***, *dizaine*, *dizaines*, *la puissance 10*, *quinzième*, *17e*, ***50***, ***50ème***, ***100***, ***centième***, *cent cinquante*, *400e*, ***mille***, ***million***, ***millions***, *36 millionième*
	jpn	*100*
	nob	*femtende*, *50*, *to og femti*, *57*, ***100***, ***hundrede***, *tusen*, *million*
	swe	***hundrade***, *sjuttifjärde*, ***miljonte***, ***miljoner***
Hyperbolic quasi-numeral expressions	dan	*nogen og tyvende*
	eng	*bajillion*, ***gazillion***, ***gazillionth***, ***umpteen***, ***umpteenth***
	fra	*bajillion*, ***énième***
	nob	***ente***, ***n-te***, ***nte***, ***ørten***, ***ørtende***, ***ørten millioner***
	swe	***femtioelva***, ***femtioelfte***, *femtioelva miljoner*, *hundrafemtielva*, *trehundrafemtioelfte*, *skviljontals*, *ziljon*
Non-numeral quantifiers	dan	*utallige*, *et bredt spektrum*, *et hav*, *-vis*
	fra	*chaque*, *combien*, *des lustres*, *des tas*, *je ne sais combien*, *multiples*, *plein*, *tous*
	jpn	*nan-CL-ka*, *nan-CL-mo*
	nob	*utallige*
	swe	*ett otal*, *otaliga*, *en massa*, *flera*, *många*, *några*, *-tal*

4.2 RQ1: Distribution of hyperbolic numerals and quasi-numerals

The cross-linguistic distribution of hyperbolic numerals and quasi-numeral expressions identified in this study is mapped in Figure 2.

Expressions matching the criteria described in section 3.4 were identified in 46 of the 50 analyzed languages. Both hyperbolic numerals and quasi-numeral expressions were identified in a majority of languages in the sample, and over two-thirds of the languages in which either type of expression was found. In 15 languages, primarily in South and Southeast Asia, only hyperbolic numerals were identified. Finally, in Bengali, Catalan, Hindi and Georgian, no expressions of either type could be extracted from the parallel data.

Notably, there is no language in the sample in which only hyperbolic quasi-numeral expressions (and no hyperbolic numerals) were found. In all languages where expressions matching the criteria described in section 3.4 were identified, these expressions include at least one numeral expression with an exact numerical value. Based on this observation, the following preliminary implicational universal is proposed:

(10) The Universal of Hyperbolic Quantifier Types
If a language uses quasi-numeral expressions for quantification in hyperbolic contexts, at least one numeral expression with a numerical value is also conventionalized in hyperbolic contexts.

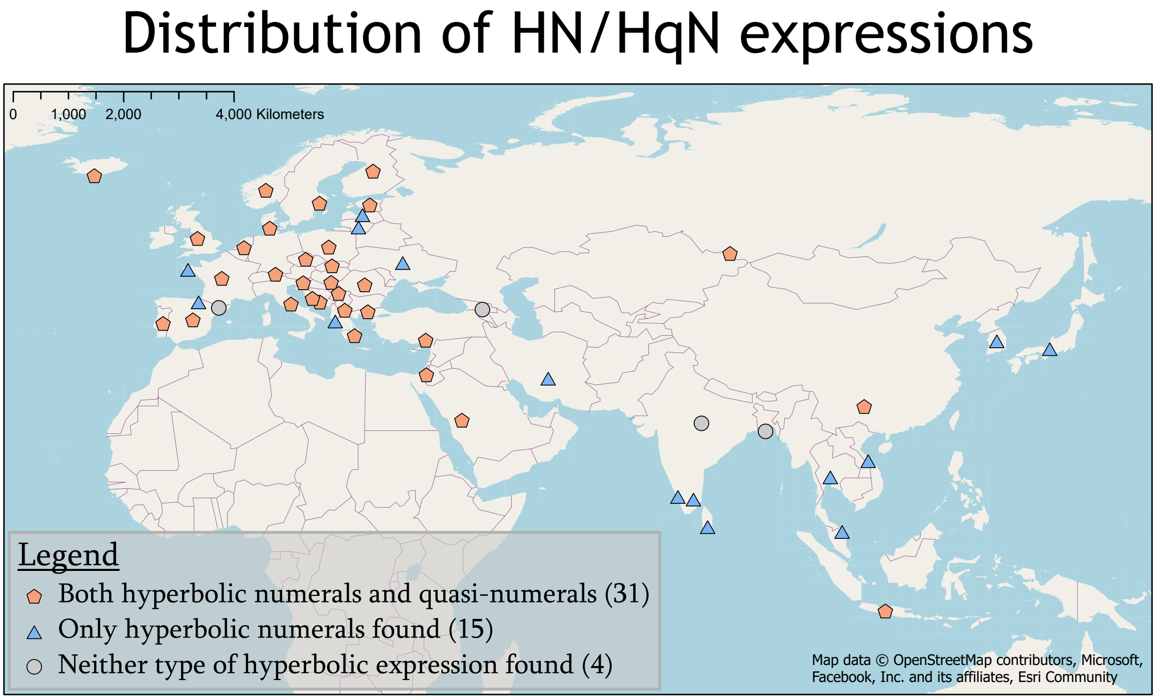

Distribution of HN/HqN expressions
0
1,000
2,000
4,000 Kilometers
Legend
Both hyperbolic numerals and quasi-numerals (31)
Only hyperbolic numerals found (15)
Neither type of hyperbolic expression found (4)
Map data © OpenStreetMap contributors, Microsoft, Facebook, Inc. and its affiliates, Esri Community

Figure 2: Cross-linguistic distribution of hyperbolic numerals and quasi-numerals

4.3 RQ2: Numerical value of hyperbolic numerals

The values of the numeral expressions identified in hyperbolic contexts in this study, along with the languages in which they were identified, are listed in Table 7 (for round numerals, as defined in section 3.4.2) and 8 (for all other numerals).

Among round hyperbolic numerals, considerable variation was found in both value and cross-linguistic prevalence, as can be seen in Table 7.

The most frequently observed values, both occurring in 37 languages in the sample, were 100 and 10^6. Other powers of 10 also occurred, to varying degrees of prevalence – in total, 12 of the 16 different values observed in this category were powers of 10. With only a few exceptions, values above 10^3 patterned as 10^{3n} with decreasing cross-linguistic prevalence as n increased (aside from a local maximum at 10^{18}). Potential explanations for this pattern are discussed further in section 5.2.

Of the observed round hyperbolic numeral values other than powers of 10, 50 was the most cross-linguistically prevalent, occuring in hyperbolic contexts in 11 languages in the sample. The remaining values only occurred in single languages, with the exception of 400. However, as hyperbolic uses of expressions denoting 400 were only found in a single shared context in three of the four languages in which this value was observed (Danish, English and French), this finding may be attributable to source language inference rather than genuine conventionalisation of 400 as a hyperbolic numeral. 20, notably, only occurred as a hyperbolic numeral value in Breton, which has a partly vigesimal numeral system.

The distribution of identified values outside of the roundness criteria defined in 3.4.2 differs from the round values in Table 7 in a number of ways, as can be seen in Table 8. Although a greater variety of non-round hyperbolic numerals were found, most of the identified values occurred only in singular languages. The most cross-linguistically prevalent value was 12, occurring in 4 languages, followed by 15 and 17 which occurred in 3 languages each.

The majority of the observed values are fairly evenly distributed across the range 11–198,

Table 7: Observed hyperbolic numeral values (category 1: round numerals)

Value	Number of lanugages	Languages (ISO 639-3)
10	13	arb, bul, bos, ell, eng, fin, fra, hrv, isl, nld, por, spa, srp
20	1	bre
40	1	tur
50	11	ces, dan, eng, fra, hun, kor, nob, pol, rus, srp, tha, tur
100	37	als, arb, bul, bos, ces, cmn dan, deu, ell, eng, est, eus, fin, fra, heb, hrv, hun, ind, isl, ita, jpn, kat, lit, msa, nld, nor, pes, pol, por, rus, slv, spa, srp, swe, tur, ukr, vie
400	4	ces, dan, eng, fra
10^3	29	arb, bul, ces, cmn, dan, deu, ell, eng, eus, fin, fra, heb, hrv, hun, ind, ita, lit, mal, mkd, nld, nob, pes, pol, por, rus, spa, srp, swe, tur
10^4	3	dan, ell, tha
10^6	37	als, arb, bul, bos, cat, ces, dan, deu, ell, eng, est, eus, fas, fin, fra, geb, hrv, hun, ind, ita, lav, mkd, nld, nor, pol, por, rus, sin, slk, slv, spa, srp, swe, tha, tur, ukr, vie
10^8	1	cmn
10^9	23	bul, ces, dan, deu, ell, eng, fin, fra, heb, hrv, hun, ind, ita, mkd, nld, pol, ron, rus, slv, srp, swe, tur, vie
10^{11}	1	tha
10^{12}	22	arb, bul, ces, dan, deu, ell, est, fra, heb, hrv, hun, ita, nld, nob, pol, por, ron, rus, spa, srp, swe, tur
10^{15}	4	ces, hrv, ron, swe
10^{18}	14	ces, dan, deu, fra, hrv, hun, isl, ita, nld, nob, pol, por, spa, swe
10^{21}	5	dan, ita, lit, nob, pol

with higher numerals either being derived from other hyperbolic numerals (such as *rɔ́ɔi-bpɛ̀ɛt-pan-gâao* ‘108009’ in Thai or *trente-six millionième* ‘36 millionth’ in French) or constructed using monomorphemic numerals denoting higher powers of 10 (such as *80 million*).

Notably, the observed expressions are the least morphologically complex representations of their respective values. For instance, while the value $8 \cdot 10^7$ is represented in English as *80 million* ’$80 \cdot 10^6$ using the monomorphemic numeral *million* ‘10^6’, no equivalent monomorphemic numeral exists in Japanese, which represents $8 \cdot 10^6$ as *happyaku-man* ‘$800 \cdot 10^4$’ using the monomorphemic numeral *man* ‘10^4’.

Partly compounded by the lack of values occurring in hyperbolic contexts in more than one language, no clear patterns of occurrence were found in the values of lower-value non-round hyperbolic numerals; although 7 appeared in several hyperbolic numeral expressions among Germanic languages (*fifty-seventh* ‘57th’ in English, *sjuttifjärde* ‘74th’ in Swedish and *hundredesyttende* ‘117th’ in Danish), non-round hyperbolic numerals not containing 7 were also identified in English and Danish. Among the higher-value non-round hyperbolic numerals, two patterns involving the number 8 emerged: in both Mandarin and Thai, hyperbolic numeral expressions are formed with numerals denoting 108 (which also occurred hyperbolically on its own in Thai), and the numeral 8 appeared in the English and Japanese hyperbolic numeral constructions discussed above.

Table 8: Observed hyperbolic numeral values (category 2: non-round numerals)

Value	Number of lanugages	Languages (ISO 639-3)	Value	Number of lanugages	Languages (ISO 639-3)
11	1	eus	108	1	tha
12	4	ben, dan, eng, tha	117	1	dan
15	3	eng, fra, nob	150	1	fra
17	3	dan, eng, fra	157	1	eng
36	1	fra	198	1	deu
37	1	tam	600	1	pes
46	1	eng	$108 \cdot 10^3$	1	cmn
52	2	dan, nob	108009	1	tha
57	2	eng, nob	$800 \cdot 10^4$	1	jpn
74	1	swe	$36 \cdot 10^6$	1	fra
			$80 \cdot 10^6$	1	eng

4.4 RQ3: Construction of hyperbolic quasi-numeral expressions

A total of 277 unique candidate quasi-numeral expressions were collected from the extraction and elicitation procedures, of which 208 occurred at least once in their respective monolingual web corpora (as listed in Appendix A). In total, hyperbolic quasi-numeral expressions were identified (according to the criteria in section 3.4.3) in 31 languages.

Among these, several lexically and morphologically distinguished clusters of expressions were identified. Tables 9 through 12 list the lemmas of all identified expressions within each cluster, and the distinguishing properties of each cluster are described below. To simplify further analysis, each cluster is also assigned a category label and index.

Category 3 – Semi-numeral expressions

A majority of the identified quasi-numeral expressions consist partly, but not entirely, of morphemes and pseudo-morphemes from the respective language's numeral series. Lemmatized forms of the identified expressions in this category are listed in Table 9.

Among these semi-numeral expressions, two common types of construction patterns were identified, distinguished by the morphological properties of the numeral constituent. In the most commonly observed type of pattern (*Numeral-pattern quasi-numerals*, category 3a), expressions resemble numerals either through morphemes used in the construction of complex numerals or through pseudo-morphemes derived from metanalysis of numerals (such as the *-illion* in English *zillion*, patterning after *million* '10^6' and *billion* '10^9'). While some expressions of this type, such as Finnish *ziljoona*, pattern after the native forms of higher powers of 10 (such as *miljoona* '10^6' and *biljoona* '10^{12}'), others do not conform to this pattern – compare, for instance, Czech *zillion* with *milión* '10^6' and *bilión* '10^{12}'. In a number of languages, including Czech, both conforming and non-conforming expressions of this type occur.

Various initial consonants and clusters were also observed in these constructions. *z-* and *j-* (as in English *zillion* and *jillion*) were particularly common, but other language-specific clusters like *ts-* (in Finnish *tsiljoona*) and *skr-* (in Icelandic *skrilljón*) also occurred. Intensifying prefixes such as *ba-* and *ga-* in *bajillion* and *gazillion* also occured cross-linguistically to some extent, although their use in the analyzed expressions seems to be limited to *z-* and *j*-initial base expressions.

Table 9: Observed hyperbolic quasi-numeral expressions in categories 3a (numeral-pattern quasi-numerals) and 3b (compound quasi-numerals), by language

Category	Language	Lemmas of expressions
Numeral-pattern quasi-numerals (3a)	arb	*dishaliuwn, jilyun, zilyun, zilywn*
	bos	*bazillion, zilion*
	bul	*zilion*
	ces	*bambilion, bazilión, bžilion, gazilión, zilión, zillion*
	dan	*fantasillion, gajillion*
	deu	*bazillion, drölf, drölfzig, zig, zillion*
	eng	*bajillion, gajillion, gazillion, jillion, umpteen, zillion*
	est	*ziljon*
	fin	*tsiljoona, ziljoona*
	fra	*bajillion, bazillion, gajillion, gazillion, jillion, squillion, zillion*
	heb	*bzilion, gzilion, zilion*
	hrv	*bazilion, gajillion, gazilijon, gazilijun, gazilion, zilijun, zilion, zillion*
	hun	*csilliárd, csillió, zillió, zsillió*
	isl	*skrilljón*
	ita	*fantastiliardo, fantastilione, zilione*
	mkd	*zilion*
	nld	*gazillion, kazillion, tig, ziljoen*
	nob	*fantasilion, gazillion, ørten, zillion*
	pol	*bazylion, gazylion, kazylion, pierdylion, zillion, zylion*
	por	*bajillion, bazilhão, gazilhão, gazillion, milhentos , porrilhão, zilhão, ziliāo, zilion*
	rus	*bazillion, dzhillion, gazillion, zillion*
	slk	*bazilión, zilióny*
	slv	*gazillion, zilijon, ziljon*
	spa	*bazillion, gajillion, gazillion, jillion, zillon*
	srp	*džilion, gajillion, gazilion, zilijun, zilion, zillion*
	swe	*baziljon, ziljon, zillion*
	tur	*gazilyon, zibilyon, zilyon*
Compound quasi-numerals (3b)	deu	*drölfhundert, drölftausend, drölfzigtausend, zig Milliarden, zig Millionen, zigmilliarden, zigmillionen, zigtausend*
	est	*mustmiljon*
	heb	*malantalafim*
	hun	*kismillió, sokmillió*
	nob	*ørten millioner*
	por	*trocentas*
	ron	*jdemii*
	spa	*chorrocientas, tropecientas, tropecientos millones*
	srp	*mali milion*

Some expressions of this type are constructed using an initial non-numerical morpheme rather than single consonants or clusters. Examples of this type include Italian *fantastilione* (analyzed by Chrisomalis (2016, p. 7) as *fantastico* 'fantastic' + *-ilione*) and *fantastiliardo*, as well as the corresponding expressions *fantasillion* in Danish and *fantasilion* in Norwegian. Vulgar terms were also commonly identified in the construction of these expressions in several languages, such as Polish *pierdylion* and Portuguese *porillhão*, which are likely derived from *pierdolić* 'fuck' and *porra* 'penis' respectively.

Expressions in this cluster do not exclusively use variants of the *-illion* pseudo-morpheme – in languages with numeral systems using the long scale, some expressions are constructed using variants of *-illiard*, patterning after 10^9, 10^{15} and so on. A number of constructions also use the *-teen* and *-ty* suffixes used to form certain numerals between 11 and 99, including English *umpteen*, Norwegian *ørten* and German *drölfzig* – a construction using another hyperbolic quasi-numeral, *drölf*, which seems to pattern after *zwölf* '12'. The *-ty* suffix is also used as a hyperbolic quasi-numeral on its own in German (*zig*, compare with *zwanzig* '20') and Dutch (*tig*, compare with *twintig* '20').

The second cluster (*Compound quasi-numerals*, category 3b) consists of expressions which use full numeral expressions denoting exact values, typically powers of 10, rather than complex numeral-constructing morphemes or pseudo-morphemes such as *-illion*. These numerals are used to form compounds with other non-numerical morphemes, as in Estonian *mustmiljon* 'black million' or Hungarian *kismillió* 'small million'. In some cases, the non-numerical morphemes concerned are other hyperbolic quasi-numerals, such as in German *zigmillionen*, Spanish *tropecientos millones* and Swedish *femtioelva miljoner*.

While numerals denoting '10^6' occurred most frequently in this type of construction, as seen in the examples above, these compound quasi-numeral expressions also commonly involved numerals denoting 100 (such as Portuguese *trocentas*) or 10^3 (such as Hebrew *malantalafim*). An especially interesting construction of this type is German *drölfzigtausend*, in which all three constituents can also be used as hyperbolic quantifiers on their own or in any combination.

A structural commonality across all semi-numeral expressions is the order of their numeral and non-numeral constituents. In all identified expressions, the numeral constituent follows the non-numeral constituent. This pattern also holds for compound quasi-numerals consisting of more than one numeral constituent, such as Spanish *tropecientos millones* where both numeral constituents follow the non-numeral.

Category 4 – Pseudo-numeral expressions

A number of identified expressions are composed entirely of numeral constituents, resembling complex numerals, but violate conventional complex numeral syntax. Russian *stopjatsot* is a compound of *sto* '100' and *pjatsot* '500', and Swedish *femtioelva* 'fifty-eleven' and French *quarante-douze* 'forty-twelve' follow the same atypical pattern of construction as English *fortyleven. femtioelva* also occurred in complex (quasi-)numeral constructions with other numerals, as in *trehundrafemtioelva* 'three hundred and fifty-eleven'.

Table 10: Observed hyperbolic quasi-numeral expressions in category 4 (pseudo-numerals), by language

Category	Language	Lemmas of expressions
Pseudo-numerals (4)	fra	*quarante-douze*
	hun	*millió-billió*
	ita	*millemila*
	rus	*stopjat'sot*
	swe	*femti-elva, femtielva, femtioelva, hundrafemtielva, trehundrafemtioelva*

Category 5 – Algebraic variable expressions

Another set of identified quasi-numeral expressions, presented in Table 11, consists of numeral derivations using common algebraic variable denotations in place of numerals. These expressions are predominantly ordinal, and less commonly frequentative (such as Mandarin *N cì* and German *x-mal.*

Table 11: Observed hyperbolic quasi-numeral expressions in category 5 (algebraic variable expressions), by language

Category	Language	Lemmas of expressions
Algebraic variable expressions (5)	bul	*en-ti, enti*
	ces	*ixté, xté*
	cmn	*N cì*
	deu	*x-mal, x-ten*
	eng	*nth*
	fra	*énième*
	hun	*ikszedik*
	ita	*ennesima*
	nob	*ente, n-te, nte*
	pol	*enty, n-ty*
	por	*enésima*
	spa	*enésima*
	srp	*enti*

The main contrast between these expressions is in the variable denotation used. All identified variable expressions use either *n*, such as Serbian *enti*, or *x*, such as Czech *xté*, with *n* occurring in a majority of expressions. No language in the sample had variable expressions using both *n* and *x*.

Orthographic conventions for these expressions varied: in some cases, such as German *x-ten*, the algebraic variable is transparent and often delimited by a word boundary or hyphen, while the boundary is blurred in other expressions like Portuguese *enésima.*

Category 6 – *many-th* expressions

Finally, a number of identified quasi-numeral expressions consist of a non-numerical quantifier and an otherwise numeral-exclusive derivational affix, as in Hungarian *sokadik* 'many-th'.

Expressions of this type are listed in Table 12.

These expressions are exclusively ordinal, and seem to be exclusively used in hyperbolic contexts. There is some variation in which non-numerical quantifier is used, but in most cases it is equivalent to *many*.

Table 12: Observed hyperbolic quasi-numeral expressions in category 6 (*many-th* expressions), by language

Category	Language	Lemmas of expressions
***many-th* expressions (6)**	ces	*bůhvíkolikátý, několikáté*
	ell	*pollostí*
	heb	*hmy-ywdʿ-kmh*
	hun	*sokadik*
	ind	*kesekian*
	nld	*zoveelste*

4.5 RQ4: Distribution of types of HN/HqN expressions

Table 13 shows the number of languages per genus in which hyperbolic numerals or quasi-numeral expression corresponding to each category presented in 4.3 and 4.4 could be identified. Maps visualizing the areal distributions of hyperbolic numerals and quasi-numerals within each category are presented in figures 3 through 7.

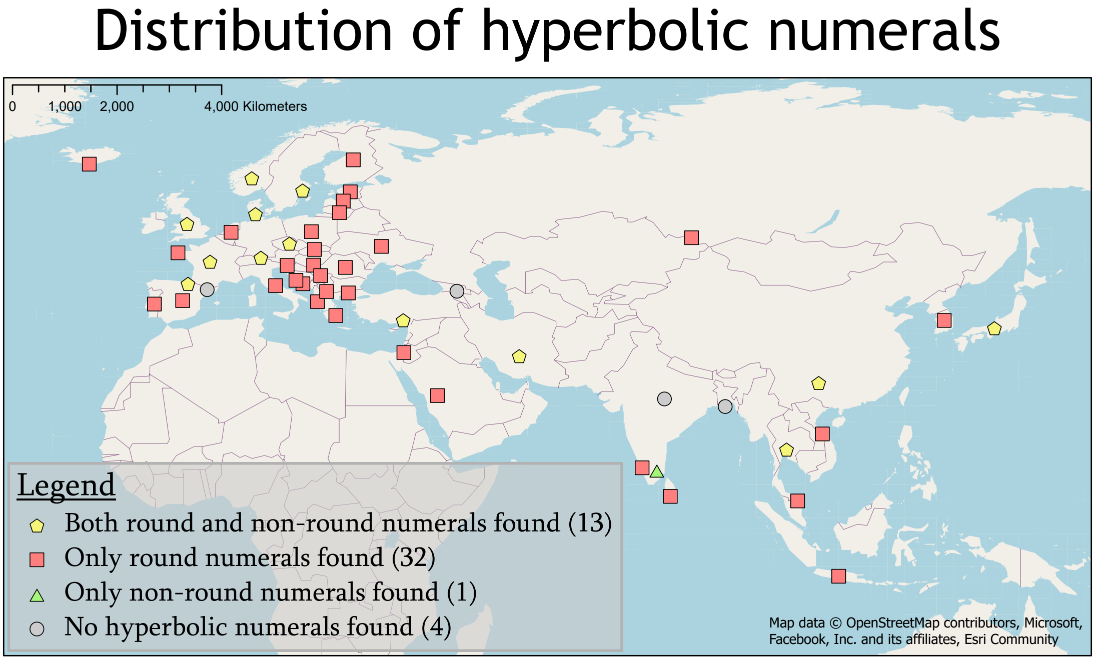

Distribution of hyperbolic numerals
0
1,000
2,000
4,000 Kilometers
Legend
⬠ Both round and non-round numerals found (13)
🟥 Only round numerals found (32)
△ Only non-round numerals found (1)
○ No hyperbolic numerals found (4)
Map data © OpenStreetMap contributors, Microsoft, Facebook, Inc. and its affiliates, Esri Community

Figure 3: Cross-linguistic distribution of hyperbolic numerals in categories 1 and 2

The distribution of round and non-round hyperbolic numerals is visualized in Figure 3. While round numerals were identified in almost all languages in the sample, non-round numerals were more unevenly distributed. Germanic languages were particularly well-represented, with non-round numerals identified in 5 out of 7 languages in the sample.

With only a single exception (Tamil), all languages where non-round hyperbolic numerals were identified also had round numerals occur in hyperbolic contexts.

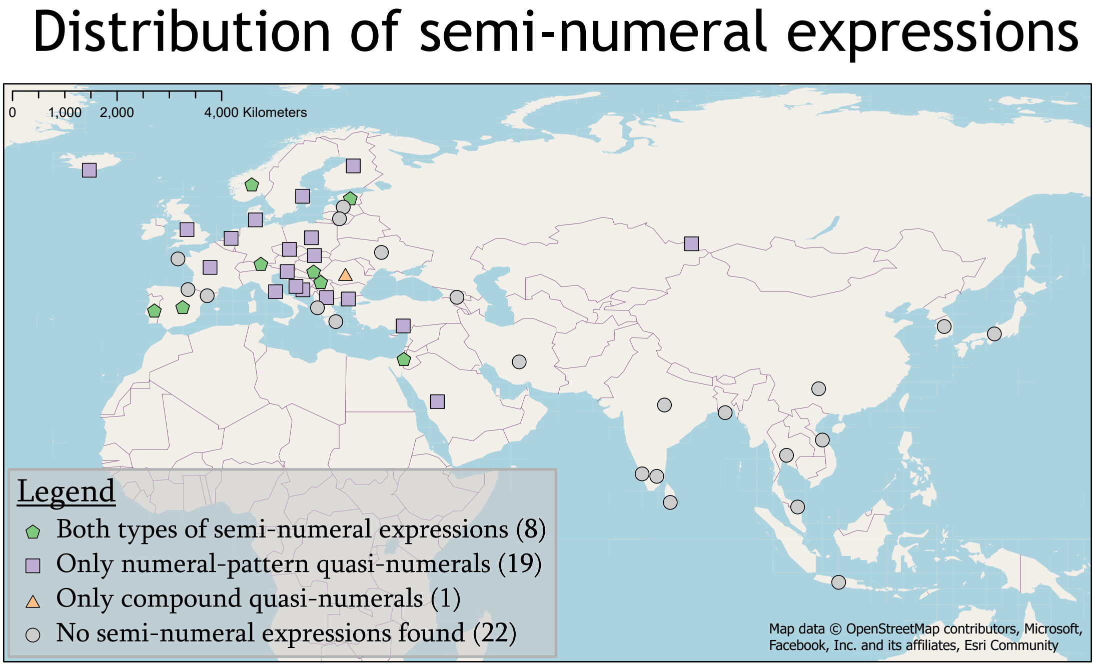

Distribution of semi-numeral expressions
0 1,000 2,000 4,000 Kilometers
Legend
⬠ Both types of semi-numeral expressions (8)
□ Only numeral-pattern quasi-numerals (19)
△ Only compound quasi-numerals (1)
○ No semi-numeral expressions found (22)
Map data © OpenStreetMap contributors, Microsoft, Facebook, Inc. and its affiliates, Esri Community

Figure 4: Cross-linguistic distribution of hyperbolic quasi-numeral expressions in categories 3a (numeral-pattern quasi-numerals) and 3b (compound quasi-numerals)

A clear areal trend is visible in the distribution of semi-numeral expressions, as shown in Figure 4. Semi-numerals were predominantly found in European languages, and were particularly prevalent among Germanic, Slavic and Romance languages in the sample. Prevalence across these three Indo-European genera differed between the two types of semi-numerals – while numeral-pattern quasi-numerals could be identified in nearly all languages belonging to these genera, compound quasi-numerals were proportionally more well-represented among Romance languages than among Germanic or Slavic languages.

Similarly to the observed relationship between round and non-round hyperbolic numerals, all languages but one (Romanian) in which compound quasi-numerals were identified also had numeral-pattern quasi-numerals. These near-universal patterns are analyzed further in 5.4.

As with semi-numeral expressions, pseudo-numeral expressions like *quarante-douze* were primarily found in European languages, as can be seen in Figure 5. Since these expressions were only identified in 5 of the 50 languages in the sample, no consistent areal or genealogical tendencies aside from this could be observed.

The distribution of algebraic variable expressions, displayed in Figure 6, is also skewed towards Europe, with the notable exception of Mandarin. Prevalence was comparable among the three most well-represented genera, Germanic, Romance and Slavic, with algebraic variable expressions identified in a slightly higher proportion of Romance languages (66%) than Germanic (43%) and Slavic (36%) languages.

Finally, Figure 7 shows the distribution of *many-th* expressions. These expressions were identified across several genera and across the entire macroarea, although once again with a skew towards Indo-European languages spoken primarily in Europe. As with pseudo-numeral expressions, too few expressions of this type were found for any more fine-grained and consistent areal or genealogical patterns to be observable. Unlike the other categories, however, *many-th* expressions were not identified in more than one language in any single genus.

Table 13: Cross-genus distribution of hyperbolic numerals and quasi-numeral expressions by category

Language		Number of	HN/HqN by category						
family	Genus	languages	1	2	3a	3b	4	5	6
Afro-Asiatic	Semitic	2	2		2	1			1
Austroasiatic	Viet-Muong	1	1						
Austronesian	Malayo-Sumbawan	2	2						1
Dravidian	Southern Dravidian	2	1	1					
Basque (isolate)	Basque	1	1	1					
Indo-European	Albanian	1	1						
—"—	Baltic	2	2						
—"—	Celtic	1	1						
—"—	Germanic	7	7	5	7	2	1	3	1
—"—	Greek	1	1						1
—"—	Indic	3	1						
—"—	Iranian	1	1	1					
—"—	Romance	6	5	1	4	3	2	4	
—"—	Slavic	11	11	1	10	1	1	4	1
Japonic	Japanese	1	1	1					
Kartvelian	Kartvelian	1							
Koreanic	Korean	1	1						
Sino-Tibetan	Chinese	1	1	1				1	
Tai-Kadai	Kam-Tai	1	1	1					
Turkic	Turkic	1	1	1	1				
Uralic	Finnic	2	2		2	1			
—"—	Ugric	1	1		1	1	1	1	1
	Sum	50	45	14	27	9	5	13	6

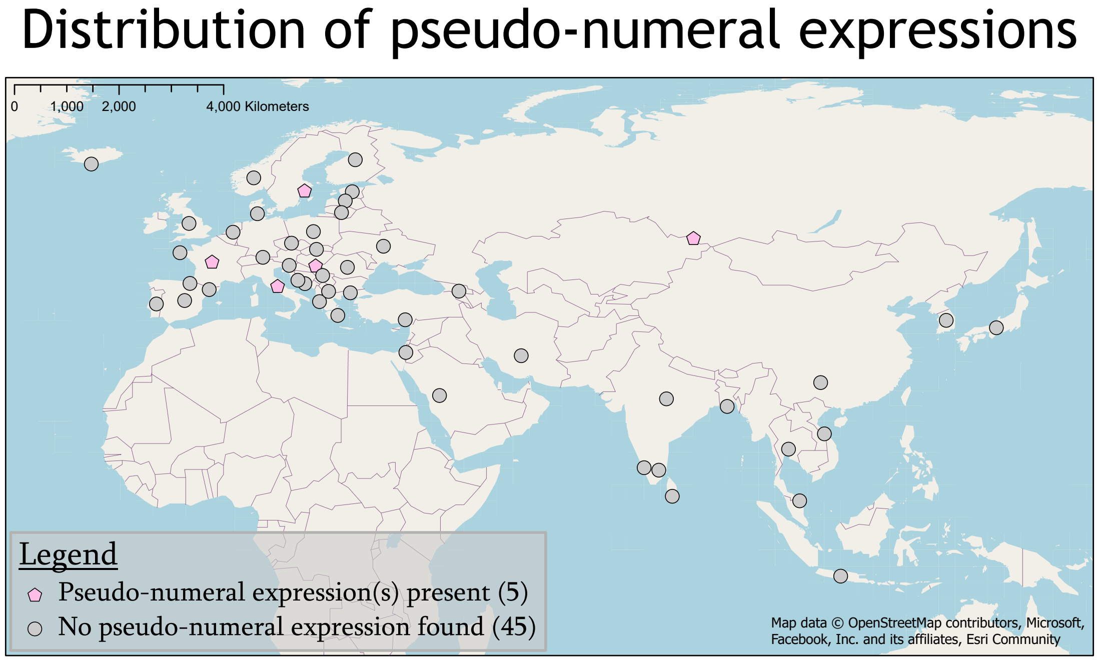

Distribution of pseudo-numeral expressions
0 1,000 2,000 4,000 Kilometers
Legend
⬠ Pseudo-numeral expression(s) present (5)
○ No pseudo-numeral expression found (45)
Map data © OpenStreetMap contributors, Microsoft, Facebook, Inc. and its affiliates, Esri Community

Figure 5: Cross-linguistic distribution of hyperbolic quasi-numeral expressions in category 4 (pseudo-numeral expressions)

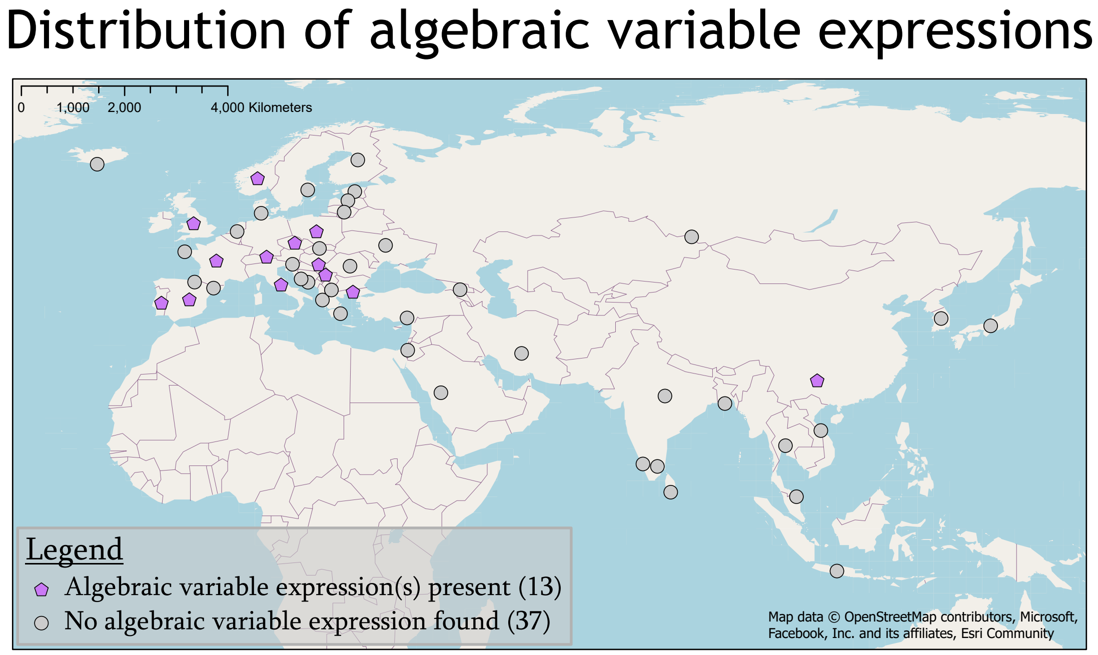

Distribution of algebraic variable expressions
0 1,000 2,000 4,000 Kilometers
Legend
⬠ Algebraic variable expression(s) present (13)
○ No algebraic variable expression found (37)
Map data © OpenStreetMap contributors, Microsoft, Facebook, Inc. and its affiliates, Esri Community

Figure 6: Cross-linguistic distribution of hyperbolic quasi-numeral expressions in category 5 (algebraic variable expressions)

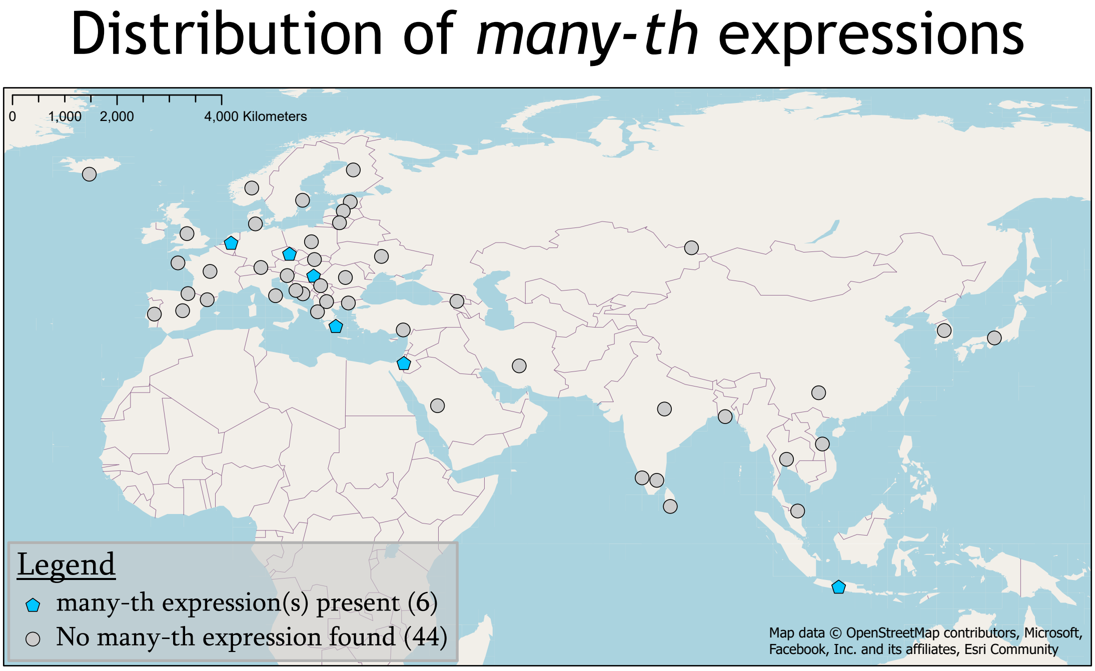

Distribution of *many-th* expressions
0
1,000
2,000
4,000 Kilometers
Legend
many-th expression(s) present (6)
No many-th expression found (44)
Map data © OpenStreetMap contributors, Microsoft, Facebook, Inc. and its affiliates, Esri Community

Figure 7: Cross-linguistic distribution of hyperbolic quasi-numeral expressions in category 6 (*many-th* expressions)

4.6 RQ5: Patterns in usage of HN/HqN expressions

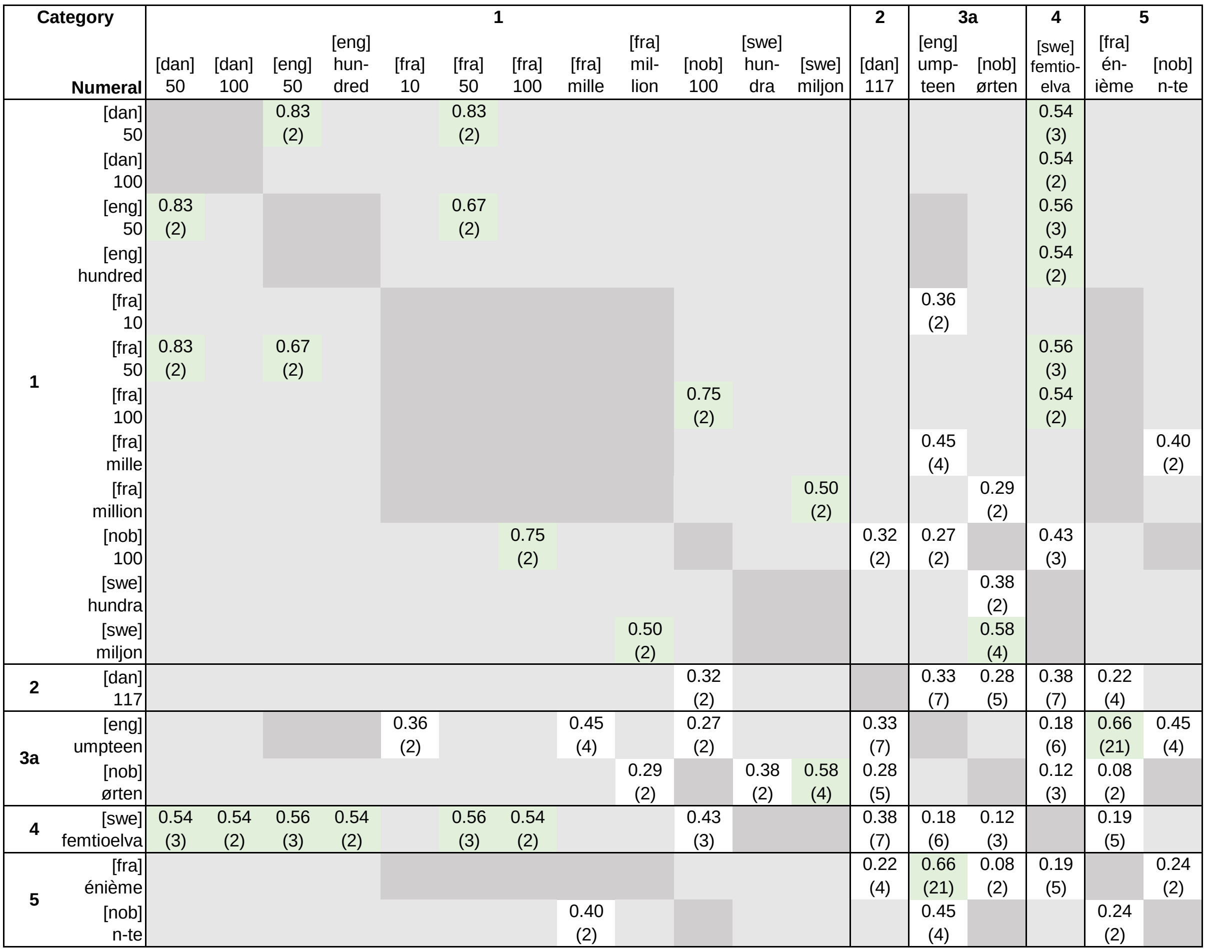
Figure 8 shows a matrix of pairwise translation probabilities (as defined in 3.4.5) of hyperbolic numeral and quasi-numeral expressions which occurred in the pilot data. Each non-blank cell represents a pair of expressions which were translational equivalents of each other in at least one pilot context. Two values are listed in each cell: the translation probability (above) and the absolute co-occurrence frequency of the translation pair in the pilot data (below, in parentheses). Pairs with a translation probability of 0.5 or higher are highlighted in green, and pairs of expressions in the same language (which logically cannot co-occur with each other in cross-language parallel data) are marked with a darker gray shade.

The four pairs of expressions with the highest translation probabilities are all hyperbolic numerals denoting identical values. Following these, the next highest-probability translation pair is English *umpteen* and French *énième*. This pair also had a considerably higher absolute frequency of co-occurrence in the pilot data than any other pair, occurring together in 21 of the 99 pilot contexts – both expressions were also among the most frequent in the pilot data on their own.

The remaining pairs with translation frequencies above 0.50 all include one of two quasi-numerals: Swedish *femtioelva* and Norwegian *ørten*. These expressions have clearly differing patterns of co-occurrence – while *femtioelva* most frequently occurred in the same contexts as hyperbolic numerals denoting the values 50 and 100, *ørten* co-occurred more often with numerals denoting 10^6.

Danish *hundredesytten* '117' was the only non-round hyperbolic numeral identified in the pilot study. Notably, *hundredesytten* co-occurred to a significantly greater extent with hyperbolic quasi-numerals than with other (round) hyperbolic numerals.

Finally, the variable expressions in category 5, French *énième* and Norwegian *n-te*, co-occurred most consistently with *umpteen* and with each other, although *n-te* also has an unexpectedly high (albeit potentially spurious) translation probability with French *mille* '1000'.



Category	Numeral	1												2	3a		4	5	
		[dan] 50	[dan] 100	[eng] 50	[eng] hundred	[fra] 10	[fra] 50	[fra] 100	[fra] mille	[fra] million	[nob] 100	[swe] hundra	[swe] miljon	[dan] 117	[eng] umpteen	[nob] ørten	[swe] femtioelva	[fra] énième	[nob] n-te
1	[dan] 50			0.83 (2)			0.83 (2)										0.54 (3)		
	[dan] 100																0.54 (2)		
	[eng] 50	0.83 (2)					0.67 (2)										0.56 (3)		
	[eng] hundred																0.54 (2)		
	[fra] 10														0.36 (2)				
	[fra] 50	0.83 (2)		0.67 (2)													0.56 (3)		
	[fra] 100										0.75 (2)						0.54 (2)		
	[fra] mille														0.45 (4)				0.40 (2)
	[fra] million												0.50 (2)			0.29 (2)			
	[nob] 100							0.75 (2)						0.32 (2)	0.27 (2)		0.43 (3)		
	[swe] hundra															0.38 (2)			
	[swe] miljon									0.50 (2)						0.58 (4)			
2	[dan] 117										0.32 (2)				0.33 (7)	0.28 (5)	0.38 (7)	0.22 (4)	
3a	[eng] umpteen					0.36 (2)			0.45 (4)		0.27 (2)			0.33 (7)			0.18 (6)	0.66 (21)	0.45 (4)
	[nob] ørten									0.29 (2)		0.38 (2)	0.58 (4)	0.28 (5)			0.12 (3)	0.08 (2)	
4	[swe] femtioelva	0.54 (3)	0.54 (2)	0.56 (3)	0.54 (2)		0.56 (3)	0.54 (2)			0.43 (3)			0.38 (7)	0.18 (6)	0.12 (3)		0.19 (5)	
5	[fra] énième													0.22 (4)	0.66 (21)	0.08 (2)	0.19 (5)		0.24 (2)
	[nob] n-te								0.40 (2)						0.45 (4)			0.24 (2)	

Figure 8: Translation probability matrix for HN/HqN expression pairs in the pilot data (all expressions lemmatized)

5 Discussion

In this chapter, the results presented in chapter 4 are discussed and related to prior work, and issues relating to methodology are explored. In sections 5.1 through 5.5, results relating to each research question are discussed. Section 5.6 contains a discussion of the data types, data sources, language sample and methods used in this study. Finally, potential directions for future work on the cross-linguistic properties of hyperbolic numerals and quasi-numeral expressions are explored in section 5.7.

5.1 RQ1: Distribution of hyperbolic numerals and quasi-numerals

The distribution of hyperbolic numerals and quasi-numeral expressions, as charted in Figure 2, shows that expressions of these types were identified in 46 languages, making up over 90 percent of the language sample. While hyperbolic numerals were the most cross-linguistically prevalent of the two types, quasi-numeral expressions also occurred in a majority of languages in the sample. This evidence clearly contradicts the notion of quasi-numerals being a cross-linguistically rare phenomenon, as is presupposed by Chrisomalis (2016, p. 6).

Some initial areal tendencies can be observed in Figure 2: languages in which both hyperbolic numerals and quasi-numerals were identified are proportionally more common in Europe than in Asia. The areal distributions of the various categories of these expressions are analyzed in further detail in section 5.4. However, a crucial factor with the potential to influence this distribution is the amount of data analyzed for each language. As noted in section 3.2, the amount of *OpenSubtitles* data varied considerably between languages, and all 19 languages in which one or both types of expressions could not be found are among the languages with the least such data available. In Tamil, no *OpenSubtitles* data could be analyzed at all, meaning that only elicited expressions were included in the analysis.

This distribution cannot be fully explained by data variance, of course – for instance, both types of expressions were found in Icelandic, a language with one of the smallest *OpenSubtitles* subcorpora. Nevertheless, it is important to note that even languages where such expressions were not found within this study may still have a number of conventionalized hyperbolic numerals and quasi-numeral expressions. As an example of this, the (fairly uncommon) Korean expression *golbaek* fits the lexical and morphological criteria for hyperbolic quasi-numeral expressions (National Institute of Korean Language 2022) but did not occur in any of the analyzed *OpenSubtitles* parallel texts and as such could not be included in the analysis. Had this expression co-occurred with one of the seed expressions in the parallel data, Korean would have been classified differently in Figure 2.

This context is also important to keep in mind when considering the initial implicational universal, proposed in (10) and restated here:

(11) The Universal of Hyperbolic Quantifier Types
If a language uses quasi-numeral expressions for quantification in hyperbolic contexts, at least one numeral expression with a numerical value is also conventionalized in hyperbolic contexts.

This universal is proposed on relatively weak grounds – a language sample with limited coverage, and varying amounts of data for the languages included – and should as such only be seen as an initial suggestion, to be tried and evaluated against further data in a broader and more representative set of languages.

5.2 RQ2: Numerical value of hyperbolic numerals

Following the extraction and analysis of hyperbolic numeral expressions, two types of hyperbolic numerals with distinct properties were identified:

1. **Round hyperbolic numerals** – generally multiples or powers of the language's numeral base; extensive approximative usage outside of hyperbolic contexts
   Examples: [fin] *miljardi* '10^9'; [hun] *millió* '10^6'

2. **Non-round hyperbolic numerals** – most commonly denote values below 200; no approximative usage outside of hyperbolic contexts
   Examples: [dan] *hundredesytten* '117', [fra] *trente-six* '36'; [tha] *rɔ́ɔi-bpɛ̀ɛt* '108'

Two notable trends were observed among the identified round hyperbolic numerals. First, powers of 10 (particularly 10^2 and 10^6) were cross-linguistically common as hyperbolic numerals. This pattern, including the specific peak in prevalence at 10^2 and 10^6, is entirely aligned with McCarthy and Carter's (2004) quantitative findings for hyperbolic usage in English. The pattern of peaks in cross-linguistic prevalence of every third power of 10 (10^3, 10^6, 10^9, 10^{12}, 10^{15}, 10^{18} and 10^{21}) can be explained by many decimal-system languages having particularly morphologically simple representations of these values. The observed expressions denoting powers of 10 which do not fit this pattern (10^4, 10^8 and 10^{11}) were also denoted by monomorphemic (or, in the case of Thai *s̄æn l̂ān* '10^{11}', dimorphemic) expressions in their respective languages, further supporting a morphological complexity-based explanation.

Secondly, the numeral base of a language's numeral system seemed to make a difference for which numerals are used hyperbolically: Breton, one of the only hybrid vigesimal-decimal languages in the sample, is also the only language in which hyperbolic usage of a numeral denoting 20 was found. This result is in line with the cross-linguistic findings of Krifka 2009, and highlights the necessity of accounting for numeral base when investigating hyperbolic quantification, approximation or other phenomena relating to numeral 'roundness'. As only a limited number of languages with numeral bases other than 10 could be included in the sample of this study, a further exploration of hyperbolic numeral values in languages with bases 20 and 12 (as well as in languages with a restricted numeral system) would evidently be valuable.

Among non-round hyperbolic numerals, no similarly consistent cross-linguistic patterns of value were identified. The numeral 108, which appeared as an independent hyperbolic numeral in Thai and as a constituent in complex hyperbolic numerals in Mandarin and Thai, has particular significance in Buddhism which could explain its particular prevalence. Although the numeral 7 (which occurred as a constituent in several hyperbolic numeral expressions in Germanic languages) also has various cultural and religious connotations, it is unclear whether its use in hyperbolic numeral expressions is culturally motivated.

The most cross-linguistically prevalent hyperbolic numeral in this category, 12, is also particularly significant in some European languages, both culturally and in morphological form (Veselinova 2020). Given that 12 is the only identified non-power-of-10 numeral to use an alternate representation (for example, English *dozen*), and consequently the only non-power-of-10 numeral to appear in plural in the analyzed data, its classification as a non-round numeral is arguable. However, 12 is also distinguished from other hyperbolic numerals in the extent of its non-hyperbolic approximative usage – unlike plural powers of 10, *dozens* was not observed in non-hyperbolic approximative contexts either in this study or in McCarthy and Carter 2004. Evidently, 12 is an exceptional value among both hyperbolic and non-hyperbolic numerals.

A majority of observed non-round hyperbolic numerals denote values in the range 11–198, with most values (including the most cross-linguistically prevalent) found in the low end of this range. In addition, as is noted in 4.3, higher non-round hyperbolic numerals used the least morphologically complex representations possible. This suggests that minimizing morphological complexity remains a cross-linguistically consistent priority even among non-round hyperbolic numerals.

While some of the identified expressions follow the patterns of hyperbolic numeral derivation identified by Lavric (2010), other patterns (such as $N + 1$ and $N{-}1$) are entirely unrepresented in this study. This may be a result of numerals following these pattern being restricted to certain hyperbolic domains not covered by the seed expressions used in this study (as briefly discussed in 2.2.2), or by their hyperbolic uses being relatively infrequent. Regardless, the potential cross-linguistic hyperbolic usage of numerals like 99 and 1001 warrants further investigation.

5.3 RQ3: Construction of hyperbolic quasi-numeral expressions

A number of common patterns of construction were identified for hyperbolic quasi-numerals:

3. **Semi-numeral expressions** – quasi-numeral expressions partly consisting of numerals, bound numeral morphemes or pseudo-morphemes from numeral series

   3a. **Numeral-pattern quasi-numerals** – pattern after existing numerals, commonly higher powers of 10
   Examples: [fin] *tsiljoona*, [ita] *fantastiliardo*, [nob] *ørten*

   3b. **Compound quasi-numerals** – consist of both numerals and non-numerical constituents
   Examples: [deu] *drölftausend*, [est] *mustmiljon*, [heb] *malantalafim*

4. **Pseudo-numeral expressions** – composed entirely of numeral constituents but in violation of conventional numeral syntax
   Examples: [fra] *quarante-douze*, [rus] *stopjat'sot*, [swe] *femtioelva*

5. **Algebraic variable expressions** – quasi-numeral constructions involving an algebraic variable letter in place of a numeral; commonly ordinal or frequentative constructions
   Examples: [ces] *x-krát*, [hun] *ikszedik*, [ita] *ennesima*

6. ***many-th* expressions** – exclusively hyperbolic ordinal derivations of non-numerical quantifiers
   Examples: [hun] *sokadik*, [ind] *kesekian*, [nld] *zoveelste*

The number of quasi-numeral expressions identified in each language varied considerably. While this must be viewed in relation to the highly uneven amounts of data analyzed in each language, variation was found even between languages with comparable *OpenSubtitles2018* subcorpus sizes – for instance, while only 1 quasi-numeral expression was extracted from the Greek subcorpus of 850 million tokens, 9 such expressions were extracted from the 625-million-token Hungarian subcorpus. Variation was also observed in the number of expressions in each cluster identified in each language – having multiple category-3a expressions was considerably more common among the analyzed languages than having multiple category-3b expressions.

The boundary between the two types of semi-numerals is not entirely clear in the cases of expressions like *umpteen*, *zig* and *ørten* – it can be argued that the suffixes *-teen*, *-zig* and *-ten* do represent the specific numerical value 10 in complex numeral constructions, and that these expressions should thus be considered compound quasi-numerals similar to Spanish *tropecientas*. However, unlike *cientas* 'hundreds', which can be used on its own to denote a numerical value, the numerical use of *-teen* is restricted to the formation of complex numerals. Veselinova (2020) also points out that the English *-teen* suffix is used on its own as a root morpheme in words like *teenager*, and the same goes for Norwegian *-ten* which occurs as a root morpheme in the equivalent *tenåring* 'teenager'. In both of these cases, the lexical content of the morpheme is not 10 but rather a range of complex numerals which can be constructed using it. These properties distinguish quasi-numerals like *umpteen*, *zig* and *ørten* from compound quasi-numerals enough to warrant their exclusion from this cluster, although it could be discussed further whether they should be considered as an altogether separate cluster from either compound or numeral-pattern quasi-numerals.

Femtioelva and *quarante-douze* exemplify a particularly interesting construction pattern for pseudo-numerals, which is also followed by the English expression *forty-leven* as described in Chrisomalis 2016, pp. 9–10. These expressions are firmly categorized as quasi-numerals, since they only occur in hyperbolic usage despite theoretically being valid representations of an exact value. They all violate Hurford's (1975, p. 67) Packing Strategy, which could be a way of emphasizing their hyperbolic meaning (rather than their theoretical exact numerical value). Curiously, *femtioelva* is the only member of this category to occur significantly more frequently in web corpora than its standardized counterpart representation of 61, *sextioett*, which speaks to the degree of conventionalisation of this particular expression.

All three of these expressions have similarly functioning but less frequently occurring variants, in which the value of one or more constituent is changed (such as *fifty-eleven* and *seventy-eleven*, as mentioned by Chrisomalis (2016, pp. 9–10)). Across all languages, variance in the first constituent (for instance *trente-douze* 'thirty-twelve') seems to be much more common than in the second (such as *quarante-onze* 'forty-eleven'). Similar principles of variation seem to apply to some non-round hyperbolic numerals as well. As reported by Lavric (2010, p. 136), $46,000$ and $56,000$ seem to occur in hyperbolic contexts in French in addition to $36,000$ (and, as identified in this study, 36 and $36 \cdot 10^6$). These variation schemas in the intersection of hyperbolic numerals and pseudo-numerals warrant closer cross-linguistic examination.

Finally, the amount of intra-language variety in hyperbolic expressions is also worth discussing in relation to prior research. Although the large inventory of hyperbolic quasi-numeral expressions in English has already been well described, a number of other languages such as German and Hungarian also seem to have a particularly diverse register of hyperbolic quasi-numerals, several of which occured with particularly high frequency in monolingual web corpora. These findings suggest that, in contrast to Chrisomalis's assertion that "no other language has such an extensive lexicon of IHN as English, and in no other language is the use of IHN so unmarked and ubiquitous" (Chrisomalis 2016, p. 7), English is only one of several languages across multiple language families in which a wide variety of hyperbolic quasi-numeral expressions is used.

5.4 RQ4: Distribution of types of HN/HqN expressions

The clearest observed areal tendency was found in the distribution of semi-numeral expressions. As Figure 4 shows, semi-numerals were nearly exclusively found in Europe, and numeral-pattern quasi-numerals were identified in an overwhelming majority of languages in Europe. This distribution is likely a result of language contact – particularly with English, as evidenced by many numeral-pattern quasi-numerals in non-English languages using the English *-illion* pseudo-morpheme instead of equivalent pseudo-morphemes reanalyzed from the own-language numeral system. Genealogically shared patterns of construction also occurred, such as numeral-pattern quasi-numerals using bound morphemes equivalent to *-teen* and *-ty* which only appeared in Germanic languages.

The two near-universal implicational tendencies identified across the observed categories are also worth examining in greater detail. Romanian is the only language in which a compound quasi-numeral expression was identified but no numeral-pattern quasi-numeral was found. The compound quasi-numeral concerned, *jdemii*, was also the only quasi-numeral identified in Romanian. Although the morphological analysis of *jdemii* is somewhat unclear, two factors suggest that it clusters with other compound quasi-numerals: *mii* 'thousands' is the numeral term used in construction of complex cardinal numerals above 1999 in Romanian, and *jdemii* is compared to the German compound quasi-numeral *zigtausend* in a Romanian reference grammar (Iliescu and Popovici 2013, p. 198). Thus, the classification of Romanian as an exceptional language in this regard seems to hold, and no categorically valid implicational universal can be proposed regarding the distribution of semi-numeral (category 3) expressions based on the data in this study.

Within the domain of hyperbolic numerals, the only language in which only non-round numerals were found in hyperbolic contexts is Tamil. While this exception prevents the positing of a preliminary universal regarding the distribution of types of hyperbolic numerals based on the present data, it is important to note the exceptional data situation of Tamil – as no suitable contexts were found in the *OpenSubtitles* parallel data, the analysis of Tamil relied entirely on the single elicited hyperbolic expression, which happened to be a non-round numeral.

The consequences of uneven data availability are important to keep in mind when considering other distributions as well, particularly that of non-round hyperbolic numerals – although they could be identified in 5 of 7 Germanic languages, 4 of these were notably part of the pilot study, and a considerable portion of these numerals were collected through elicitation rather than extraction. This suggests that either the extraction method or the *OpenSubtitles* parallel data may not be well suited for identifying expressions of this type. Similarly, algebraic variable expressions (category 5) were identified mainly in languages with large amounts of analyzed data. This tendency, coupled with the low frequency of occurrence of several of these expressions in monolingual web corpora, may indicate that algebraic variable expressions form an especially peripheral or domain-restricted category of hyperbolic quasi-numerals, and are difficult to identify without sufficient parallel data and suitable seed expressions.

5.5 RQ5: Patterns in usage of HN/HqN expressions

The first result which relates to the contexts in which hyperbolic numerals and quasi-numeral expressions occur is the quantitative analysis of noun collocates within the pilot study, as detailed in Table 4. Bearing in mind the limited number of contexts and languages analyzed within the scope of the pilot study, the results nevertheless align relatively well with Lavric's (2010, pp. 131–132) cross-linguistic findings:

> Not very surprisingly, a large majority of [noun collocates] represent measuring units of different types, namely:
>
> - Iteration, Repetition: *times*
> - Probability: *chances*

- Distance(s): *steps, meters, miles*
- Time: *seconds, minutes, hours, days, years*
- Money: *pennies, cents, dollars*

The categories of Iteration, Time and Money were all present (with varying degrees of prominence) in the pilot data, and Lavric's observation that hyperbolic quantification can occur 'virtually with any type of countable noun' (Lavric 2010, p. 132) is consistent with the wide variety of nouns and noun categories identified in the pilot study.

Notably, however, the categories of Probability and Distance(s) did not appear in the pilot data. While this may simply be a result of the relatively low number of contexts analyzed (given that frequentative constructions made up nearly two thirds of pilot contexts), the types of hyperbolic expressions analyzed may also be a relevant factor. In Lavric 2010, the majority of hyperbolic numerals collocating with nouns categorized as Distance(s) are minimizing (in contexts like French *à deux pas d'ici* 'two steps from here'), whereas all of the seed expressions used in the pilot study are typically maximizing. This suggests that a set of contexts assembled using minimizing rather than maximizing hyperbolic expressions would yield a different distribution of nouns and categories.

The second result of relevance to research question 5 is the translation probability analysis of the pilot data, presented in section 4.6 and Figure 8. While a number of frequently coocurring translation pairs were expected, such as pairs of hyperbolic numerals with corresponding values, there was also some unexpected intra-category variation and clustering.

The most consistent inter-category tendency is the pseudo-numeral *femtioelva* co-occurring with hyperbolic numerals denoting 50 and 100, to a greater extent than any other analyzed expression. The numeral-pattern quasi-numeral expressions *umpteen* and *ørten* do not share this tendency, instead patterning most frequently with algebraic variable expressions and with higher-value hyperbolic numerals, respectively. The only potential cluster in which these quasi-numeral expressions occur together is with *hundredesytten* '117', which co-occurred far more often with nearly all quasi-numeral expressions than with other hyperbolic numerals. This pattern suggests, albeit with a highly limited amount of supporting data, that *hundredesytten* behaves more like a prototypical quasi-numeral than *femtioelva* does, and more generally that the functional distinction between non-round hyperbolic numerals and hyperbolic quasi-numerals is not particularly strong or meaningful.

The discrepancy between the expressions *femtioelva* and *ørten* in the numerical value of their commonly co-occurring hyperbolic numerals is also particularly interesting, as it suggests that these expressions (similarly to hyperbolic numerals with numerical values) denote hyperbolic quantities of different magnitudes. Although this comparative property would seem to run counter to the inherently indefinite nature of hyperbolic quantification, it is worth noting that it is already unquestionably a property of hyperbolic numerals with numerical values. For instance, given the two English hyperbolic sentences *I've called you a hundred times* and *I've called you a million times*, the hyperbolic expression with greater numerical value also conveys a greater hyperbolic quantity, despite said quantity remaining indefinite.

This finding is also in line with Chrisomalis's (2016, p. 7) assertion that major IHNs have a 'larger yet still indefinite referent' than minor IHNs. As mentioned briefly in section 2.2.3, Chrisomalis (2016, p. 25) also claims that intensifier prefixes such as *ba-* and *ga-* in English increase the perceived magnitude of a hyperbolic quantity. This property could be investigated further through a similar co-occurrence analysis including a greater variety of quasi-numeral expressions in each included language.

In contrast to the high variance among the other analyzed quasi-numerals, the algebraic variable expressions *énième* and *n-te* had relatively similar co-occurrence tendencies, primarily occurring together and with *umpteen*. A likely contributing factor to this similarity is the ordinal-restricted nature of most variable expressions – since both *énième* and *n-te* can only occur in ordinal contexts, they are naturally more likely to co-occur with each other as well as other expressions which frequently occur as ordinals (such as *umpteen* and to a lesser extent *hundredesytten*).

These tendencies must all be viewed in the context of the pilot data's limitations – a broader cross-linguistic investigation of significantly more contexts and hyperbolic expressions in a wider range of languages would be needed to reliably identify anything more than preliminary cross-linguistic clustering trends. In particular, a broader translation probability study of co-occurrence tendencies in hyperbolic numerals and quasi-numerals should include expressions belonging to all categories identified in this study, including the presently unrepresented categories 3b (compound quasi-numerals) and 6 (*many-th* expressions).

5.6 Method discussion

In this section, some central methodological issues are considered and discussed. A discussion of the types, sources and properties of the data used in this study is provided in 5.6.1, followed by an analysis of the consequences of language sample limitations in 5.6.2, and finally an evaluation of the chosen method in 5.6.

5.6.1 Data

The highly variant amounts of data between languages and language pairs, as shown in Appendix A, most likely skewed the extraction results to some degree. This was manifested most clearly in the analysis of areal and genealogical distribution in 4.5 and 5.4, in that the languages in which the fewest expressions could be identified were also among those with the least amounts of data analyzed. Although data discrepancies are not a sufficient explanation for all of the observed areal and genealogical tendencies, this potential confounding variable is nevertheless a methodological problem. To ensure the validity of the conclusions of a comparative analysis, a sufficient amount of comparable data should be available for all languages in the sample.

Issues of validity and generalisability also appear in the nature of the parallel corpus used. Stolz (2007, p. 102) cautions against equating 'non-authentic language' (such as translated texts) with natural language, as it is difficult to control for the inherent possibility of source language inference and variation in translation approach. The *OpenSubtitles* parallel data is translational and (equally crucially) scripted, which poses an inherent problem for analyzing and estimating the prevalence of any given phenomenon in natural language. It is, for instance, possible that the the dominance of calques and direct loans from English among extracted semi-numeral expressions (see Table 9) may be at least partly a result of translational artifacts rather than authentic usage (despite the post-extraction step of searches in monolingual corpora to attempt to verify authentic usage). Controlling for this type of source language inference was particularly difficult since the source language of the translated material was not known. Although information about the original language of each subtitled media is present in the *OpenSubtitles* corpus, this information is difficult to access through the available corpus analysis tools, and it is not certain that a given set of subtitles was actually translated from the original media language (rather than previously translated subtitles in a third language).

It is also important to note that the *OpenSubtitles.org* database is entirely open – anyone can upload subtitles, and there is no manual quality control process before subtitles are added to the database. While the *OpenSubtitles* corpus available through OPUS has been preprocessed to detect and clean up certain categories of errors (as described further in Tiedemann 2016 and Lison and Tiedemann 2016), variable translation quality is nevertheless a factor potentially impacting the validity of this study's results.

Despite these serious issues, spoken informal parallel data is clearly valuable for cross-linguistic investigation of this phenomenon (particularly in combination with other monolingual primary and secondary data sources), and the value of such data increases the more languages are sufficiently represented. Various other potential parallel data sources might be interesting for future investigations, such as subtitles for unscripted TV shows. Although neither unscripted nor necessarily authentically spoken, manuscripts of comic strips such as Donald Duck may also be an interesting data source. As exemplified in Chrisomalis 2016, pp. 20, 23, the specific linguistic domain of comic strips seems to be both a common origin of new quasi-numeral expressions and a domain where quasi-numerals occur in abundance.

5.6.2 Language sample

Vast variation among hyperbolic numerals and quasi-numerals was potentially left uncaptured by a lack of diversity in the language sample. Although this study was not intended to be an exhaustive survey of hyperbolic numerals and quasi-numeral expressions, the results and conclusions should nevertheless not be presumed to be cross-linguistically generalisable beyond the genera and macroareas which were well represented in the sample.

Even within represented genera, the uneven distribution of languages remains an issue for the validity of conclusions based on cross-genus comparison. For instance, Turkish, the only Turkic language in the sample, is most probably not a prototypical Turkic language from a feature perspective due to extensive contact with other European languages, yet it is the genus's sole representative in the *OpenSubtitles2018* parallel corpus [5] and thus by extension in the language sample of this study.

A heavily biased convenience sample was deemed sufficient (and unavoidable for parallel data availability reasons) for this exploratory study of a relatively unexplored phenomenon. However, further work aiming to obtain anything more than preliminary, questionably generalisable conclusions regarding the distribution of these expressions would require a sample balanced in a number of aspects. In addition to areal and genealogical stratification, a sufficiently balanced sample should also take numeral base and numeral system productivity into account, and (as mentioned above) it should be ensured that sufficient data is available for each included language.

5.6.3 Procedure and analysis

Most methodological issues are tied to the data issues discussed in section 5.6.1, since the method used for extraction of candidate expressions relies on the availability of a sufficient number of parallel sentences containing known hyperbolic (quasi-)numeral expressions in the parallel corpus. The use of seed expressions in multiple languages does increase the amount of parallel data that can be analyzed for each language, of course, and another extraction round using all hitherto extracted (quasi-)numerals as seed expressions would likely yield a variety of candidate expressions that were missed in this study.

[5] with the exception of Kazakh, which had too little data available for the language to be included in this study.

The seed expressions used in the extraction process are in and of themselves important to discuss as a methodological factor, as the expressions targeted by the chosen extraction method are first and foremost translational equivalents of the specific seed expressions. In the pilot extraction, only quasi-numeral seed expressions were used: *umpteen*, *ørten* and *femtioelva*. This delimitation was a necessity for the extraction procedure, as quasi-numerals (in contrast to hyperbolic numerals with exact values) are exclusively used in hyperbole and as such would not return non-hyperbolic contexts. However, the limited and somewhat overlapping set of seed expressions used is still a potential methodological weakness – given the tendencies toward inter-category variance observed in the translation probability analysis (see section 4.6), using additional quasi-numerals belonging to other categories as seed expressions would likely have resulted in a larger and more varied set of pilot contexts. This was addressed in the full extraction procedure as a greater variety of quasi-numeral expressions were included as seed expressions.

The semantic relationship between seed expressions and extracted expressions is also worth discussing in relation to the elicited data. While the extracted expressions are (as discussed above) by necessity translational equivalents of the seed expressions, expressions obtained through elicitation did not have this restriction. Although the example sentences provided in the elicitation query (see Appendix D) were constructed using the seed expressions used in extraction, replies did not always limit themselves to the specific example contexts, providing hyperbolic expressions that are commonly used in (or restricted to) other domains and contexts as well. This may have resulted in the set of expressions obtained in elicitation-and-extraction languages being more diverse than for extraction-only languages, and an extended study should aim to both conduct a larger and more systematic elicitation procedure and include a broader range of contexts in extraction.

Even with the aid of word alignments, parallel text extraction procedures which require manual verification and analysis of extracted expressions are time-consuming and difficult to scale. Although the phrase translation probability threshold for extracted expressions could be raised to filter out more inaccurate correspondences which would otherwise need to be manually identified, doing so would also risk excluding peripheral or infrequent (yet still highly relevant) expressions from the extraction.

In statistical machine translation, increasing the size of the training corpus typically improves performance (Koehn, Och et al. 2003). While additional data in each language would contribute to a more reliable extraction, as discussed in 5.6.1, parallel data with some form of morphosyntactic annotation would lead to even greater improvements in both efficiency and validity, as the burden of manual analysis of each extracted expression would decrease considerably.

5.7 Further research

A number of suggestions for future work regarding hyperbolic numerals and quasi-numerals have already been proposed. Most fundamentally, the extraction procedure could be extended to a broader language sample with better coverage and more data, to identify further hyperbolic numeral and quasi-numeral expressions which may have been missed entirely in this study. Similarly, the distributional analysis could be repeated with a balanced and stratified language sample (and comparable data) to identify cross-linguistic patterns with a higher degree of validity and generalisability. As discussed in 5.2 and reiterated in 5.6.2, the distribution of these expressions in languages with numeral bases other than 10 has yet to be thoroughly investigated.

Using the expressions gathered in this study, a deeper and broader investigation of the co-occurrence of hyperbolic numerals and quasi-numeral expressions could also be performed. Cross-linguistic comparisons of shared contexts similar to 4.6 but including a greater number and variety of expressions might reveal whether the preliminary categorical distinctions proposed in this study are functionally relevant or exclusively structural in nature. In addition to the translation probability matrix approach taken in this study, massively parallel data also enables cross-linguistic analysis of lexical similarity through semantic maps created using multidimensional scaling methods (as in Wälchli and Cysouw 2012).

Finally, a systematic investigation of the domain specificity of particular expressions may yield interesting results. Lavric (2010, p. 137) identifies some hyperbolic numeral expressions which are to varying degrees restricted to certain hyperbolic contexts, and domain-specific quasi-numeral expressions also seem to occur – for instance, the Hebrew quasi-numeral expression *tarapapu* only occurs in hyperbolic contexts relating to years in the distant past (D. Gil, personal communication, February 22, 2022) and was deemed too limited to this domain to include in the analysis of this study. As parallel text extraction may not be a suitable method of identifying strictly domain-specific hyperbolic quantifiers, an expanded elicitation procedure may provide more sufficient grounds for further analysis.

6 Conclusions

In this exploratory study, numerals and quasi-numeral expressions used in hyperbolic contexts across a sample of 50 languages were collected through parallel text extraction and elicitation, analyzed and grouped according to their functional and structural properties.

The main contribution of the study is a systematically assembled database of hyperbolic numerals and quasi-numeral expressions in the 46 languages in which such expressions could be identified. The full table of identified hyperbolic expressions, organized by language, can be found in Appendix B.

Based on the distribution of these two types of hyperbolic expressions across the languages in the sample, the following initial suggestion of an implicational universal is formulated:

(12) The Universal of Hyperbolic Quantifier Types
If a language uses quasi-numeral expressions for quantification in hyperbolic contexts, at least one numeral expression with a numerical value is also conventionalized in hyperbolic contexts.

In addition, a number of patterns in value and morphological construction with varying cross-linguistic prevalence were observed and discussed. Based on these observed clusters, an initial classification framework (presented in greater detail in 5.2 and 5.3) is proposed:

Hyperbolic numerals

1. **Round hyperbolic numerals**
2. **Non-round hyperbolic numerals**

Hyperbolic quasi-numeral expressions

3. **Semi-numeral expressions**
   - 3a. **Numeral-pattern quasi-numerals**
   - 3b. **Compound quasi-numerals**
4. **Pseudo-numeral expressions**
5. **Algebraic variable expressions**
6. ***many-th* expressions**

Areal and genealogical distributions of various categories of hyperbolic numerals and quasi-numeral expressions were also visualized and discussed. Despite an unstratified and unbalanced language sample with a strong Eurasian and Indo-European bias, some tendencies were observed that may provide direction for future typological work on these expressions.

Finally, a limited analysis of co-occurrence frequencies among extracted hyperbolic numeral and quasi-numeral expressions in translational parallel data was performed. This analysis revealed that specific expressions of these types are not necessarily used interchangeably, and that different expressions (despite functional and structural similarities) may be restricted to certain domains or contexts, or convey different magnitudes of hyperbolic quantification.

While there is doubtlessly vast further cross-linguistic variation within the domain of hyperbolic numerals and quasi-numeral expressions which has not been captured by the limited language diversity and scope of this study, the results nevertheless suggest that these hyperbolic expressions constitute a cross-linguistically common phenomenon, with internal variation in both function and form.

References

Agnihotri, Rama Kant (2007). *Hindi: an essential grammar.* Routledge essential grammars. OCLC: ocm72799377. London ; New York: Routledge, 2007.

Aulamo, Mikko, Umut Sulubacak, Sami Virpioja and Jörg Tiedemann (2020). OpusTools and Parallel Corpus Diagnostics. In: *Proceedings of the 12th Language Resources and Evaluation Conference* (2020), p. 8.

Benson, Morton (1998). *Standard English-SerboCroatian, SerboCroatian-English dictionary.* Cambridge: University Press, 1998.

BNC Consortium (2007). *The British National Corpus, XML Edition.* 2007. URL: http://hdl.handle.net/20.500.12024/2554 (visited on 05/04/2022).

Bradley, Peter T. and I. E. Mackenzie (2004). *Spanish: an essential grammar.* Routledge Essential grammars. OCLC: ocm53090990. London ; New York: Routledge, 2004.

Chan, Eugene (2022). *Numeral Systems of the World's Languages.* 2022. URL: https://lingweb.eva.mpg.de/channumerals/ (visited on 04/04/2022).

Chandralal, Dileep (2010). *Sinhala.* Vol. 15. London Oriental and African language library. OCLC: ocn460935998. Amsterdam, The Netherlands ; Philadelphia: John Benjamins Pub. Co, 2010.

Channell, Joanna (1994). *Vague language.* Oxford: Oxford Univ. Press, 1994.

Chrisomalis, Stephen (2016). Umpteen Reflections on Indefinite Hyperbolic Numerals. In: *American Speech* 91.1 (Feb. 2016), pp. 3–33.

Comrie, Bernard (2013). Numeral Bases. In: *The World Atlas of Language Structures Online.* Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. URL: https://wals.info/chapter/131.

Cysouw, Michael and Bernhard Wälchli (2007). Parallel texts: using translational equivalents in linguistic typology. In: *Language Typology and Universals* 60.2 (July 2007), pp. 95–99.

David, Anne Boyle (2015). *Descriptive grammar of Bangla.* Ed. by Thomas J. Conners and Dustin A. Chaćon. Mouton-CASL Grammar Series 2. Maryland: De Gruyter Mouton, 2015.

Dehaene, Stanislas and Jacques Mehler (1992). Cross-linguistic regularities in the frequency of number words. In: *Cognition* 43.1 (1992), pp. 1–29.

Dryer, Matthew S. (1992). The Greenbergian Word Order Correlations. In: *Language* 68 (1992), pp. 18–138.

Dryer, Matthew S. and Martin Haspelmath (2013). Genealogical Language List. In: *The World Atlas of Language Structures Online.* Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. URL: https://wals.info/languoid/genealogy.

Erjavec, Tomaž, Ştefan Bruda, Ivan Derzhanski, Ludmila Dimitrova, Radovan Garabík, Peter Holozan, Nancy Ide, Heiki-Jaan Kaalep, Natalia Kotsyba, Csaba Oravecz, Vladimír Petkevič, Greg Priest-Dorman, Igor Shevchenko, Kiril Simov, Lydia Sinapova, Han Steenwijk, Laszlo Tihanyi, Dan Tufiş and Jean Véronis (2010). *MULTEXT-East free lexicons 4.0.* ISSN: 2820-4042. 2010. URL: http://hdl.handle.net/11356/1041.

Francis, W.N. and H. Kučera (1982). *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton Mifflin, 1982.

Glinert, Lewis (2016). *Modern Hebrew: an essential grammar.* 4th edition. Routledge Essential Grammars. London and New York: Routledge, 2016.

Grice, Herbert Paul (1975). Logic and Conversation. In: *Speech Acts.* Ed. by Peter Cole and Jerry L. Morgan. Vol. 3. Syntax and Semantics. New York: Academic Press, 1975, pp. 41–58.

Göksel, Aslı and Celia Kerslake (2011). *Turkish: an essential grammar.* Routledge essential grammars. Milton Park, Abingdon ; New York: Routledge, 2011.

Hammarström, Harald (2010). Rarities in numeral systems. In: *Rethinking Universals.* Ed. by Jan Wohlgemuth and Michael Cysouw. De Gruyter Mouton, Mar. 2010, pp. 11–60.

Hewitt, B. G. (2005). *Georgian: a learner's grammar.* 2nd ed. Essential grammars. London: Routledge, 2005.

Holton, David, Peter Mackridge, Irene Philippaki-Warburton and Vassilios Spyropoulos (2012). *Greek: A Comprehensive Grammar of the Modern Language.* 2, revised. OCLC: 1053830990. Florence: Routledge, 2012.

Hurford, James R. (1987). *Language and number: the emergence of a cognitive system.* Oxford, UK ; New York, NY, USA: B. Blackwell, 1987.

— (1975). *The linguistic theory of numerals.* Cambridge: Cambridge U.P., 1975.

Hutchinson, Amélia P., Janet Lloyd and Maria Cristina Marques dos Santos Sousa (2019). *Portuguese: an essential grammar.* Third edition. Routledge essential grammars. New York: Routledge, Taylor & Francis Group, 2019.

Iliescu, Maria and Victoria Popovici (2013). *Rumänische Grammatik.* Hamburg: Helmut Buske Verlag, 2013.

Institutet för de inhemska språken (2022). *Stora finsk-svenska ordboken.* 2022. URL: https://kaino.kotus.fi/finsk-svensk/ (visited on 08/04/2022).

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý and Vít Suchomel (2013). The TenTen Corpus Family. In: *Abstract Book of the 7th international Corpus Linguistics conference.* Lancaster: UCREL, 2013, pp. 125–127.

Johansson, Stig (1980). Word frequencies in British and American English: Some prelimiary observations. In: *ALVAR: Stockholm Papers in Language and Literature.* Ed. by J Allwood and M. L. Jung. Stockholm, 1980, pp. 56–74.

K Dictionaries Ltd. (2015). *PASSWORD English–Malay Learner's Dictionary.* Dictionary. 2015. URL: https://dictionary.cambridge.org/dictionary/english-malaysian/ (visited on 09/04/2022).

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel (2014). The Sketch Engine: ten years on. In: *Lexicography* 1.1 (July 2014), pp. 7–36.

Kilgarriff, Adam, Siva Reddy, Jan Pomikálek and Avinesh Pvs (2010). A Corpus Factory for many languages. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10).* Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias. Valletta, Malta: European Language Resources Association (ELRA), 2010, p. 7.

Koehn, Philipp, Franz Josef Och and Daniel Marcu (2003). Statistical Phrase-Based Translation. In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.* Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 48–54.

Koehn, Philipp, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen and Christine Moran (2007). Moses: open source toolkit for statistical machine translation. In: *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL '07.* Prague, Czech Republic: Association for Computational Linguistics, 2007, p. 177.

Krifka, Manfred (2009). Approximate Interpretations of Number Words: A Case for Strategic Communication. In: *Theory and Evidence in Semantics* 189 (2009). Ed. by Erhard W. Hinrichs and John Nerbonne, p. 16.

Lavric, Eva (2010). Hyperbolic Approximative Numerals in Cross-Cultural Comparison. In: *New Approaches to Hedging*. Ed. by Gunther Kaltenböck, Wiltrud Mihatsch and Stefan Schneider. Vol. 9. Studies in Pragmatics. Bingley: Emerald Group Publishing Limited, Jan. 2010, pp. 123–164.

Lison, Pierre and Jörg Tiedemann (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)* (2016), pp. 923–929.

McCarthy, Michael and Ronald Carter (2004). "There's millions of them": hyperbole in everyday conversation. In: *Journal of Pragmatics* 36.2 (Feb. 2004), pp. 149–184.

National Institute of Korean Language (2022). *Korean-English Learners' Dictionary*. Online dictionary. 2022. URL: https://krdict.korean.go.kr/ (visited on 04/04/2022).

Neijmann, Daisy L. (2021). *Icelandic: an essential grammar*. Routledge essential grammars. London ; New York: Routledge, 2021.

Newmark, Leonard, Philip Hubbard and Peter R. Prifti (1982). *Standard Albanian: a reference grammar for students*. Stanford, Calif: Stanford University Press, 1982.

Ngô, Binh (2020). *Vietnamese: an essential grammar*. Routledge essential grammars. Abingdon, Oxon ; New York, NY: Routledge, 2020.

Oosterhoff, Jenneke A. (2015). *Modern Dutch grammar: a practical guide*. Routledge Modern grammars. London ; New York: Routledge, 2015.

Peet, Joseph (2008). *A Grammar of the Malayalam Language*. OCLC: 1122455825. Piscataway: Gorgias Press, LLC, 2008.

Press, Ian (1987). *A grammar of modern Breton*. Mouton grammar library 2. Berlin ; New York: Mouton de Gruyter, 1987.

Proudfoot, Anna and Francesco Cardo (2013). *Modern Italian grammar: a practical guide*. 3rd ed. Routledge modern grammars. London ; New York: Routledge, 2013.

Rosch, Eleanor (1975). Cognitive reference points. In: *Cognitive Psychology* 7.4 (Oct. 1975), pp. 532–547.

Rounds, Carol (2009). *Hungarian: An Essential Grammar*. OCLC: 1058455284. Hoboken: Taylor & Francis, 2009.

Ryding, Karin C. (2005). *A reference grammar of modern standard Arabic*. New York: Cambridge University Press, 2005.

Smyth, David (2014). *Thai: an essential grammar*. Second edition. Routledge essential grammars. Milton Park, Abingdon, Oxon ; New York: Routledge, 2014.

Sneddon, James N. (2010). *Indonesian Reference Grammar*. 2nd ed. OCLC: 630655881. Crows Nest, N.S.W: Allen & Unwin, 2010.

Spektors, Andrejs, Ilze Auzina, Roberts Dargis, Normunds Gruzitis, Peteris Paikens, Lauma Pretkalnina, Laura Rituma and Baiba Saulite (2016). Tēzaurs.lv: the Largest Open Lexical Database for Latvian. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2568–2571.

Språkrådet and University of Bergen (2022). *«ørten»*. 2022. URL: https://ordbokene.no/bm/42824/%C3%B8rten (visited on 04/04/2022).

Stolz, Thomas (2007). Harry Potter meets Le petit prince – On the usefulness of parallel corpora in crosslinguistic investigations. In: *Language Typology and Universals* 60.2 (July 2007), pp. 100–117.

Stolz, Thomas and Ljuba N. Veselinova (2013). Ordinal Numerals. In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. URL: https://wals.info/chapter/53.

Stump, Gregory (2010). The derivation of compound ordinal numerals: Implications for morphological theory. In: *Word Structure* 3.2 (Oct. 2010), pp. 205–233.

Tárnyiková, Jarmila (2010). Bags of Talent, a Touch of Panic, and a Bit of Luck: The Case of Non-Numerical Vague Quantifiers. In: *Linguistica Pragensia* 20.2 (Jan. 2010), pp. 71–85.

Tiedemann, Jörg (2016). Finding Alternative Translations in a Large Corpus of Movie Subtitles. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)* (2016), p. 5.

— (2012). Parallel Data, Tools and Interfaces in OPUS. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)* (2012), p. 5.

— (2008). Synchronizing Translated Movie Subtitles. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'2008)* (2008), p. 5.

Veselinova, Ljuba N. (2004). *Cross-linguistic distribution of numeral derivatives*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, Mar. 2004. URL: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1313560&dswid=7981.

— (2020). *Numerals in Morphology*. Mar. 2020. URL: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-559.

Wheeler, Max, Alan Yates and Nicolau Dols (1999). *Catalan: a comprehensive grammar*. OCLC: 252800880. London; New York: Routledge, 1999.

Wälchli, Bernhard and Michael Cysouw (2012). Lexical typology through similarity semantics: Toward a semantic map of motion verbs. In: *Linguistics* 50.3 (Jan. 2012).

Yip, Po-ching and Don Rimmington (2021). *Chinese: an essential grammar*. Third edition. Routledge essential grammars. London ; New York: Routledge, 2021.

Appendix A Languages, corpora and additional sources

Table 14: All languages and data sources in the sample, organized by ISO 639-3 label

ISO		Genealogical	Numeral	*OpenSubtitles*	Web crawl corpus		Additional
639-3	Language [6]	affiliation	base [7]	corpus size [8]	Name	Size	source(s) [9]
als	Albanian	Indo-European, Albanian	10	24.3M	-	-	Newmark et al. (1982)
arb	Arabic	Afro-Asiatic, Semitic	10	458.4M	arTenTen12	8.3G	Ryding (2005)
ben	Bengali	Indo-European, Indic	10	3.7M	bnWaC	13.8M	David (2015)
bos	Bosnian	Indo-European, Slavic	10	215.4M	bsWaC 1.2	286.9M	Benson (1998)
bre	Breton	Indo-European, Celtic	Hybrid 10-20	0.2M	-	-	Press (1987)
bul	Bulgarian	Indo-European, Slavic	10	617.8M	bgTenTen12	843.3M	Erjavec et al. (2010)
cat	Catalan	Indo-European, Romance	10	4.6M	caTenTen14	210.6M	Wheeler et al. (1999)
ces	Czech	Indo-European, Slavic	10	864.0M	csTenTen17	12.6G	Erjavec et al. (2010)
cmn	Chinese (Simplified)	Sino-Tibetan, Chinese	10	191.4M	zhTenTen17 Simplified	16.6G	Yip and Rimmington (2021); K. Chung (pers. comm., 19/02/22)
	(Traditional)		10	66.2M	zhTenTen17 Traditional	3.0G	J.P. Gates (pers. comm., 22/02/22)
dan	Danish	Indo-European, Germanic	Hybrid 10-20	206.7M	daTenTen20	4.1G	
deu	German	Indo-European, Germanic	10	288.0M	deTenTen13	19.8G	A. McIntyre (pers. comm., 18/02/22); R.E. Cramer (pers. comm., 19/02/22); V. Minow (pers. comm., 19/02/22); S. Nordhoff (pers. comm., 22/02/22)
ell	Greek	Indo-European, Greek	10	850.2M	elTenTen14	2.0G	Holton et al. (2012)
eng	English	Indo-European, Germanic	10	3.2G	enTenTen20	43.1G	B. Palmer (pers. comm., 22/02/22)
est	Estonian	Uralic, Finnic	10	168.2M	etTenTen19	623.0M	Erjavec et al. (2010)
eus	Basque	Isolate	Hybrid 10-20	5.7M	BasqueWaC v2	123.9M	S. Eliasson (pers. comm., 20/02/22)

[6] Languages in which expressions were elicited are underlined.

[7] Numeral base classification gathered from Comrie 2013 and Chan 2022.

[8] Corpus size listed in number of tokens.

[9] Language consultants, cited below as (pers. comm.), provided expressions in their respective language(s) during the elicitation procedure, but were not consulted in regard to extracted expressions in their respective language(s) or in any other capacities.

ISO		Genealogical	Numeral	*OpenSubtitles*	Web crawl corpus		Additional
639-3	Language	affiliation	base	corpus size	Name	Size	source(s)
fin	Finnish	Uralic, Finnic	10	281.5M	fiTenTen14	1.7G	Institutet för de inhemska språken (2022)
fra	French	Indo-European, Romance	10	791.0M	frTenTen12	11.4G	J. Kokkelmans (pers. comm., 22/02/22)
heb	Hebrew	Afro-Asiatic, Semitic	10	544.0M	heTenTen21	3.2G	Glinert (2016); N. Faust (pers. comm., 18/02/22); D. Gil (pers. comm., 22/02/22)
hin	Hindi	Indo-European, Indic	10	1.0M	hiTenTen17	1.4G	Agnihotri (2007)
hrv	Croatian	Indo-European, Slavic	10	707.5M	hrWaC 2.2	1.4G	Benson (1998)
hun	Hungarian	Uralic, Ugric	10	625.9M	huTenTen12	3.2G	Rounds (2009)
ind	Indonesian	Austronesian, Malayo-Sumbawan	10	137.2M	IndonesianWaC	109.2M	Sneddon (2010)
isl	Icelandic	Indo-European, Germanic	10	12.2M	isTenTen20	595.1M	Neijmann (2021)
ita	Italian	Indo-European, Romance	10	769.5M	itTenTen20	14.5G	Proudfoot and Cardo (2013); L.F. Mazzitelli (pers. comm., 22/02/22); R. Giomi (pers. comm., 22/02/22)
jpn	Japanese	Japonic, Japanese	10	23.7M	jaTenTen11	10.3G	S. Kalyan (pers. comm., 22/02/22)
kat	Georgian	Kartvelian	Hybrid 10-20	1.7M	kaWaC	63.6M	Hewitt (2005)
kor	Korean	Koreanic, Korean	10	10.2M	koTenTen18	2.1G	National Institute of Korean Language (2022)
lav	Latvian	Indo-European, Baltic	10	3.5M	lvTenTen14	657.5M	Spektors et al. (2016)
lit	Lithuanian	Indo-European, Baltic	10	11.6M	ltTenTen14	981.5M	Erjavec et al. (2010)
mal	Malayalam	Dravidian, South Dravidian	10	2.8M	malayalamWaC	21.2M	Peet (2008)
mkd	Macedonian	Indo-European, Slavic	10	50.2M	OPUS2 Macedonian	49.1M	Erjavec et al. (2010)
mly	Malay	Austronesian, Malayo-Sumbawan	10	22.8M	MalaysianWaC	230.4M	K Dictionaries Ltd. (2015)
nld	Dutch	Indo-European, Germanic	10	752.9M	nlTenTen14	2.6G	Oosterhoff (2015); S. Gregersen (pers. comm., 19/02/22); E. Visser (pers. comm., 22/02/22)

ISO		Genealogical	Numeral	*OpenSubtitles*	Web crawl corpus		Additional
639-3	Language	affiliation	base	corpus size	Name	Size	source(s)
nob	Norwegian	Indo-European, Germanic	10	86.1M	noTenTen17 Bokmål	2.9G	Språkrådet and University of Bergen (2022)
pes	Persian	Indo-European, Iranian	10	78.8M	TalkBank Persian	549.2M	F. Sabouri (pers. comm., 23/02/22)
pol	Polish	Indo-European, Slavic	10	1.4G	plTenTen12	9.4G	Erjavec et al. (2010); M. Dąbkowski (pers. comm., 22/02/22)
por	Portuguese (European)	Indo-European, Romance	10	804.7M	ptTenTen11	4.6G	Hutchinson et al. (2019)
	(Brazilian)		10	1.7G	ptTenTen11	4.6G	
ron	Romanian	Indo-European, Romance	10	1.3G	roTenTen16	3.1G	Iliescu and Popovici (2013)
rus	Russian	Indo-European, Slavic	10	290.6M	ruTenTen11	18.3G	Erjavec et al. (2010); D. Teptiuk (pers. comm., 22/02/22)
sin	Sinhala	Indo-European, Indic	10	5.7M	-	-	Chandralal (2010)
slk	Slovak	Indo-European, Slavic	10	103.4M	skTenTen11	876.0M	Erjavec et al. (2010)
slv	Slovenian	Indo-European, Slavic	10	360.7M	slTenTen15	988.5M	Erjavec et al. (2010)
spa	Spanish	Indo-European, Romance	10	1.5G	esTenTen18	19.6G	Bradley and Mackenzie (2004)
srp	Serbian	Indo-European, Slavic	10	1.1G	srWaC	516.5M	Benson (1998)
swe	Swedish	Indo-European, Germanic	10	243.0M	svTenTen14	3.9G	
tam	Tamil	Dravidian, South Dravidian	10	0.2M [10]	TamilWaC	32.9M	S. Kalyan (pers. comm., 22/02/22)
tha	Thai	Tai-Kadai, Kam-Tai	10	18.8M	thTenTen18	695.9M	Smyth (2014); R. Dockum (pers. comm., 22/02/22)
tur	Turkish	Turkic	10	962.5M	trTenTen12	4.1G	Göksel and Kerslake (2011); D. Coşkun (pers. comm., 28/02/22)
ukr	Ukranian	Indo-European, Slavic	10	7.9M	ukTenTen14	2.7G	Erjavec et al. (2010)
vie	Vietnamese	Austroasiatic, Viet-Muong	10	41.8M	VietnameseWaC	129.8M	Ngô (2020)

[10] As no expressions could be extracted from the Tamil subcorpus of *OpenSubtitles2018* due to a lack of contexts matching the seed expressions used in the extraction procedure, only elicited expressions were analyzed for Tamil.

Appendix B Full table of extractions

The full table of extracted and elicited hyperbolic numerals and quasi-numeral expressions has been packaged as a csv file.

This file contains five semicolon-delimited fields with the following content:

- Field 1: ISO 639-3 language label
- Field 2: Extracted word form of expression
- Field 3: Category label
- Field 4: Value (if applicable)
- Field 5: Origin(s) of expression in the study (pilot study, extraction and/or elicitation)

Appendix C Regular expression search terms

The following regular expressions were used in the extraction of hyperbolic numerals and quasi-numerals as described in section 3.3.

For each regular expression detailed below, the specific word forms it was intended to capture are listed in italics.

Pilot extraction

- `(umpteen(th)?)` – [eng] *umpteen, umpteenth*
- `(^ørten(de)?)` – [nob] *ørten, ørtende*
- `(femtio?-?el(va|fte))` – [swe] *femtielva, femtielfte, femtioelva, femtioelfte, femti-elva, femti-elfte, femtio-elva, femtio-elfte*

Full extraction

- `(117\.|hundredesyttende)` – [dan] *117., hundredesyttende*
- `((umpteen|[jz]illion)(th)?` – [eng] *bajillion, gazillion, gazillionth, umpteen, umpteenth*
- `(énième|bajillion)` – [fra] *énième, bajillion*
- `(^ørten(de)?|^e?n-?te$)` – [nob] *ente, n-te, nte, ørten, ørtende, ørten millioner*
- `(femtio?-?el(va|fte)|(z|skv)iljon(tals)?)` – [swe] *femtielva, femtielfte, femtioelva, femtioelfte, femti-elva, femti-elfte, femtio-elva, femtio-elfte, femtioelva miljoner, skviljontals, ziljon*

Appendix D Elicitation query

The following query was distributed through LINGUIST List on February 18th, 2022, and through Lingtyp on February 22nd, 2022.

> I'm looking for corresponding expressions (in any language) to the approximative numeral phrases in the examples below – conventionalised numeral expressions which typically express larger approximate numeric quantities and are used to encode various emotive functions. I'm interested in the composition and value of these numerals, as well as their emphatic and emotive functions – if there are other expressions in the numeral domain in your language(s) that carry a similar illocutionary force, I would love to hear about them as well!
>
> Swedish [swe] (from Bloggmix 2013, accessed through
> http://spraakbanken.gu.se/korp)
>
> Det finns nämligen femtioelva sorters myror
> 'There are actually many types of ants' (lit. 'There are actually fifty-eleven types of ants')
>
> French [fra] (Lavric 2010)
>
> Il n'y a pas trente-six façons de voir la chose
> 'There aren't very many ways of seeing the thing' (lit. 'There aren't thirty-six ways of seeing the thing')
>
> Danish [dan] (from OpenSubtitles2018, accessed through http://opus.nlpl.eu)
>
> Han fortalte mig 117 gange, at han ikke gjorde hende noget
> 'He told me a thousand times that he didn't do anything to her' (lit. 'He told me 117 times that he didn't do anything to her')
>
> English [eng] (from OpenSubtitles2018, accessed through http://opus.nlpl.eu)
>
> For the umpteenth time, we are not getting a dog
>
> Thank you very much in advance for any tips, examples or comments!

Stockholm University
SE-106 91 Stockholm, Sweden
Telephone +46 (0)8 16 20 00
https://www.su.se/