Imperial College London

Numerolinguistics

Imperial College London
Department of Mathematics
Danilo Jr Dela Cruz, Julian Yu

Abstract

In the Gresham College lecture "Where do Mathematical Symbols Come From?" [1], Sarah Hart introduces the subject of Numerolinguistics which concerns itself with the relationship between the value and character length of numbers in a given language. We aim to answer some of the "Homework" questions provided at the end of the lecture by proposing a model that places an upper bound on the length of numbers.

June 2021

GitHub Repository: https://github.com/djd19/numerolinguistics

Contents

1 Introduction

Numerolinguistics concerns itself with the relationship between the value length of a given number. The **length** of a number refers to the number of characters the word used to describe the number has after removal of blacklisted characters such as spaces and hyphens. Naturally, this depends on the language we are using. Therefore, we will study the function $f_{\text{lang}} : \mathbb{N} \to \mathbb{N}$ which returns the length of a number in the language lang.

For example,

$$f_{\text{English}}(3) = \text{length}(\text{three}) = 5$$
$$f_{\text{English}}(123) = \text{length}(\text{onehundredtwentythree}) = 21.$$

We consider the directed graph that arises when we point from a number n to $f_{\text{lang}}(n)$.

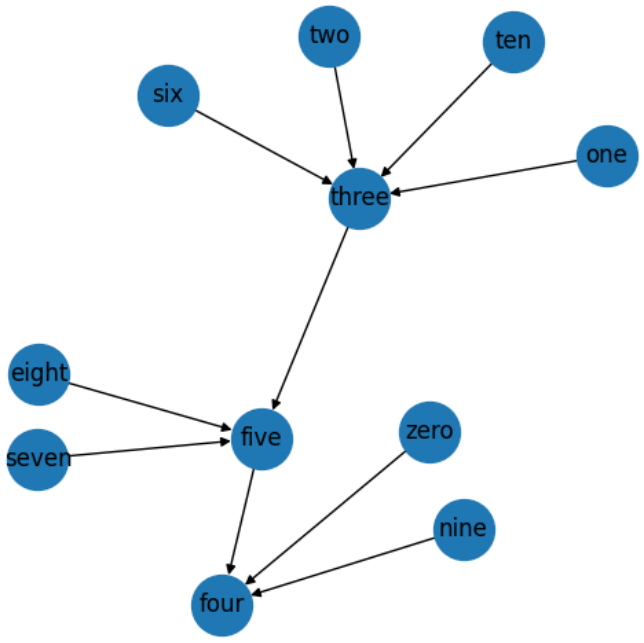

Figure 1: A directed graph of f_{English} applied on the first 10 numbers. "four" is a fixed point.

Upon repeated application of f, we may eventually stabilise at a **cycle**. Cycles are a collection of points that eventually repeat under f. Formally a collection of points $C \neq \emptyset$ is a cycle if:

$$\exists k \in \mathbb{N} : \forall n \in C, f^k(n) = n$$
$$\wedge \forall x, y \in C : \exists t \in \mathbb{N} : f^t(x) = y.$$

A cycle is said to have **length** k if k is the smallest integer such that $f^k(n) = n$ for some n in the cycle.

A cycle is said to be a **fixed point** if it has length one and cycles of length greater than one are **proper cycles**. Examples can be seen in Figures 1 and 2 respectively.

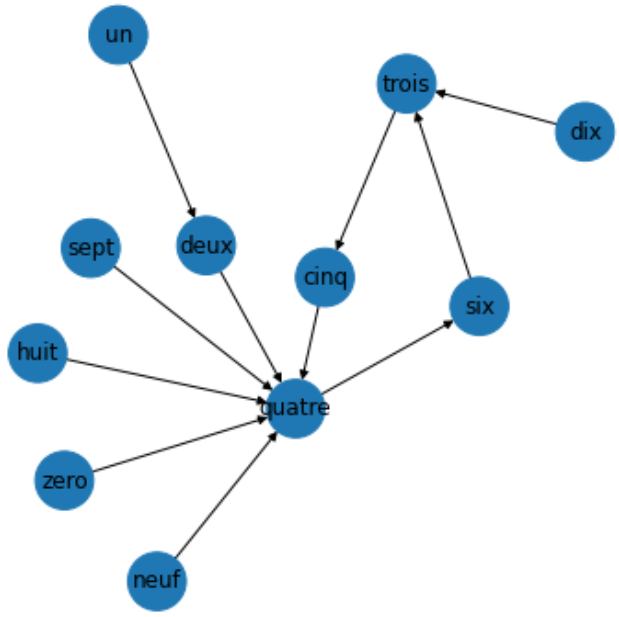

Figure 2: A directed graph of f_{French} applied on the first 10 numbers. We have a cycle of length 4: "trois" $\rightarrow$ "cinq" $\rightarrow$ "quatre" $\rightarrow$ "six".

A **safe testing threshold** (STT) refers to the value N such that $[0, N]$ contains all cycles of f_{lang}. Sarah Hart proposes that $N = 100$ is often sufficient [1].

Homework

Given this mathematical background, we can now state the homework questions.

- What's a safe testing threshold for all languages?
- What's the highest fixed point in any language?
- What is the longest cycle in any language?

Though technically not a language, all numbers in the tally system are fixed points as the length of numbers grow as fast as their value. Since the tally system has fixed points that can be arbitrarily large, it can be said that it has the highest fixed point. However, modern languages do not use the tally system to represent their numbers as it renders discussion of large numbers unfeasible. Instead, most languages arrive at a system where the length grows much slower than value of numbers. This assumption allows us to demonstrate the existence and find estimates of STTs. More will be explored in the next section.

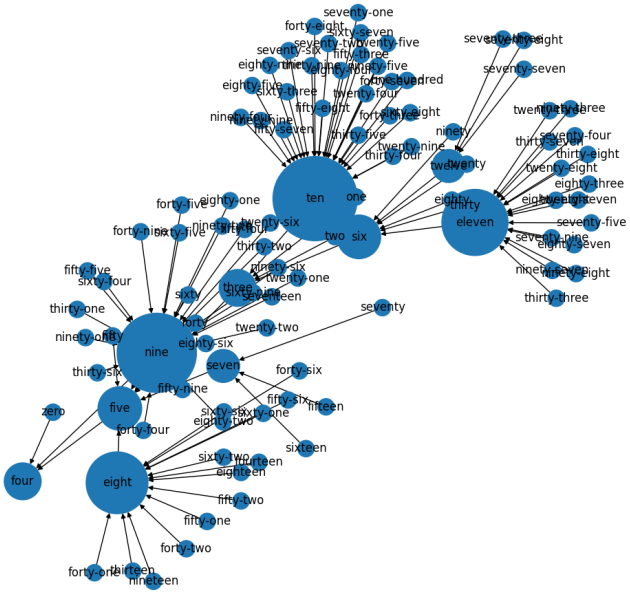

Figure 3: A directed graph of f_{English} applied on the first 100 numbers (where the size of a node is proportional to its degree). At least for the first 100 numbers there are no cycles. Is it possible for cycles to exist above one hundred?

2 Safe Testing Threshold

If we can estimate the length of numbers, then we may be able to determine if the length grows slow enough to allow for the existence of STTs. In this section, we propose a model for the length of numbers and study its implications.

Number Length Model

Assuming that a language obeys a decimal place value system, we can form a correspondence between blocks of words and each digit of the number. E.g.,

$$\underbrace{\text{one hundred}}_{1}\ \underbrace{\text{twenty}}_{2}\ \underbrace{\text{three}}_{3} = 100 + 20 + 3 = 123.$$

Let us further assume that we have an upper bound m for the length of these blocks. Therefore, an upper bound for the length of a number n is given by

$$f(n) \leq m \cdot (\text{number of digits of}\, n) = m \cdot (\lfloor \log_{10} n \rfloor + 1) := \text{ul}(n).$$

For mathematical simplicity, we will often consider

$$\text{ul}(n) \leq m \cdot (\log_{10} n + 1) := \widehat{\text{ul}}(n).$$

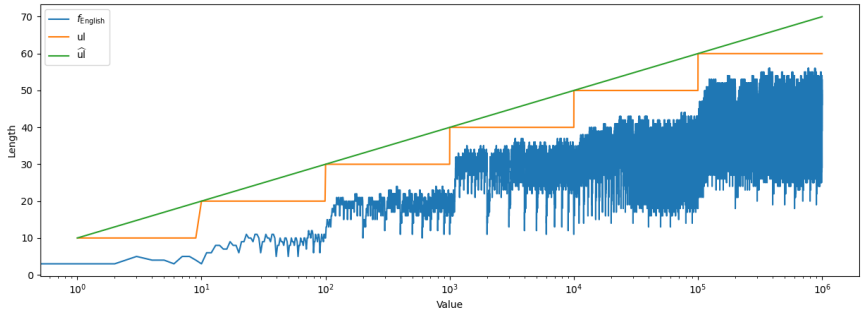

Figure 4: Comparison of model bounds to real lengths in English for $m = 10$.

Invariant Sets and the Existence of Cycles

Lemma 1. *Given a finite set S and a function $f : S \to S$. Then $\forall s \in S, (f^i(s))_{i=0}^{\infty}$ must stabilise in a cycle.*

Proof. Fix $s \in S$. By the pigeonhole principle, the sequence $f^i(s)$ must eventually repeat an element. This induces a cycle because f has a unique output value. □

Lemma 2. *Assuming the Number Length Model, if there exists $N \in \mathbb{R} : \forall n \geq N, \widehat{\text{ul}}(n) < n$ then:*

1. *$[0, N]$ is an invariant set.*
2. *f must have a cycle.*
3. *All cycles are contained in $[0, N]$.*

Proof.

1. Because $\widehat{\text{ul}}$ is an increasing function, $\forall n \leq N$

$$f(n) \leq \widehat{\text{ul}}(n) \leq \widehat{\text{ul}}(N) < N$$

   Therefore $f([0, N]) \subset [0, N]$.

2. This follows directly from Lemma 1 by considering the sequence $f^i(0)$. Note that 1 requires the domain to be finite. But because $f : \mathbb{N} \to \mathbb{N}$, we can consider the finite restriction of $[0, N] \cap \mathbb{N}$.

3. Consider any cycle C. By the well ordering principle, this must have a minimal element which we denote with n_0. $n_0 < N$, otherwise we would contradict its minimality. Because $[0, N]$ is invariant, we deduce $C \subset \{f^i(n_0)\} \subset f([0, N]) \subset [0, N]$.

□

Lemma 3. *There exists* $N \in \mathbb{N} : \forall n \geq N, \widehat{\text{ul}}(n) < n$.

Proof. Define $g(n) := \widehat{\text{ul}}(n) - n$.

$$\frac{dg}{dn} = \frac{m}{n \cdot \ln 10} - 1, \qquad \forall n > m/\ln 10, g'(n) < 0$$

If $\exists N > m/\ln 10 : (g(N) < 0$, we deduce that $\forall n \geq N, g(n) < 0$. Because ul is logarithmic, g must eventually be negative and so such an N must exist. □

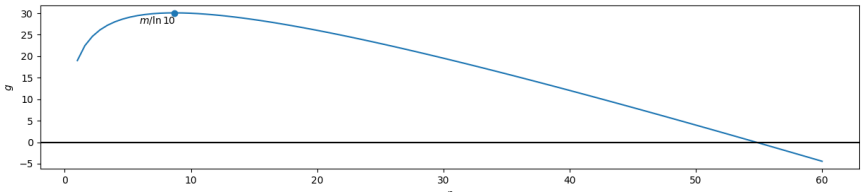

Figure 5: A plot of g for $m = 20$.

Safe Testing Threshold

Lemma 2 tells us the conditions required for a safe testing threshold N and Lemma 3 tells us how to find N: we simply need to find $N > m/\ln 10$ such that $g(N) < 0$.

We define the **minimal** N for a given m to be $N > m/\ln 10 : g(N) = 0$. Plotting this against m, we see that $4m$ is a sufficient heuristic for obtaining a STT.

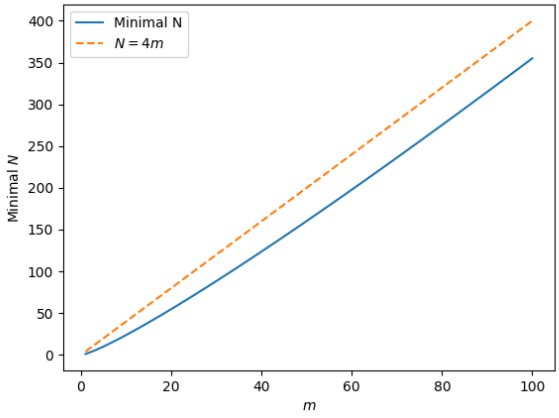

Figure 6: A plot of minimal N for "realistic" values of m.

It is important to note that this model has some oversights. In section 4, we critique the current model and study modifications which yield more realistic models.

3 Results

Using the heuristic from section 2, we deduce that $N = 100$ is a STT if $m \leq 25$. $m = 25$ is pretty extreme but it affirms that checking up to 100 is sufficient to detect all cycles.

Obtaining a list words for 1-100 in various languages from Airnet [2] and Lexis Rex [3], we have found that:

- Alamblak has a fixed point at 28, ("yima yohtti tir yohtti hosfirpat") narrowly beating Zulu's fixed point at 27 ("amashumi amabili nesikhombisa").
- French is unmatched in terms of maximum cycle length.
- Zulu has the most fixed points with 6 fixed points.
- Arabic has the most cycles with 2 cycles.

In our search, we have decided to ignore hyphens and use the full blacklist. Relaxing this leads us to a fixed point at 33 in Nahuatl ("cem-pohualli-om-mahtlactli-om-eyi"). Though it seems that spaces were simply replaced with hyphens and it's questionable if it *counts*.

Blacklist	**HFP**	**MCL**	**MFP**	**MPC**
()	(ndom, 34)	(alamblak, 4)	(ganda, 6)	(cuzco_quechua, 3)
(S, A, B, C)	(nahuatl, 33)	(french, 4)	(zulu, 6)	(arabic, 2)
(H, S, A, B, C)	(alamblak, 28)	(french, 4)	(zulu, 6)	(arabic, 2)

Table 1: The results for the Highest (HFP) Fixed Point, Maximum Cycle Length (MCL), Most Fixed Points (MFP), Most Proper Cycles (MPC) under a given blacklist.

Blacklist Key: H – Hyphen, S – Space, A – Apostrophe, B – Backtick, C – Comma

For fuller results, see the appendix. In appendix A we list the top results under all subsets of the full blacklist. And in appendix B we show how each language performs in terms of cycles under the full blacklist.

4 Collapse of the Upper Bound

The model is only guaranteed to be valid when the upper bound m is valid. However, there is nothing to stop m becoming invalid. Indeed the length of a block is unbounded if we use finitely many symbols (such as the English alphabet) because there are infinitely many integers. Therefore, m will eventually become invalid. But does it matter?

A case study in English

In Figure 4, we chose m to be 10. However, larger numbers may have blocks longer than 10 letters. Consider the number $777,333$. The blocks will be as follows:

[seven hundred] [seventy] [seven thousand] [three hundred] [thirty] [three]

Here we have 2 blocks of length 12 and 1 block of length 13 and therefore the upper bound $m = 10$ is invalid.

In English, blocks only come in three forms:

- A multiple of 100 between 100 and 900
- A multiple of 10 between 20 and 90, or a number between 11 and 19)
- A number between 1 and 9, plus a suffix representing a power of 10^3 such as "thousand" or "million".

The first and second of these only contain a few numbers, and the longest blocks in the first two cases can be found by checking all the cases. The third case, however, contains an unknown suffix which can contain any number of letters. In English, the longest blocks belonging to these two groups are "three hundred" and "seven hundred".

In English, the longest blocks in numbers below 10^4 are "three thousand" and "seven thousand". These have 13 letters each. In fact, the longest blocks in numbers below 10^{69} are "three quattuordecillion" and "seven quattuordecillion", which have 17 letters each. From this, we can guarantee that $m = 25$ clears the possibility of any cycles occurring in $(10^2, 10^{69})$. But this still leaves the possibility for cycles to exist above 10^{69}.

However, it is possible that even when this bound is surpassed that the value of a number has far surpassed the length of the number thus eliminating the possibility of cycles.

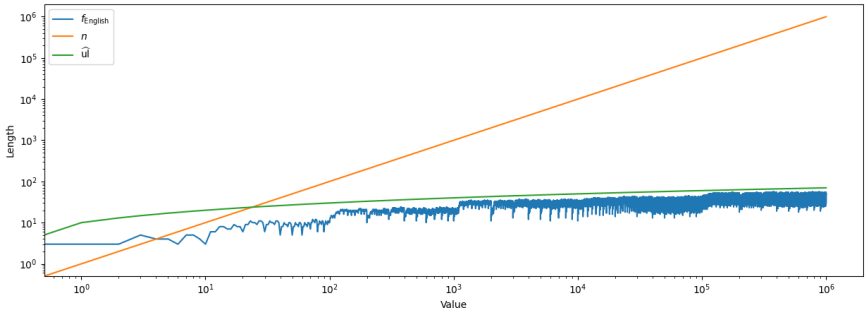

Figure 7: A plot of value against length of English numbers.

Allowing m to grow

Consider blocks of the third type (as defined in the previous section). There is no limit on the length of suffixes but it is unlikely they would start growing fast enough to catch up to their value. Let us suppose that they grow no more than 5 letters every power of 10^3.

Since the largest block below 1000 in English has length 12, we get the following results:

Bound	Number of Blocks	Largest Block Size
$< 10^3$	≤ 3	12
$< 10^6$	≤ 6	≤ 17
$< 10^9$	≤ 9	≤ 22
$\vdots$	$\vdots$	$\vdots$
$< 10^{3n}$	$\leq 3n$	$\leq 5n + 7$

Note that an upper bound for the length of a word is equal to the number of blocks multiplied by the largest possible block size. This means that under this model, for all $k < 10^{3n}$,

$$f_{\text{English}}(k) \leq 3n(5n+7).$$

We claim that it is sufficient to only check numbers up to 36.

Lemma 4. *Under this model, for all integers $k > 36$, $f_{\text{English}}(k) < k$.*

Proof. First suppose $k \geq 1000$. This means that there is an integer $n \geq 2$ such that $10^{3(n-1)} \leq k < 10^{3n}$. Furthermore, it is easy to show that $3n(5n+7) < 10^{3(n-1)}$ for all $n \geq 2$, since the former grows quadratically and the latter grows exponentially. Hence,

$$f_{\text{English}}(k) \leq 3n(5n+7) < 10^{3(n-1)} \leq k.$$

For $k < 1000$, we know from the table that $f_{\text{English}}(k) < 3 \cdot 12 = 36$. This means that for $36 < k < 1000$, we have $f_{\text{English}}(k) < 36 < k$. □

This model can be adapted to other languages by changing the largest block size for numbers $k < 10^3$ and the increase in word length after moving to a new power of 10^3.

5 Conclusion

In terms of a safe testing threshold, most modern languages will tend to have numbers whose length grow slow enough to make discussion of large numbers tractable. For this reason they will have a relatively low STT and often 100 will be more than sufficient.

The highest fixed points are likely to occur in languages that don't obey the decimal system and allow length to grow faster. This may occur by using a lower base system like Ndom which uses base 6 or like Alamblak which doesn't use a place-value system at all!

Here we have only considered 48 languages, but there could still be languages that beat Alamblak in terms of fixed points and French in terms of cycles. There are sources such as:

- Omniglot [4] which have a wider range of languages from 1-100 but is harder to parse.
- Zompist [5] which have 1-10 in 5000 languages. But 10 is not a STT and it is possible we may miss out a lot of cycles.

References

[1] Sarah Hart. Where do mathematical symbols come from?, May 2021. URL https://youtu.be/Edewyp87W-Q?t=3510.

[2] URL http://www.sf.airnet.ne.jp/~ts/language/number.

[3] URL https://www.lexisrex.com/.

[4] URL https://www.omniglot.com/language/numbers/index.htm.

[5] URL https://www.zompist.com/numbers.shtml.

6 Appendix

A Results by Blacklist

Blacklist	HFP	MCL	MFP	MPC
()	(ndom, 34)	(alamblak, 4)	(ganda, 6)	(cuzco_quechua, 3)
(H,)	(huli, 36)	(alamblak, 4)	(ganda, 6)	(cuzco_quechua, 3)
(S,)	(nahuatl, 33)	(french, 4)	(cuzco_quechua, 6)	(arabic, 2)
(A,)	(ndom, 34)	(alamblak, 4)	(aymara, 5)	(zulu, 3)
(B,)	(ndom, 34)	(alamblak, 4)	(ganda, 6)	(cuzco_quechua, 3)
(C,)	(huli, 36)	(alamblak, 4)	(ganda, 6)	(cuzco_quechua, 3)
(H, S)	(alamblak, 28)	(french, 4)	(zulu, 6)	(arabic, 2)
(H, A)	(huli, 36)	(alamblak, 4)	(ganda, 5)	(cuzco_quechua, 3)
(H, B)	(huli, 36)	(alamblak, 4)	(ganda, 6)	(cuzco_quechua, 3)
(H, C)	(ndom, 34)	(alamblak, 4)	(ganda, 6)	(cuzco_quechua, 3)
(S, A)	(nahuatl, 33)	(french, 4)	(zulu, 6)	(arabic, 2)
(S, B)	(nahuatl, 33)	(french, 4)	(cuzco_quechua, 6)	(arabic, 2)
(S, C)	(nahuatl, 33)	(french, 4)	(cuzco_quechua, 6)	(arabic, 2)
(A, B)	(ndom, 34)	(alamblak, 4)	(aymara, 5)	(zulu, 3)
(A, C)	(huli, 36)	(alamblak, 4)	(aymara, 5)	(zulu, 3)
(B, C)	(huli, 36)	(alamblak, 4)	(ganda, 6)	(cuzco_quechua, 3)
(H, S, A)	(alamblak, 28)	(french, 4)	(zulu, 6)	(arabic, 2)
(H, S, B)	(alamblak, 28)	(french, 4)	(zulu, 6)	(arabic, 2)
(H, S, C)	(alamblak, 28)	(french, 4)	(zulu, 6)	(arabic, 2)
(H, A, B)	(huli, 36)	(alamblak, 4)	(ganda, 5)	(cuzco_quechua, 3)
(H, A, C)	(ndom, 34)	(alamblak, 4)	(ganda, 5)	(cuzco_quechua, 3)
(H, B, C)	(ndom, 34)	(alamblak, 4)	(ganda, 6)	(cuzco_quechua, 3)
(S, A, B)	(nahuatl, 33)	(french, 4)	(zulu, 6)	(arabic, 2)
(S, A, C)	(nahuatl, 33)	(french, 4)	(zulu, 6)	(arabic, 2)
(S, B, C)	(nahuatl, 33)	(french, 4)	(cuzco_quechua, 6)	(arabic, 2)
(A, B, C)	(huli, 36)	(alamblak, 4)	(aymara, 5)	(zulu, 3)
(H, S, A, B)	(alamblak, 28)	(french, 4)	(zulu, 6)	(arabic, 2)
(H, S, A, C)	(alamblak, 28)	(french, 4)	(zulu, 6)	(arabic, 2)
(H, S, B, C)	(alamblak, 28)	(french, 4)	(zulu, 6)	(arabic, 2)
(H, A, B, C)	(ndom, 34)	(alamblak, 4)	(ganda, 5)	(cuzco_quechua, 3)
(S, A, B, C)	(nahuatl, 33)	(french, 4)	(zulu, 6)	(arabic, 2)
(H, S, A, B, C)	(alamblak, 28)	(french, 4)	(zulu, 6)	(arabic, 2)

B Cycles by Language

Language	MFP	Highest Fixed Point	MCL	Longest Cycle
ainu	7	arwanpe	1	arwanpe
alamblak	28	yima yohtti tir yohtti hosfirpat	1	tir hosfirpati hosf
arabic	-	-	2	thalathata 'ashar → khamsata 'ashar
assyrian	7	shaawaa	1	ishtaa
aymara	11	tunka-mayani	2	suxta → phisca
basque	9	bederatzi	2	hiru → lau
chinese	3	San	1	Er
chinook_wawa	14	tahtlum pe lakit	2	tahtlum pe sinamokst → tahtlum pe stotekin
croatian	3	tri	2	cetiri → sest
cuzco_quechua	15	chunka phisqa-yoq	2	chunka kinsa-yoq → chunka tawa-yoq
danish	4	fire	1	tre
dutch	4	vier	1	vier
english	4	four	1	four
esperanto	4	kvar	1	tri
estonian	4	neli	1	neli
finnish	5	viisi	2	kahdeksan → yhdeksan
french	-	-	4	quatre → six → trois → cinq
ganda	12	kkumi na bbiri	1	nnya
georgian	5	khuti	1	khuti
german	4	vier	1	vier
greek	5	pente	2	tessera → epta
hawaiian	11	ʻumikumakahi	1	ʻelima
hindi	2	do	2	chhuh → paanch
huli	5	duria	1	duria
hungarian	4	negy	2	ot → ketto
indonesian	-	-	2	empat → lima
italian	3	tre	1	tre
javanese	-	-	2	papat → lima
kiribati	-	-	3	tebwi ma teuana → tebwi ma tenua → tebwi ma uoua

Continued on next page

Language	MFP	Highest Fixed Point	MCL	Longest Cycle
latin	-	-	2	octo → quattuor
maltese	5	hamsa	1	hamsa
manx	7	shiaght	1	queig
nahuatl	10	mahtlactli	1	eyi
ndom	-	-	3	mer an thef → mer abo ithin → mer abo meregh
norwegian	4	fire	1	to
ojibwa	-	-	2	nishwaaswi → midaaswi
portuguese	5	cinco	2	seis → quatro
romanian	5	cinci	1	cinci
russian	11	odinnadtsat'	1	tri
spanish	5	cinco	2	cuatro → seis
swahili	-	-	2	tatu → nne
swedish	4	fyra	1	fyra
tagalog	4	apat	1	apat
thai	-	-	2	sahm → see
vietnamese	-	-	2	hai → ba
wolof	19	fukka ak juroom nenent	2	fukka ak naar → fukka ak benna
yoruba	5	marun	1	marun
zulu	27	amashumi amabili nesikhombisa	2	amashumi amabili nane → ishumi nesishiyagalolunye