Decomposing numerals

Authors: Isidor Konrad Maier, Matthias Wolff

Why?

To understand how computers may understand numerals

What?

An arithmetic-based unsupervised decomposition algorithm for natural numerals.
It divides a numeral into its stem and its input subnumerals.
Required algorithm input:

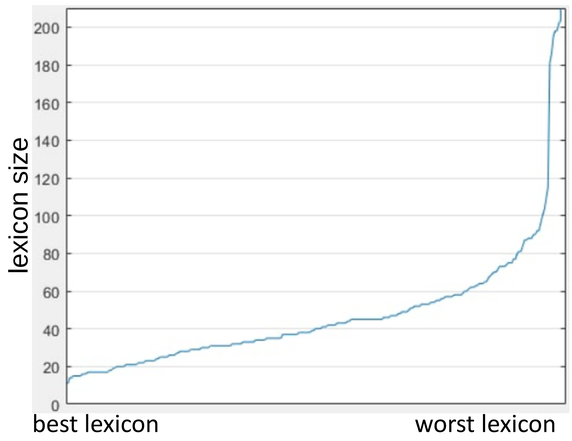
- a numeral word,
- its number value and
- a lexicon of numeral-number-pairs that could appear as subnumerals.

What for?

- The **stems** of a dataset of numerals are a lexicon to **restore the whole dataset**
- **Machine learning** of numerals in value-ascending order
- Stems, such as ‚_ hundred and _‘ are **affine linear** functions.

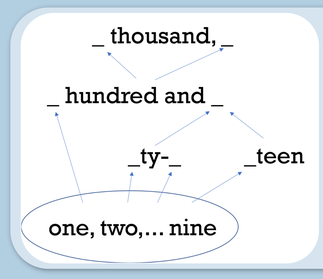
Results

From 255 languages we decomposed datasets of numerals. The graph shows **which amount of different stems** the determined lexica have, in order to **generate** the datasets, which mostly contain **1000 numerals**.



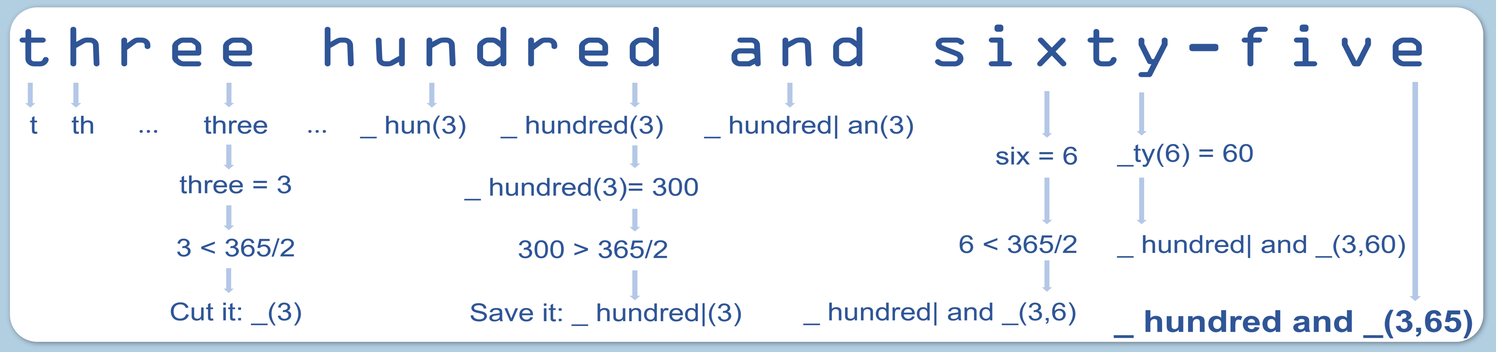
Brandenburg University of Technology Cottbus - Senftenberg

Sources and related work:
https://tinyurl.com/5n83b7ay



Hi, I have drawn this graph with the stems of numeral words. These stems, like '_ hundred and _' can also store meaning, e.g. (3)hundred and(65)=3*100+65.
Is there a way I can read out the stem of a numeral word, so I can find out that ‘three hundred and sixty-five’ means 3*100+65? 🤔

I would do it this way:



I scan the numeral from left for subnumerals to cut out. I extend the cuts as long as the value of the cut out numeral is less than half the whole numeral. And when we find a subnumeral larger than half the numerals value - like 300>365/2 in the example - we save the current decomposition and no longer scan this first part.

I didn’t understand everything, but it seems like I do not even have to know that the stem '_ hundred and _' exists in order to assign 365 to it. So, we could even apply this as a learning algorithm. 🤔

Exactly. The algorithm only needs to know the value of the numeral to decompose. And obviously, it needs to have heard about 'three' and 'sixty-five', otherwise these subwords do not mean anything to it. For more examples – also in other languages – check out this app. https://tinyurl.com/jwtawcsn 😉

Scan to try out the algorithm

Aah, so this does not only work in English? 😮

Yes, we do not even rely on base 10. The algorithm is tested on 255 languages. For some languages the performance was better, for others worse. We have provided an analysis of the performances in https://tinyurl.com/mr4yy92u

Scan to see performance of the algorithm

Great, I'll check it out! 👍