How to extract numbers from sentences written as phrases

Here’s what the data looks like:

|  |  |
| --- | --- |
| **Original Sentence** | **Revised Sentence** |
| 9 years' imprisonment | 11 years and 5 months' imprisonment |
| 2 years' imprisonment suspended for 2 years | Sentence Unchanged |
| 12 years' imprisonment | 14 years' imprisonment |
| 22 months' imprisonment | Sentence Unchanged |

This is what you need to do:

* Identify \*where\* the number of years is stated
* Extract that number of years
* Identify \*where\* the number of months is stated
* Extract that number of months
* Identify other context like ‘suspended’, ‘minimum’ etc.
* Convert years and months to a number
* Convert years and months to a common measure (total months)

## Identify where the years/months are detailed

Use SEARCH to tell you where the first mention of a word appears. It’s case-insensitive:

=search("year",G14)

Translation: Search for where year appears in I2

This (written in cell I2) returns a position, e.g. 4. If it doesn’t find it, it returns #VALUE

We assume that the number of years appears 3 positions before that (one space, plus two digits). So we subtract 3 to get the position of the \*number\* of years.

=search("year",G14)-3

Translation:

## Extract the number of years/months (and correct for problems)

If it’s a single figure, we should get just the space before it, which is fine. But if it’s at the start of a line, we won’t, and will get an error instead.

So we need a new column (L) to correct for that:

=IF(I14=0,1,I14)

Now nest that within a MID function to extract the numbers at that position:

=MID(H14, IF(I14=0,1,I14),2)

This grabs characters from that position, and continues grabbing for 2 characters.

However, this might also grab spaces if only one character is a digit.

To correct that, and make sure the result is formatted as a number, nest this again in an INT function (which turns a value into an integer):

=INT(MID(I2, IF(K2=0,1,K2),2))

You can adapt and repeat the above two processes for months - and also for the revised years and months - to get the four numbers you need.

## Identify unusual words

It’s worth adding a column which just measures the length of the description cell, so you can sort it and pick out outliers:

=LEN(I2)

The longest cell on that basis is this one: “4 years and 6 months’ imprisonment with a licence extension of 2 years and 6 months” which can help you decide to add a column checking for those key words:

=COUNTIF(I2,"\*licence\*")

This will count 1 if ‘licence’ is in that cell, or 0 if it is not. The asterisks are wildcards which mean ‘any or no characters’.

You can use the same function to check whether a sentence mentions years or months:

=COUNTIF(I2,"\*year\*")

A 0 value should match up with a #VALUE error in your earlier formula. You can use filters to check that they all do, and investigate any that don’t.

Repeat for month, and for revised year and month.