Plotting alignment data

In [1]:
```python
%matplotlib inline
import pandas as pd
import utils.db_utils as db
import utils.plot_utils as plot

targetLang = 'en'
bibleType = 'en_ult'
dbPath = f'./data/{bibleType}_alignments.sqlite'

connection = db.initAlignmentDB(dbPath)
```

```
Connection to SQLite DB successful
```

In [2]:
```python
def getDataFrameForOriginalWords(connection, word, searchLemma = True):
    # print(f"searchLemma = {searchLemma}")
    words = [word]
    alignments_ = db.getAlignmentsForOriginalWords(connection, words, searchLemma)
    alignments = pd.DataFrame(alignments_[word])
    return alignments
```

In [3]:
```python
# find all alignments for this lemma

word = 'θεός'
lemmaAlignments = getDataFrameForOriginalWords(connection, word, searchLemma = True)
lemmaAlignments
```

```
updating 'θεός'
splitLemmasAndAddData - found original word 'Θεός' for lemma
splitLemmasAndAddData - found original word 'Θεὸς' for lemma
splitLemmasAndAddData - found original word 'Θεοῦ' for lemma
Replacing "God s" with "God's"
Replacing "the things that are God s" with "the things that are God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "of God s" with "of God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "of God s" with "of God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "with God s words" with "with God's words"
Replacing "God s" with "God's"
Replacing "of God s" with "of God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
splitLemmasAndAddData - found original word 'Θεόν' for lemma
splitLemmasAndAddData - found original word 'Θεὸν' for lemma
splitLemmasAndAddData - found original word 'Θεῷ' for lemma
splitLemmasAndAddData - found original word 'Θεέ' for lemma
splitLemmasAndAddData - found original word 'θεοί' for lemma
splitLemmasAndAddData - found original word 'θεοὺς' for lemma
splitLemmasAndAddData - found original word 'θεοῦ' for lemma
splitLemmasAndAddData - found original word 'θεοὶ' for lemma
splitLemmasAndAddData - found original word 'θεὸν' for lemma
splitLemmasAndAddData - found original word 'θεόν' for lemma
splitLemmasAndAddData - found original word 'θεὸς' for lemma
splitLemmasAndAddData - found original word 'θεοῖς' for lemma
```

Out[3]:

	id	book_id	chapter	verse	alignment_num	orig_lang_keys	target_lang_keys	origSpan	origWords	origWordsTxt
0	326	mat	1	23	16	,402,403,	,518,	1	[{'id': 402, 'book_id': 'mat', 'chapter': '1',...	ὁ Θεός
1	17427	mrk	2	7	10	,19162,19163,	,24951,	1	[{'id': 19162, 'book_id': 'mrk', 'chapter': '2...	ὁ Θεός
2	22758	mrk	10	18	10	,24891,24892,	,32410,	1	[{'id': 24891, 'book_id': 'mrk', 'chapter': '1...	ὁ Θεός
3	26691	mrk	15	34	17	,29156,29157,	,38064,	1	[{'id': 29156, 'book_id': 'mrk', 'chapter': '1...	ὁ Θεός
4	26693	mrk	15	34	19	,29159,29160,	,38066,	1	[{'id': 29159, 'book_id': 'mrk', 'chapter': '1...	ὁ Θεός
...	...	...	...	...	...	...	...	...	...	...
1349	72801	act	19	37	11	,77510,	,101074,	0	[{'id': 77510, 'book_id': 'act', 'chapter': '1...	θεὸν
1350	105300	2th	2	4	6	,111284,	,147759,	0	[{'id': 111284, 'book_id': '2th', 'chapter': '...	θεὸν
1351	77911	act	28	6	27	,82744,	,108220,108221,	0	[{'id': 82744, 'book_id': 'act', 'chapter': '2...	θεόν
1352	92604	2co	4	4	3	,98291,	,130171,	0	[{'id': 98291, 'book_id': '2co', 'chapter': '4...	θεὸς
1353	97161	gal	4	8	11	,102952,	,136824,	0	[{'id': 102952, 'book_id': 'gal', 'chapter': '...	θεοῖς

1354 rows × 20 columns

In [4]:
```python
# find all alignments for this original word

# word = 'θεός' # found 69
# word = 'θεὸς' # found 239
word = 'θεοῦ'  # found 712
origAlignments = getDataFrameForOriginalWords(connection, word, searchLemma = False)
origAlignments
```

```
updating 'θεοῦ'
Replacing "God s" with "God's"
Replacing "the things that are God s" with "the things that are God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "of God s" with "of God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "of God s" with "of God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "with God s words" with "with God's words"
Replacing "God s" with "God's"
Replacing "of God s" with "of God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
Replacing "God s" with "God's"
```

Out[4]:

	id	book_id	chapter	verse	alignment_num	orig_lang_keys	target_lang_keys	origSpan	origWords	origWordsTxt
0	1047	mat	3	16	16	,1202,1203,	,1537,1538,	1	[{'id': 1202, 'book_id': 'mat', 'chapter': '3'...	τοῦ Θεοῦ
1	1099	mat	4	3	9	,1259,1260,	,1611,1612,	1	[{'id': 1259, 'book_id': 'mat', 'chapter': '4'...	τοῦ Θεοῦ
2	1125	mat	4	4	19	,1287,	,1645,1646,	0	[{'id': 1287, 'book_id': 'mat', 'chapter': '4'...	Θεοῦ
3	1149	mat	4	6	6	,1311,1312,	,1676,1677,	1	[{'id': 1311, 'book_id': 'mat', 'chapter': '4'...	τοῦ Θεοῦ
4	1534	mat	5	9	7	,1738,	,2227,2228,	0	[{'id': 1738, 'book_id': 'mat', 'chapter': '5'...	Θεοῦ
...	...	...	...	...	...	...	...	...	...	...
707	129469	rev	21	10	23	,137161,137162,	,182348,	1	[{'id': 137161, 'book_id': 'rev', 'chapter': '...	τοῦ Θεοῦ
708	129473	rev	21	11	3	,137166,137167,	,182353,182354,	1	[{'id': 137166, 'book_id': 'rev', 'chapter': '...	τοῦ Θεοῦ
709	129716	rev	21	23	16	,137431,137432,	,182674,182675,	1	[{'id': 137431, 'book_id': 'rev', 'chapter': '...	τοῦ Θεοῦ
710	129796	rev	22	1	13	,137522,137523,	,182789,182790,	1	[{'id': 137522, 'book_id': 'rev', 'chapter': '...	τοῦ Θεοῦ
711	129835	rev	22	3	9	,137568,137569,	,182847,182848,	1	[{'id': 137568, 'book_id': 'rev', 'chapter': '...	τοῦ Θεοῦ

712 rows × 20 columns

In [5]:
```python
db.describeAlignments(origAlignments)
```

```
Alignments description:
         origSpan  alignmentOrigWords  targetSpan  alignmentTargetWords  \
count  712.000000          712.000000  712.000000            712.000000
mean     0.623596            1.620787    0.599719              1.599719
std      0.501928            0.494146    0.515471              0.515471
min      0.000000            1.000000    0.000000              1.000000
25%      0.000000            1.000000    0.000000              1.000000
50%      1.000000            2.000000    1.000000              2.000000
75%      1.000000            2.000000    1.000000              2.000000
max      3.000000            3.000000    4.000000              5.000000

        frequency  matchCount  origWordsBetween  targetWordsBetween
count  712.000000  712.000000        712.000000               712.0
mean     0.241384  171.865169          0.002809                 0.0
std      0.147977  105.359575          0.074953                 0.0
min      0.001404    1.000000          0.000000                 0.0
25%      0.137640   98.000000          0.000000                 0.0
50%      0.176966  126.000000          0.000000                 0.0
75%      0.410112  292.000000          0.000000                 0.0
max      0.410112  292.000000          2.000000                 0.0
fields = ['origSpan', 'alignmentOrigWords', 'targetSpan', 'alignmentTargetWords', 'frequency', 'matchCount', 'origWordsBetween', 'targetWordsBetween']

Frequency of origSpan:
1    435
0    273
2      3
3      1
Name: origSpan, dtype: int64

Frequency of alignmentOrigWords:
2    436
1    273
3      3
Name: alignmentOrigWords, dtype: int64

Frequency of targetSpan:
1    417
0    291
2      3
4      1
Name: targetSpan, dtype: int64

Frequency of alignmentTargetWords:
2    417
1    291
3      3
5      1
Name: alignmentTargetWords, dtype: int64

Frequency of matchCount:
292    292
126    126
102    102
98      98
23      23
21      21
1       20
11      11
8        8
5        5
4        4
2        2
Name: matchCount, dtype: int64

Frequency of origWordsBetween:
0    711
2      1
Name: origWordsBetween, dtype: int64

Frequency of targetWordsBetween:
0    712
Name: targetWordsBetween, dtype: int64
```

Out[5]:

```
{'desc': {'origSpan': {'count': 712.0,
   'mean': 0.6235955056179775,
   'std': 0.5019282025771814,
   'min': 0.0,
   '25%': 0.0,
   '50%': 1.0,
   '75%': 1.0,
   'max': 3.0},
  'alignmentOrigWords': {'count': 712.0,
   'mean': 1.6207865168539326,
   'std': 0.4941462676438909,
   'min': 1.0,
   '25%': 1.0,
   '50%': 2.0,
   '75%': 2.0,
   'max': 3.0},
  'targetSpan': {'count': 712.0,
   'mean': 0.5997191011235955,
   'std': 0.5154708412840628,
   'min': 0.0,
   '25%': 0.0,
   '50%': 1.0,
   '75%': 1.0,
   'max': 4.0},
  'alignmentTargetWords': {'count': 712.0,
   'mean': 1.5997191011235956,
   'std': 0.5154708412840628,
   'min': 1.0,
   '25%': 1.0,
   '50%': 2.0,
   '75%': 2.0,
   'max': 5.0},
  'frequency': {'count': 712.0,
   'mean': 0.2413836636788284,
   'std': 0.14797693103697737,
   'min': 0.0014044943820224719,
   '25%': 0.13764044943820225,
   '50%': 0.17696629213483145,
   '75%': 0.4101123595505618,
   'max': 0.4101123595505618},
  'matchCount': {'count': 712.0,
   'mean': 171.86516853932585,
   'std': 105.35957489832789,
   'min': 1.0,
   '25%': 98.0,
   '50%': 126.0,
   '75%': 292.0,
   'max': 292.0},
  'origWordsBetween': {'count': 712.0,
   'mean': 0.0028089887640449437,
   'std': 0.07495316889958614,
   'min': 0.0,
   '25%': 0.0,
   '50%': 0.0,
   '75%': 0.0,
   'max': 2.0},
  'targetWordsBetween': {'count': 712.0,
   'mean': 0.0,
   'std': 0.0,
   'min': 0.0,
   '25%': 0.0,
   '50%': 0.0,
   '75%': 0.0,
   'max': 0.0}},
 'fields': {'origSpan': [435, 273, 3, 1],
  'alignmentOrigWords': [436, 273, 3],
  'targetSpan': [417, 291, 3, 1],
  'alignmentTargetWords': [417, 291, 3, 1],
  'matchCount': [292, 126, 102, 98, 23, 21, 20, 11, 8, 5, 4, 2],
  'origWordsBetween': [711, 1],
  'targetWordsBetween': [712]}}
```

Analysis of alignments for Θεοῦ in the en_ult:

Frequency of alignments:

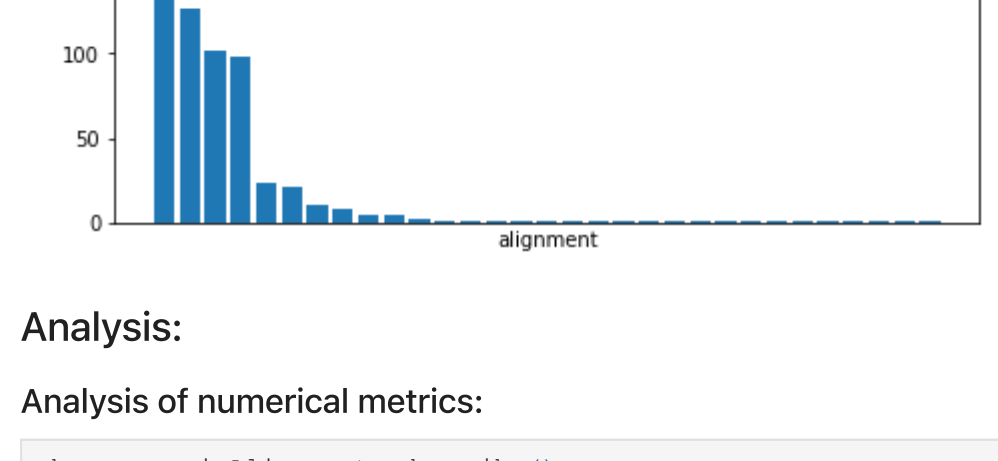
In [6]:
```python
frequency = origAlignments['alignmentTxt'].value_counts()
print(frequency)
```

```
τοῦ Θεοῦ = of God                             292
Θεοῦ = God                                    126
Θεοῦ = of God                                 102
τοῦ Θεοῦ = God                                 98
Θεοῦ = God's                                   23
τοῦ Θεοῦ = God's                               21
τοῦ Θεοῦ = of                                  11
Θεοῦ = of                                       8
Θεοῦ = of the                                   5
τοῦ Θεοῦ = from God                             4
τοῦ Θεοῦ = of God's                             2
λόγια Θεοῦ = with God's words                   1
μὲν Θεοῦ = of God                               1
Θεοῦ = of a god                                 1
ὅτι Θεοῦ = of God                               1
Θεοῦ = Father                                   1
Θεοῦ = by God                                   1
τοῦ Θεοῦ = against God                          1
Θεοῦ = for God                                  1
ἐνώπιον τοῦ Θεοῦ = and God                      1
τὴν τοῦ Θεοῦ = of God's                         1
Θεοῦ = in God                                   1
Θεοῦ = the Son                                  1
Θεοῦ = the                                      1
τοῦ Θεοῦ = about God                            1
τὰ τοῦ Θεοῦ = the things that are God's         1
Θεοῦ = because of God                           1
τοῦ Θεοῦ = for God                              1
Θεοῦ = godly                                    1
τοῦ Θεοῦ = godly                                1
τοῦ Θεοῦ = to God                               1
Name: alignmentTxt, dtype: int64
```

Notes:

- the left column is the specific alignment, and the right column is the number of times that specific alignment has been made so far in the NT.
- alignments that contain more words are more suspect.
- in future will combine "God s" to "God's" before doing analysis

In [7]:
```python
plot.plotFieldFrequency(frequency, "", 'alignment', title="Frequency of Alignments", xNumbers=False, xShowTi
```



Analysis:

Analysis of numerical metrics:

In [8]:
```python
descr = origAlignments.describe()
print(f"Alignments description:\n{descr}")
```

```
Alignments description:
         origSpan  alignmentOrigWords  targetSpan  alignmentTargetWords  \
count  712.000000          712.000000  712.000000            712.000000
mean     0.623596            1.620787    0.599719              1.599719
std      0.501928            0.494146    0.515471              0.515471
min      0.000000            1.000000    0.000000              1.000000
25%      0.000000            1.000000    0.000000              1.000000
50%      1.000000            2.000000    1.000000              2.000000
75%      1.000000            2.000000    1.000000              2.000000
max      3.000000            3.000000    4.000000              5.000000

        frequency  matchCount  origWordsBetween  targetWordsBetween
count  712.000000  712.000000        712.000000               712.0
mean     0.241384  171.865169          0.002809                 0.0
std      0.147977  105.359575          0.074953                 0.0
min      0.001404    1.000000          0.000000                 0.0
25%      0.137640   98.000000          0.000000                 0.0
50%      0.176966  126.000000          0.000000                 0.0
75%      0.410112  292.000000          0.000000                 0.0
max      0.410112  292.000000          2.000000                 0.0
```

Analysis of distance between first and last original language word:

In [9]:
```python
field = 'origSpan'
field_frequency = origAlignments[field].value_counts().sort_index()
print(f"\nFrequency of {field}:\n{field_frequency}")
```

```
Frequency of origSpan:
0    273
1    435
2      3
3      1
Name: origSpan, dtype: int64
```

Notes:

- this field is less useful because it includes aligned words, so added originalWordsBetween as more useful normalized metric (see analysis below).

Analysis of distance between first and last target language word:

In [10]:
```python
field = 'targetSpan'
field_frequency = origAlignments[field].value_counts().sort_index()
print(f"\nFrequency of {field}:\n{field_frequency}")
```

```
Frequency of targetSpan:
0    291
1    417
2      3
4      1
Name: targetSpan, dtype: int64
```

Notes:

- this field is also less useful because it includes the aligned words, so added targetWordsBetween as more useful normalized metric (see analysis below).

Analysis of original language word count:

In [11]:
```python
field = 'alignmentOrigWords'
field_frequency = origAlignments[field].value_counts().sort_index()
print(f"\nFrequency of {field}:\n{field_frequency}")
```

```
Frequency of alignmentOrigWords:
1    273
2    436
3      3
Name: alignmentOrigWords, dtype: int64
```

Notes:

- this field analysis suggests for θεός nearly all the original language word counts are tight. The word counts of 3 may need review. So we could probaby use that as a threshold for to flag for review.

Analysis of target language word count:

In [12]:
```python
field = 'alignmentTargetWords'
field_frequency = origAlignments[field].value_counts().sort_index()
print(f"\nFrequency of {field}:\n{field_frequency}")
```

```
Frequency of alignmentTargetWords:
1    291
2    417
3      3
5      1
Name: alignmentTargetWords, dtype: int64
```

Notes:

- this field analysis suggests that for θεός likely all the target language word counts are tight. The word count of 3 probably good for English (`of a god`). But still we could probaby use that as a threshold for to flag for review.

Analysis of count of extra unaligned words between aligned original language words:

In [13]:
```python
field = 'origWordsBetween'
field_frequency = origAlignments[field].value_counts().sort_index()
print(f"\nFrequency of {field}:\n{field_frequency}")
```

```
Frequency of origWordsBetween:
0    711
2      1
Name: origWordsBetween, dtype: int64
```

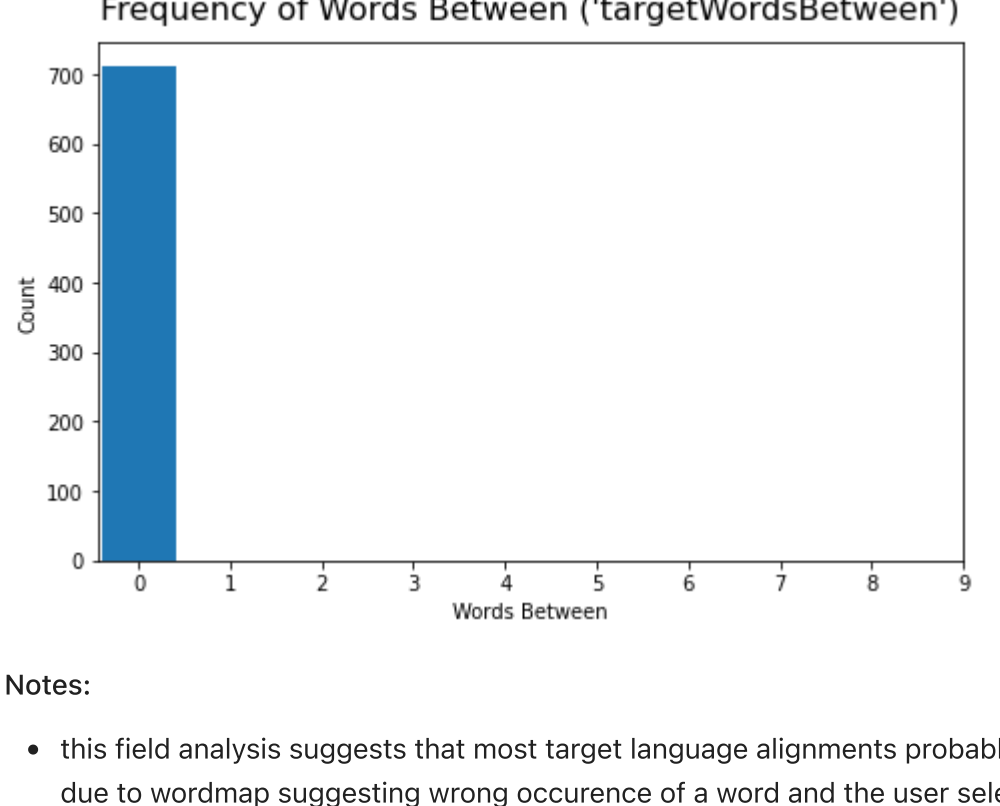
Notes:

- this field analysis suggests that most original language alignments probably good. Probably the cases of a word between (count > 0) aligned words should be reviewed.

Analysis of count of extra unaligned words between aligned target language words:

In [14]:
```python
field = 'targetWordsBetween'
field_frequency = origAlignments[field].value_counts().sort_index()
print(f"\nFrequency of {field}:\n{field_frequency}")
plot.plotFieldFrequency(field_frequency, field, f"Words Between", max=10)
```

```
Frequency of targetWordsBetween:
0    712
Name: targetWordsBetween, dtype: int64
```



Notes:

- this field analysis suggests that most target language alignments probably good. Large gaps between aligned words are likely due to wordmap suggesting wrong occurence of a word and the user selecting. Probably the cases of a word between (count > 0) aligned words should be reviewed.

In [13]:

In [ ]: