Plotting alignment data

In [1]:
```
%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np
import json
import csv
import math
import pandas as pd
import utils.db_utils as db
import utils.plot_utils as plot
import utils.file_utils as file
import config

############################################
# get configuration
cfg = config.getConfig() # configure values in config.js
############################################

targetLang = cfg['targetLang']
bibleType = cfg['targetBibleType']
tWordsTypeList = cfg['tWordsTypeList']
dbPath = cfg['dbPath']
trainingDataPath = cfg['trainingDataPath']
testamentStr = cfg['testamentStr']
baseDataPath = cfg['baseDataPath']
```

In [2]:
```
# get alignments for tW keyterms

minAlignments = 20
remove = ['ὁ', 'τὸ', 'τὰ', 'αὐτός', 'λέγω', 'ὡς', 'μέν', 'εἷς']

alignmentsForWord, filteredAlignmentsForWord = db.fetchAlignmentDataForAllTWordsCached(trainingDataPath, bib
print(f"Original Language Alignments: {len(filteredAlignmentsForWord)}")
```

```
Using cached Alignments
Unfiltered Alignments: 4368
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω in remove list
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ in remove list
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting τὸ in remove list
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting τὰ in remove list
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting ὁ alignments
getFilteredAlignmentsForWord - rejecting μέν alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting μέν in remove list
filtered alignments by original list count is 243
Size of filtered alignments by original ./data/en/ult/TrainingData/kt_en_ult_NT_alignments_by_orig_20.json i
s 7.628 MB
Size of filtered alignments by original ./data/en/ult/TrainingData/kt_en_ult_NT_alignments_by_orig_20.csv is
2.076 MB
Filtered Alignments: 243
Using cached Alignments
Unfiltered Alignments: 538
filtered alignments by original list count is 33
Size of filtered alignments by original ./data/en/ult/TrainingData/names_en_ult_NT_alignments_by_orig_20.jso
n is 1.298 MB
Size of filtered alignments by original ./data/en/ult/TrainingData/names_en_ult_NT_alignments_by_orig_20.csv
is 0.363 MB
Filtered Alignments: 33
Using cached Alignments
Unfiltered Alignments: 7380
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω in remove list
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting ὡς in remove list
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting εἷς alignments
getFilteredAlignmentsForWord - rejecting εἷς alignments
getFilteredAlignmentsForWord - rejecting εἷς in remove list
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting λέγω alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting αὐτός in remove list
getFilteredAlignmentsForWord - rejecting αὐτός alignments
getFilteredAlignmentsForWord - rejecting εἷς alignments
getFilteredAlignmentsForWord - rejecting εἷς alignments
getFilteredAlignmentsForWord - rejecting εἷς alignments
filtered alignments by original list count is 250
Size of filtered alignments by original ./data/en/ult/TrainingData/other_en_ult_NT_alignments_by_orig_20.jso
n is 6.275 MB
Size of filtered alignments by original ./data/en/ult/TrainingData/other_en_ult_NT_alignments_by_orig_20.csv
is 1.670 MB
Filtered Alignments: 250
Original Language Alignments: 429
```

Analysis of alignments for tWords in the en_ult:

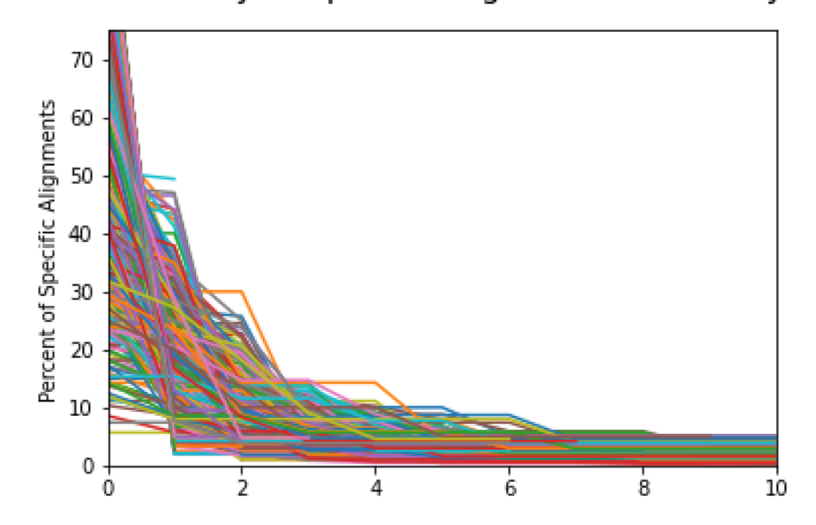
Frequency of alignments:

****Note that each line on the graphs below represents an alignment for a specific word. For example we have separate lines for 'Θεός', 'Θεὸς', or 'Θεοῦ' even though they have the same lemma. It made sense to group the alignments this way since aligners are likely to choose different target language words based on morphology of the word.*

In [3]:
```
frequenciesOfAlignments, stats = db.getFrequenciesOfFieldInAlignments(filteredAlignmentsForWord, 'alignmentT

print(f"Plotting of {len(filteredAlignmentsForWord)} tWord Alignments")
title = f"Plot of Variability of Specific Alignments in tW KeyTerms"
ylabel = "Percent of Specific Alignments"
xlimit = [0, 10]
ylimit = [0, 75]
outputTable = plot.plotFrequencies(frequenciesOfAlignments, title, ylabel, showXValues=False, xlimit=xlimit,
```

```
Plotting of 429 tWord Alignments
```

Plot of Variability of Specific Alignments in tW KeyTerms



In [4]:
```
print(f"Testing all tWords")

alignmentOrigWordsThreshold = 3
alignmentTargetWordsThreshold = 5
origWordsBetweenThreshold = 1
targetWordsBetweenThreshold = 1
alignmentFrequencyMinThreshold = 5

type_ = 'all_twords'

warningPath = f'{baseDataPath}/{type_}_{bibleType}_{testamentStr}_warnings.json'
warningData = db.generateWarnings(warningPath, type_, bibleType, filteredAlignmentsForWord, alignmentOrigWor
                                  alignmentTargetWordsThreshold, origWordsBetweenThreshold,
                                  targetWordsBetweenThreshold, alignmentFrequencyMinThreshold,
                                  tag=f'{minAlignments}')

print(f"Found {len(warningData)} alignments to check - min threshold {minAlignments}")

frequencyWarnings = warningData[warningData['frequencyWarning'].str.len() > 0]
print (f"\nFound {len(frequencyWarnings)} frequencyWarnings")
frequencyWarningsByOrigWords = frequencyWarnings['originalWord'].value_counts()
print (f"FrequencyWarnings by original word:")
frequencyWarningsByOrigWords
```

```
Testing all tWords
Found 1436 alignments to check - min threshold 20

Found 1201 frequencyWarnings
FrequencyWarnings by original word:
```

Out[4]:
```
ἐγένετο     65
Θεοῦ        50
Ἰησοῦς      37
λόγον       37
Χριστοῦ     33
            ..
Χριστός      1
Χριστῷ       1
νόμον        1
θρόνου       1
Χριστὸν      1
Name: originalWord, Length: 168, dtype: int64
```

In [5]:
```
frequencyWarningsByLemma = frequencyWarnings['lemma'].value_counts()
print (f"FrequencyWarnings by lemma:")
frequencyWarningsByLemma
```

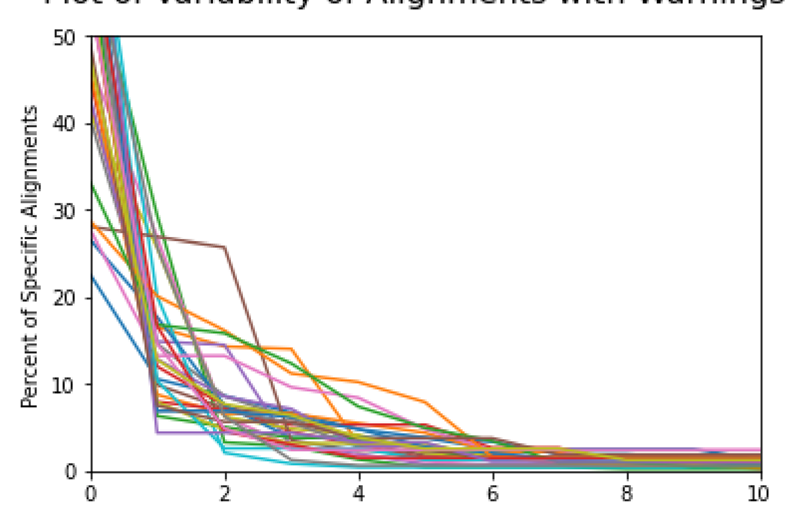
```
FrequencyWarnings by lemma:
```

Out[5]:
```
θεός          104
Ἰησοῦς         77
γίνομαι        65
λόγος          53
ὁράω           50
             ...
Σίμων           1
ἔρημος          1
θρόνος          1
δαιμόνιον       1
Ἱεροσόλυμα      1
Name: lemma, Length: 97, dtype: int64
```

In [6]:
```
minNumberOfWarnings = 10
origWordsWithWarnings = []
for key in frequencyWarningsByOrigWords.keys():
    count = frequencyWarningsByOrigWords[key]
    if count > minNumberOfWarnings:
        origWordsWithWarnings.append(key)

warningsAlignments = {}
for word in origWordsWithWarnings:
    warningsAlignments[word] = filteredAlignmentsForWord[word]
print(f"Found {len(origWordsWithWarnings)} original words with frequency warnings")

frequenciesOfAlignments, stats = db.getFrequenciesOfFieldInAlignments(warningsAlignments, 'alignmentText')

print(f"Plotting of {len(warningsAlignments)} tWord Alignments")
title = f"Plot of Variability of Alignments with Warnings"
ylabel = "Percent of Specific Alignments"
xlimit = [0, 10]
outputTable = plot.plotFrequencies(frequenciesOfAlignments, title, ylabel, showXValues=False, xlimit=xlimit)
```

```
Found 29 original words with frequency warnings
Plotting of 29 tWord Alignments
```

Plot of Variability of Alignments with Warnings



Analysis:

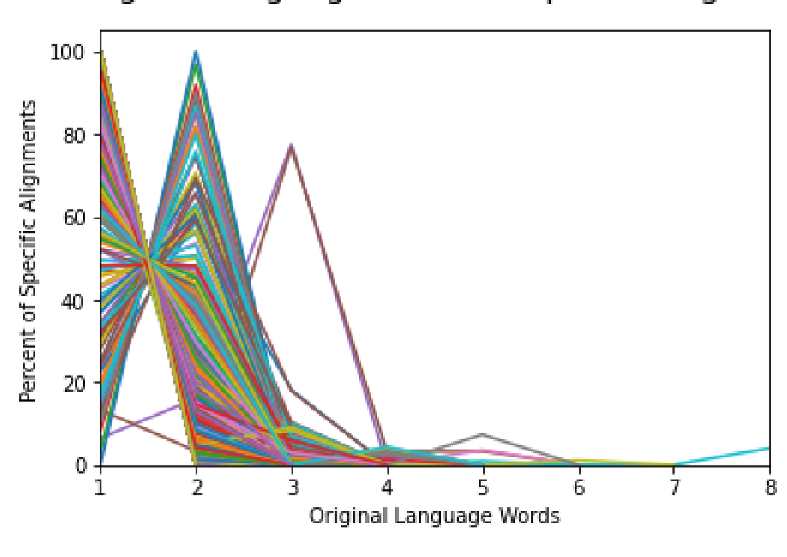
Analysis of numerical metrics:

Analysis of original language word count:

In [7]:
```
type_ = 'all'
field = 'origWordsCount'
field_frequencies, stats = db.getFrequenciesOfFieldInAlignments(filteredAlignmentsForWord, field, sortIndex
filledFrequencies = db.zeroFillFrequencies(field_frequencies)

print(f"Found {len(field_frequencies)} original language words for tW type {type_}")
title = f"Plot of number of Original Language Words in Specific Alignments in tW type {type_}"
ylabel = "Percent of Specific Alignments"
xlabel = "Original Language Words"
plot.plotXYdataDict(filledFrequencies, title, ylabel, xlabel, showXValues=True, xlimit=[1, 8])
```

```
Found 429 original language words for tW type all
```

Plot of number of Original Language Words in Specific Alignments in tW type all



Notes:

- this field analysis suggests that original word counts are tight - a threshold word count of 3 probably good for Greek to flag for review.

In [8]:
```
threshold = 4
abnormalAlignments = {}
for origWord in field_frequencies:
    frequency = field_frequencies[origWord]
    count = len(frequency)
    if count >= threshold:
        abnormalAlignments[origWord] = frequency

print(f"Out of {len(field_frequencies)}, found {len(abnormalAlignments)} original language words that have i
filledFrequencies = db.zeroFillFrequencies(abnormalAlignments)

title = f"Plot of abnormal number of Original Language Words in Specific Alignments in tW KeyTerms"
ylabel = "Percent of Specific Alignments"
xlabel = "Original Language Words"
plot.plotXYdataDict(filledFrequencies, title, ylabel, xlabel, showXValues=True, xlimit=[threshold, 10], ylim
```

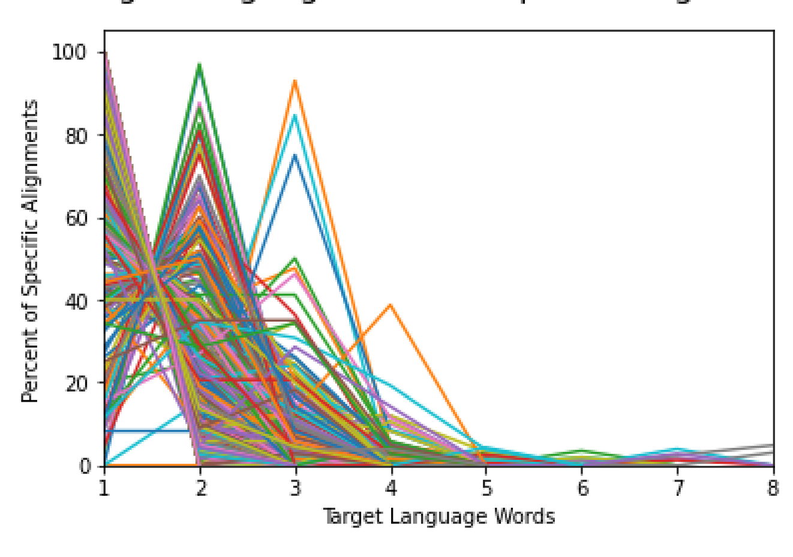
```
Out of 429, found 5 original language words that have instances with over 4 words
```

Plot of abnormal number of Original Language Words in Specific Alignments in tW KeyTerms

Analysis of target language word count:

In [9]:
```
field = 'targetWordsCount'
field_frequencies, stats = db.getFrequenciesOfFieldInAlignments(filteredAlignmentsForWord, field, sortIndex
filledFrequencies = db.zeroFillFrequencies(field_frequencies)

title = f"Plot of number of Target Language Words in Specific Alignments in tW KeyTerms"
ylabel = "Percent of Specific Alignments"
xlabel = "Target Language Words"
plot.plotXYdataDict(filledFrequencies, title, ylabel, xlabel, showXValues=True, xlimit=[1, 8])
```

Plot of number of Target Language Words in Specific Alignments in tW KeyTerms



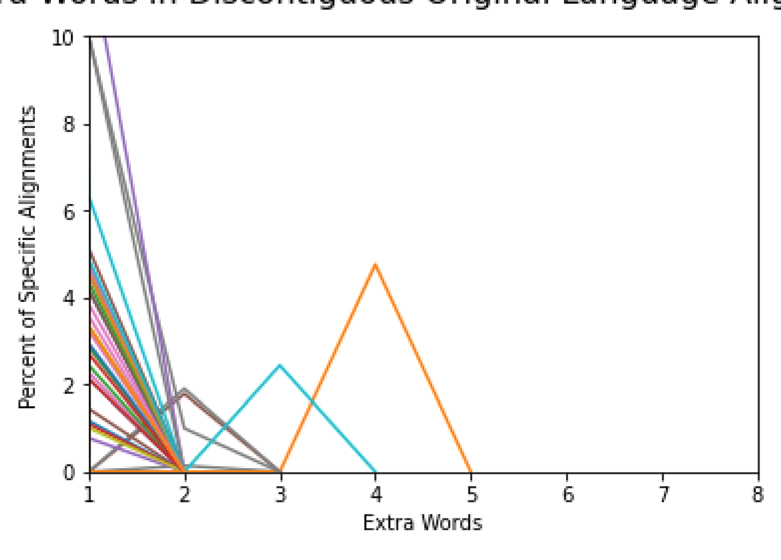
Notes:

- this field analysis suggests that a threshold word count of 3 probably good for English to flag for review.

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

In [10]:
```
field = 'origWordsBetween'
field_frequencies, stats = db.getFrequenciesOfFieldInAlignments(filteredAlignmentsForWord, field, sortIndex
filledFrequencies = db.zeroFillFrequencies(field_frequencies)

title = f"Plot of number of Extra Words in Discontiguous Original Language Alignments in tW KeyTerms"
ylabel = "Percent of Specific Alignments"
xlabel = "Extra Words"
plot.plotXYdataDict(filledFrequencies, title, ylabel, xlabel, showXValues=True, xlimit=[1, 8], ylimit=[0,10]
```

Plot of number of Extra Words in Discontiguous Original Language Alignments in tW KeyTerms



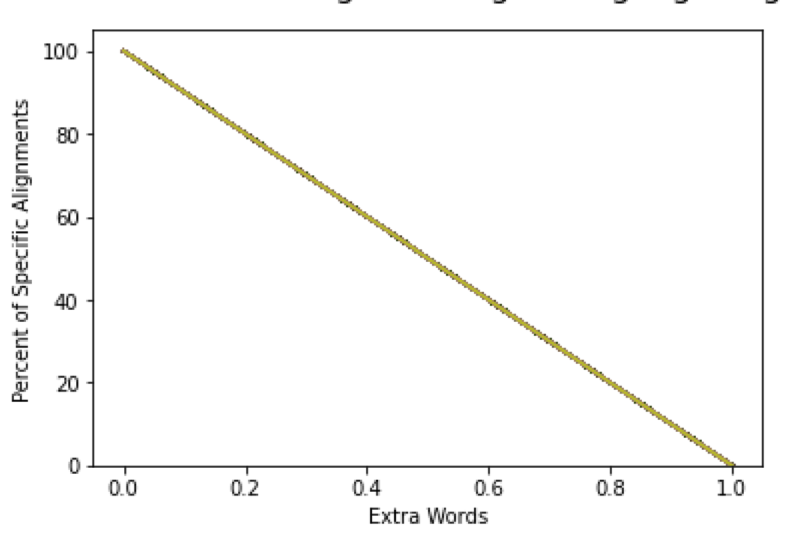
Notes:

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

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

In [11]:
```
field = 'targetWordsBetween'
field_frequencies, stats = db.getFrequenciesOfFieldInAlignments(filteredAlignmentsForWord, field, sortIndex
filledFrequencies = db.zeroFillFrequencies(field_frequencies)

title = f"Plot of number of Extra Words in Discontiguous Target Language Alignments in tW KeyTerms"
ylabel = "Percent of Specific Alignments"
xlabel = "Extra Words"
plot.plotXYdataDict(filledFrequencies, title, ylabel, xlabel, showXValues=True)
```

Plot of number of Extra Words in Discontiguous Target Language Alignments in tW KeyTerms



Notes:

- this field analysis suggests that most target language alignments are very tight.