What Is HUNSPELL ?

Hunspell Framework

Hunspell is a spell checker and morphological analyzer designed for languages with rich morphology and complex word compounding or character encoding and can use UTF-8 encoded dictionaries.

•Used by many popular applications.

- Open Office
- Mozila products - ThunderBird/ FireFox /SeaMonkey
- Opera 10
- Google Chrome
- Apple's Snow Leopard

Hunspell in Brief

Requires two files to define the language that it is spell checking.

- Dic file - a dictionary containing words for the language
- Aff file - an “affix” file that defines the meaning of special flags in the dictionary.

Hunspell in Brief Cont…

More on Affix files.

- SET - setting the character encodings of affixes and dictionary files
- TRY - sets the change characters for suggestions
- REP - replacement table for multiple character corrections
- PFX - defines prefixes
- SFX - suffix classes

Hunspell in Brief Cont…

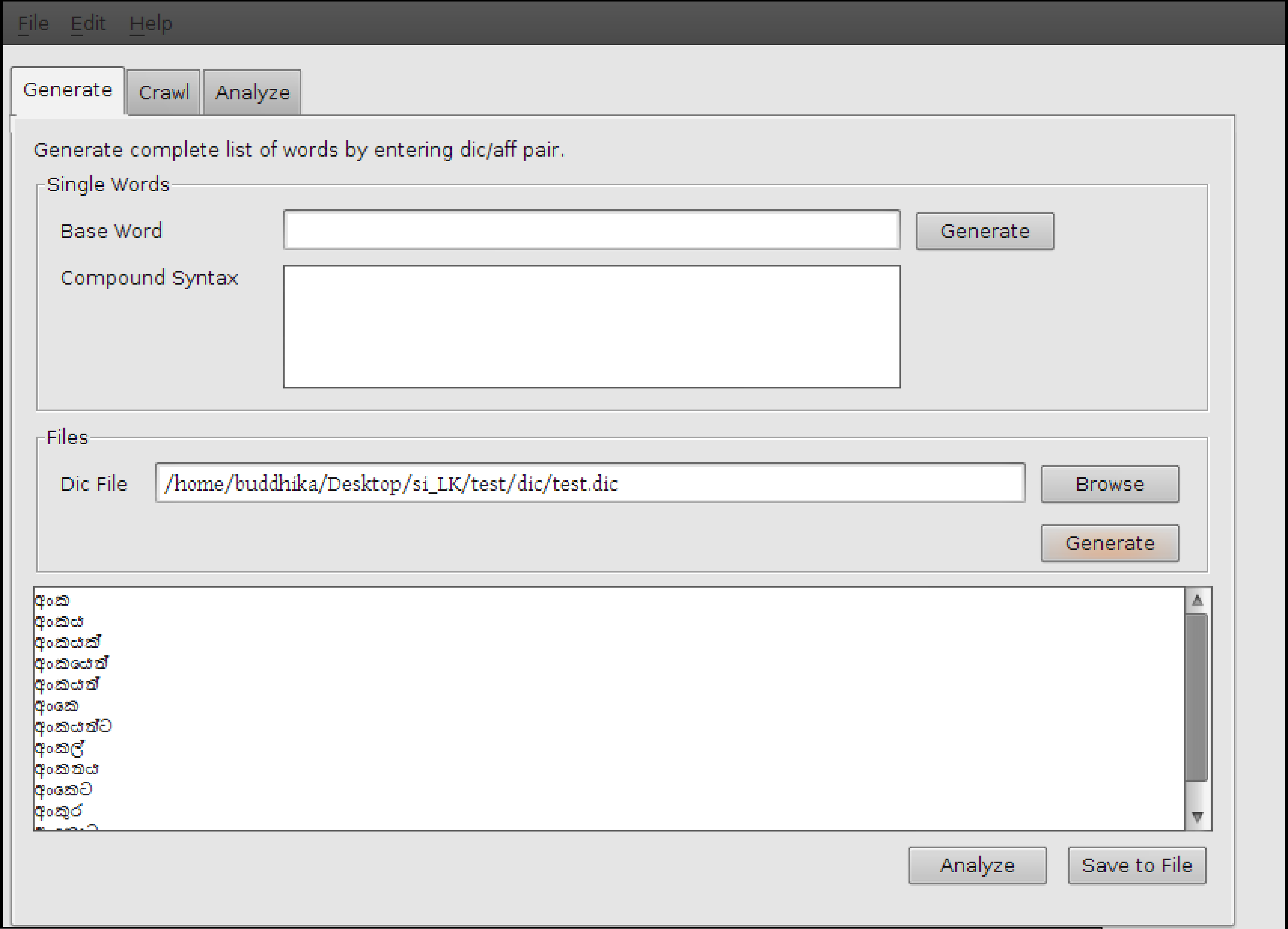
```
SET UTF-8
TRY esianrtolcdugmphbyfvkwzESIANRTOLCDUGMPHBYFVKWZ’
REP 2
REP f ph REP ph f
PFX     A     Y     1
PFX     A     0     re .
SFX     B     Y     2
SFX     B     0     ed   [^y]
SFX     B     y     ied   y
```

Aff File

```
hello
try/B
work/AB
```

Dic File

“hello", “try", “tried", “work", “worked", “rework", “reworked".



Generate complete list of words by entering dic/aff pair.

Single Words

Base Word:

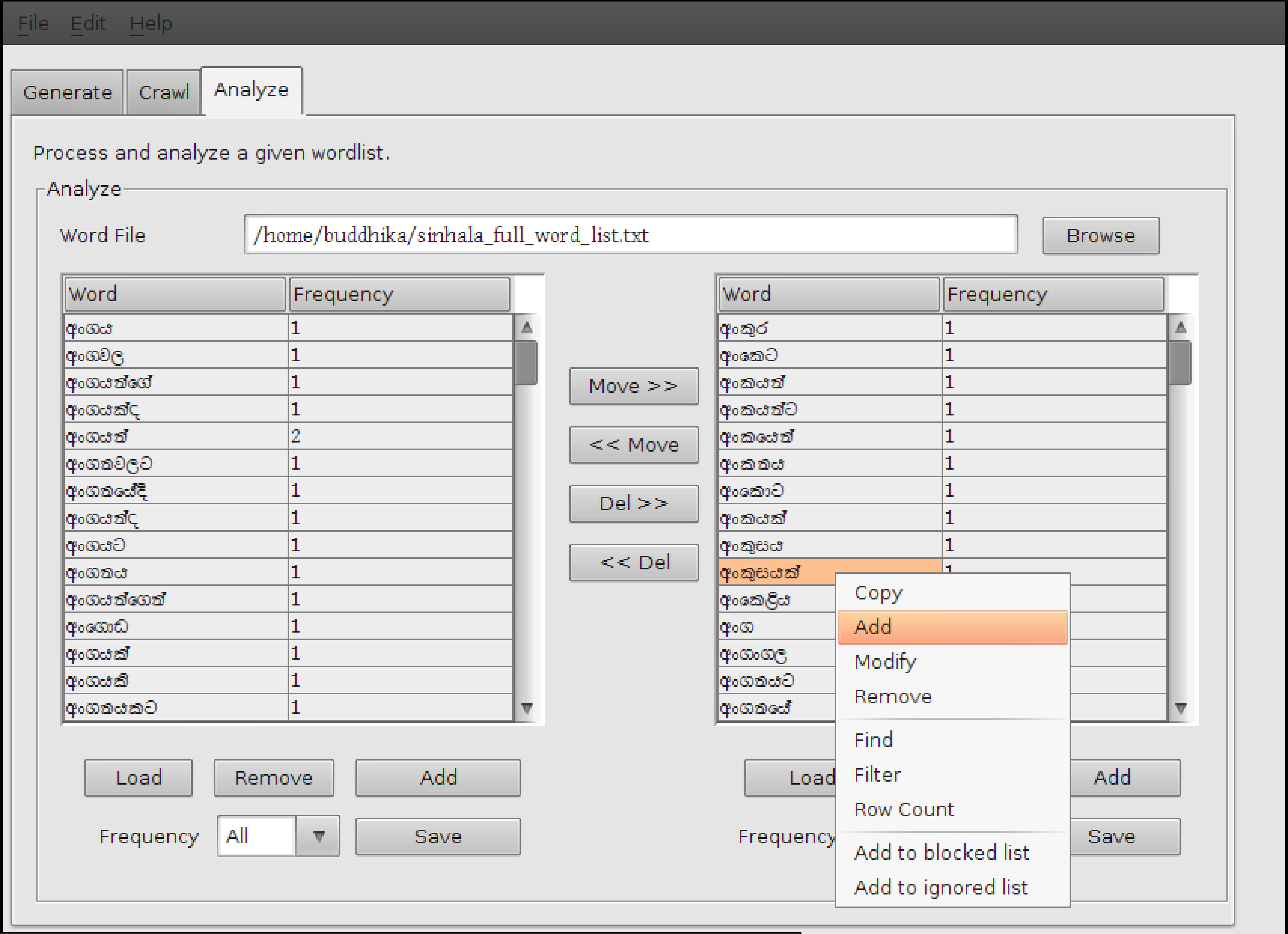
Compound Syntax:

Files

Dic File: /home/buddhika/Desktop/si_LK/test/dic/test.dic

```
අංක
අංකය
අංකයක්
අංකයෙක්
අංකයත්
අංකෙ
අංකයත්ට
අංකල්
අංකකය
අංකෙට
අංකුර
```

1.1 - Finding existing sources and bringing them to a common platform



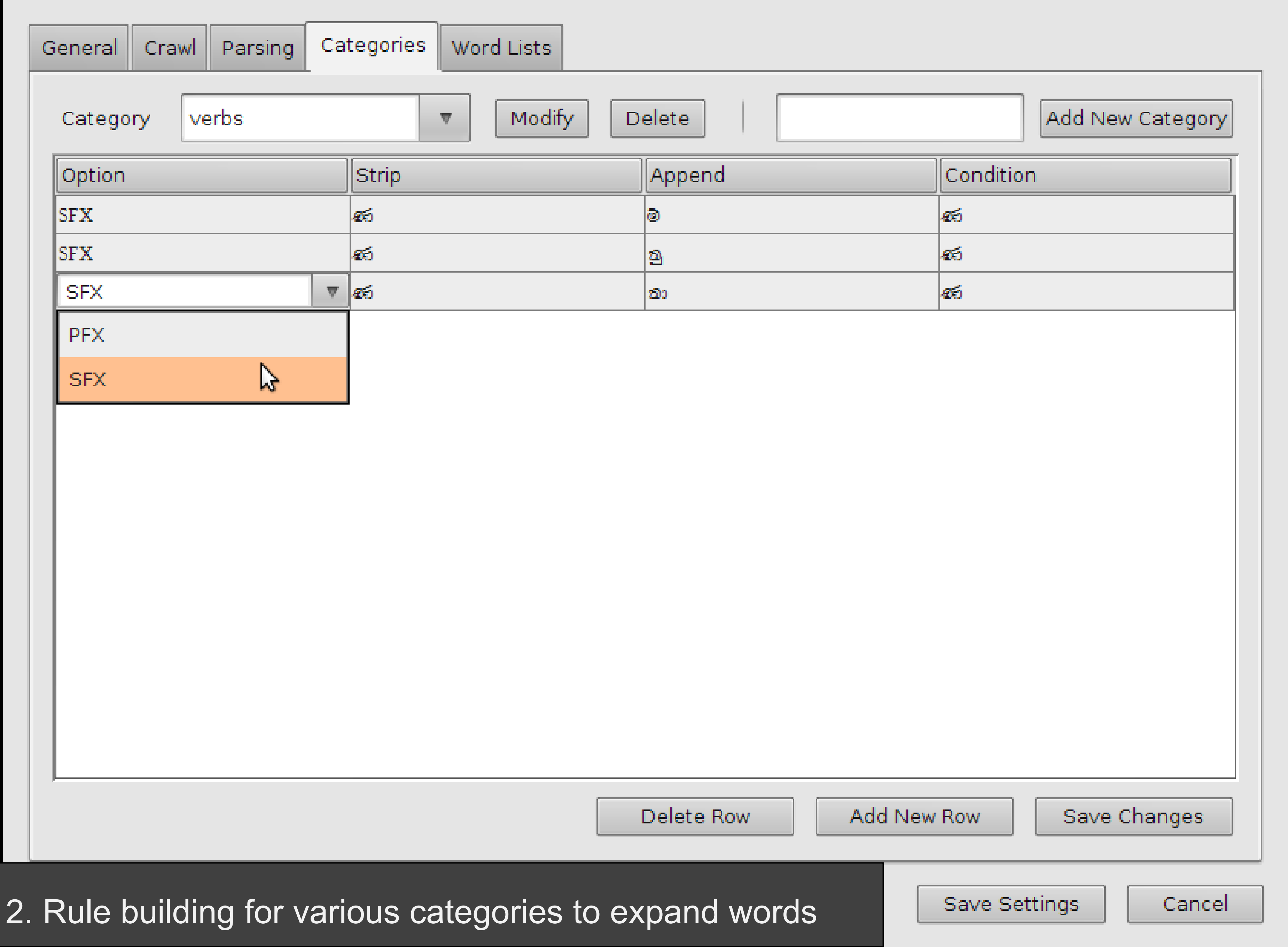
Process and analyze a given wordlist.

Word File: /home/buddhika/sinhala_full_word_list.txt

Word	Frequency
අංගය	1
අංගවල	1
අංගයන්ගේ	1
අංගයන්ද	1
අංගයන්	2
අංගනවලට	1
අංගනයේදී	1
අංගයන්ද	1
අංගයට	1
අංගනය	1
අංගයන්ගෙන්	1
අංගොඩ	1
අංගයක්	1
අංගයකි	1
අංගනයකට	1

Word	Frequency
අංකුර	1
අංකෙට	1
අංකයන්	1
අංකයන්ට	1
අංකයෙන්	1
අංකනය	1
අංකොට	1
අංකයක්	1
අංකුසය	1
අංකුසයක්	1
අංකෙලිය	
අංග	
අංගංගලෙ	
අංගනයට	
අංගනයේ	

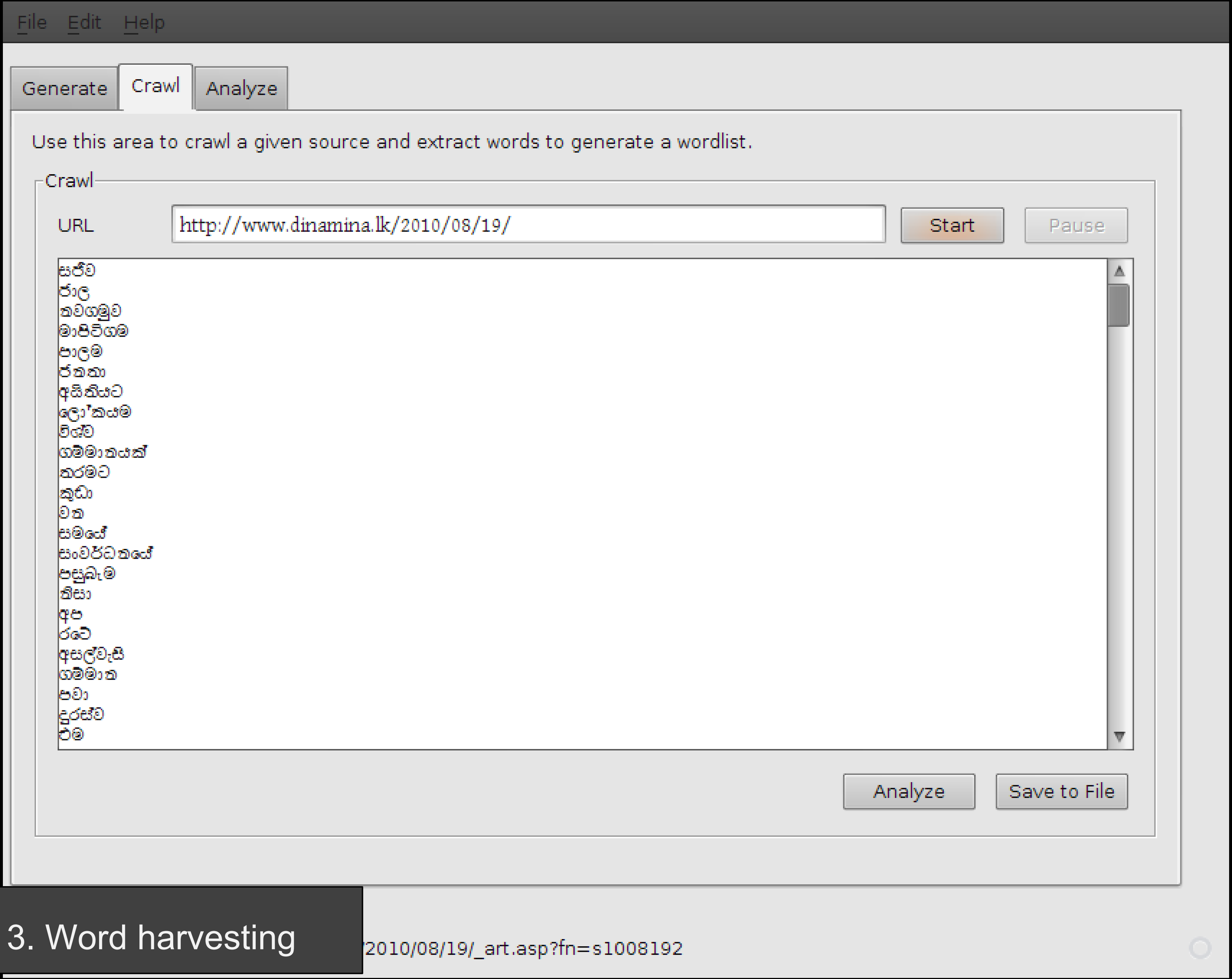
1.2 – Manually processing and error correcting



Category: verbs

Option	Strip	Append	Condition
SFX	න්	මි	න්
SFX	න්	මු	න්
SFX	න්	නා	න්

2. Rule building for various categories to expand words



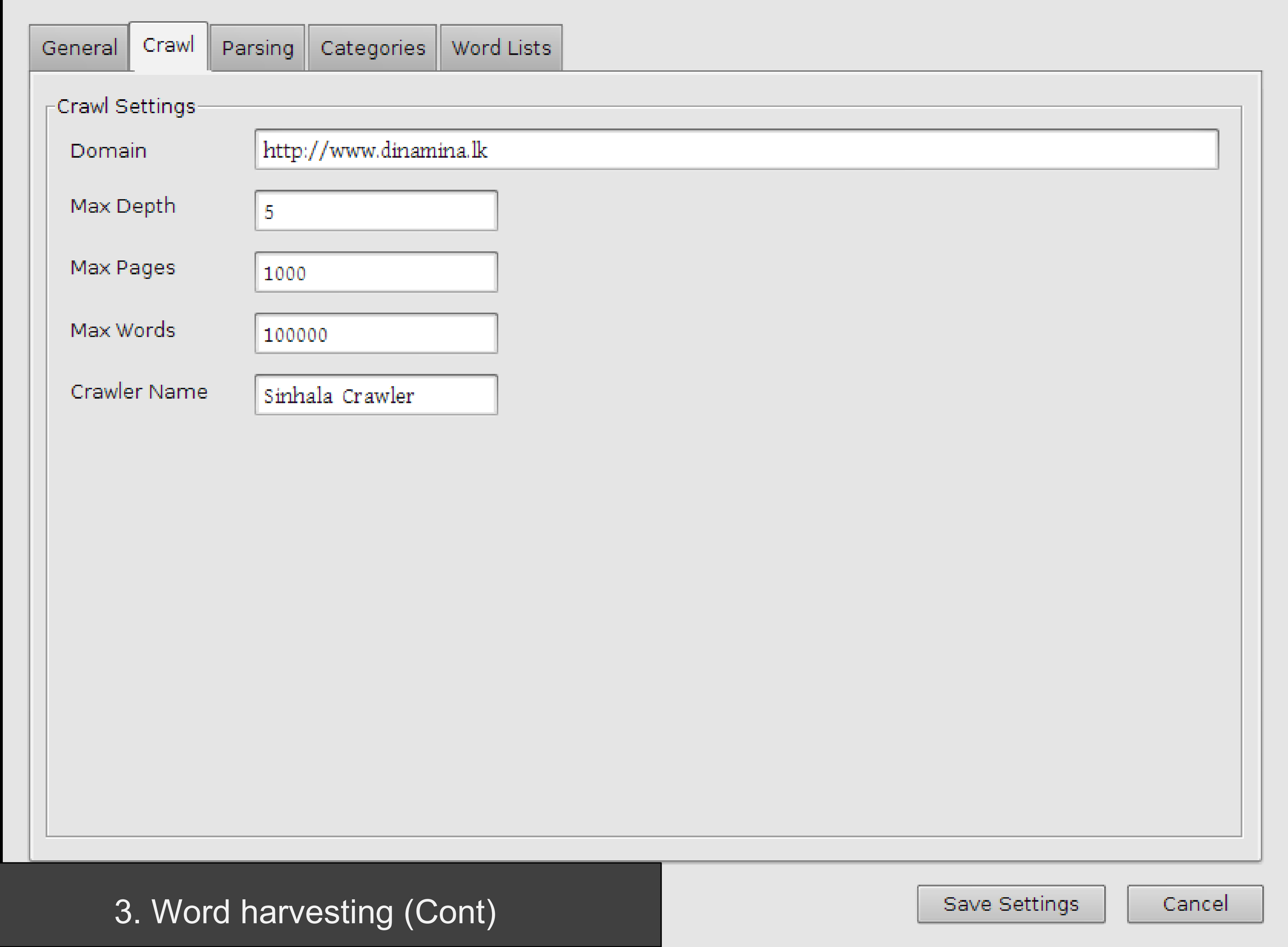
Use this area to crawl a given source and extract words to generate a wordlist.

URL: http://www.dinamina.lk/2010/08/19/

```
සජීව
ජාල
තවගමුව
මාපිටිගම
පාලම
ජනතා
අසිතියට
ලෝකයම
විශ්ව
ගම්මානයක්
තරමට
කුඩා
වන
සමයේ
සංවර්ධනයේ
පසුබෑම
නිසා
අප
රටේ
අසල්වැසි
ගම්මාන
පවා
දුරස්ව
එම
```

/2010/08/19/_art.asp?fn=s1008192

3. Word harvesting



Crawl Settings

Domain: http://www.dinamina.lk

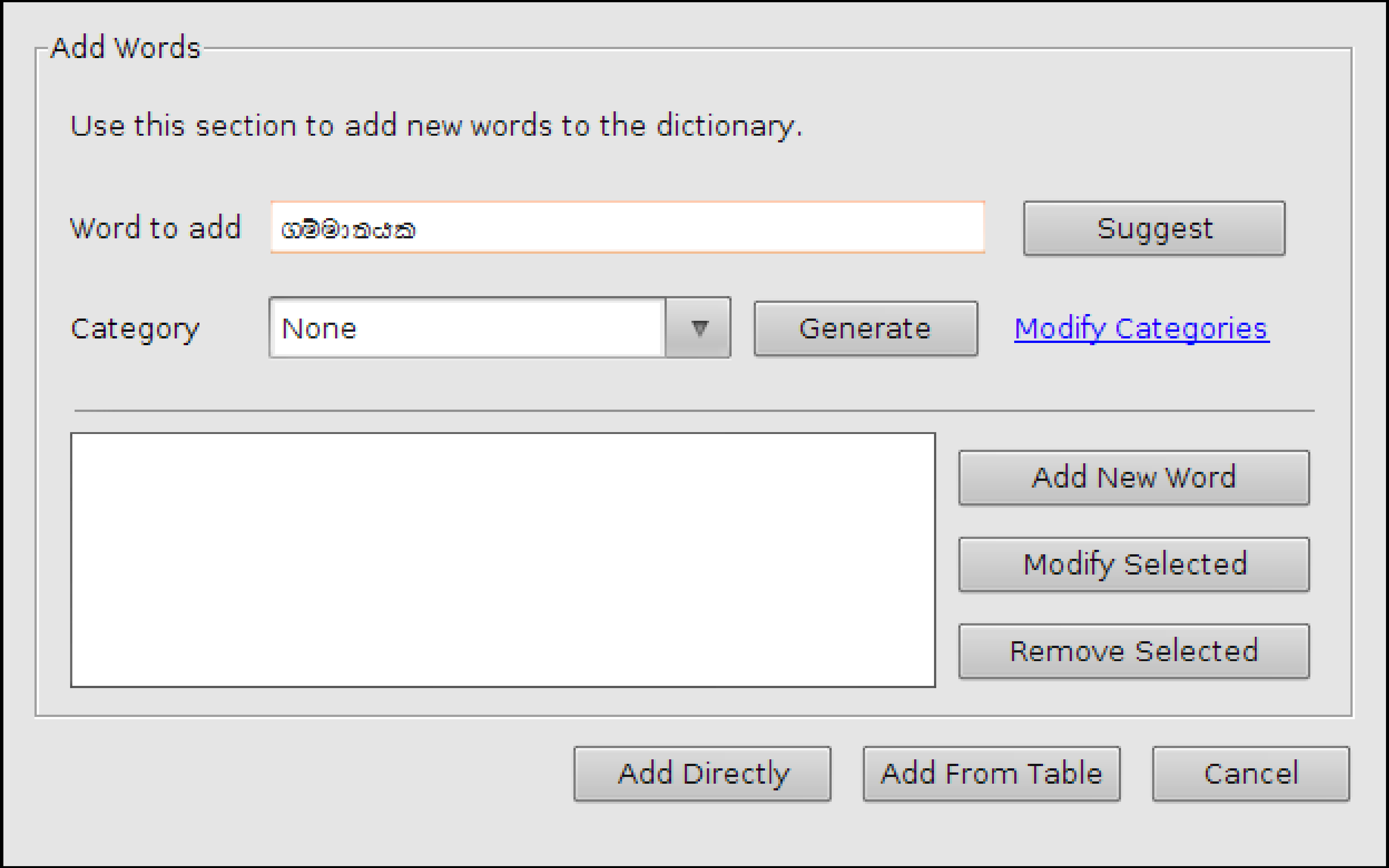
Max Depth: 5

Max Pages: 1000

Max Words: 100000

Crawler Name: Sinhala Crawler

3. Word harvesting (Cont)



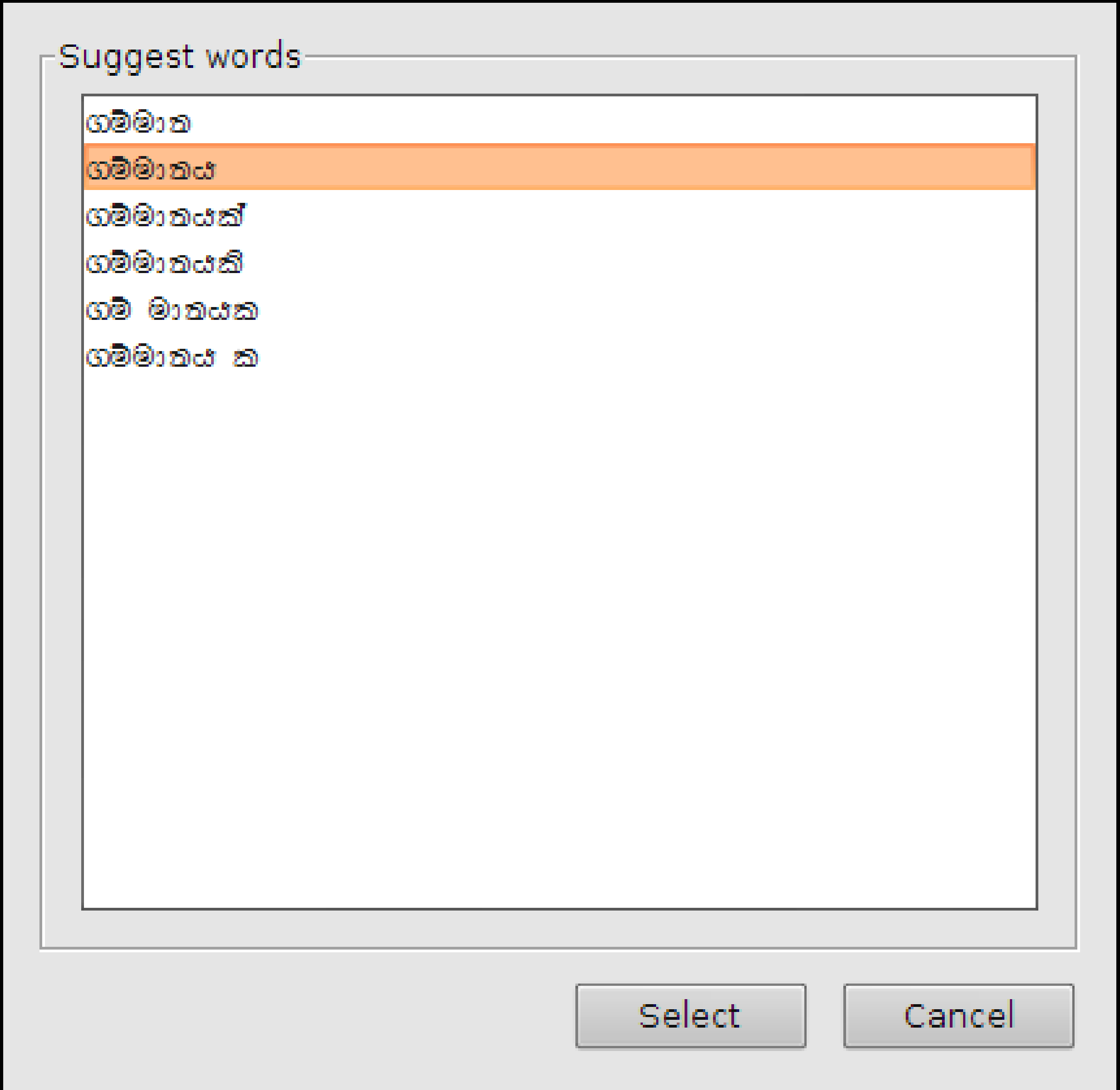
Add Words

Use this section to add new words to the dictionary.

Word to add: ගම්මානයක

Category: None

4. Find Base words



Suggest words

ගම්මාන

ගම්මානය

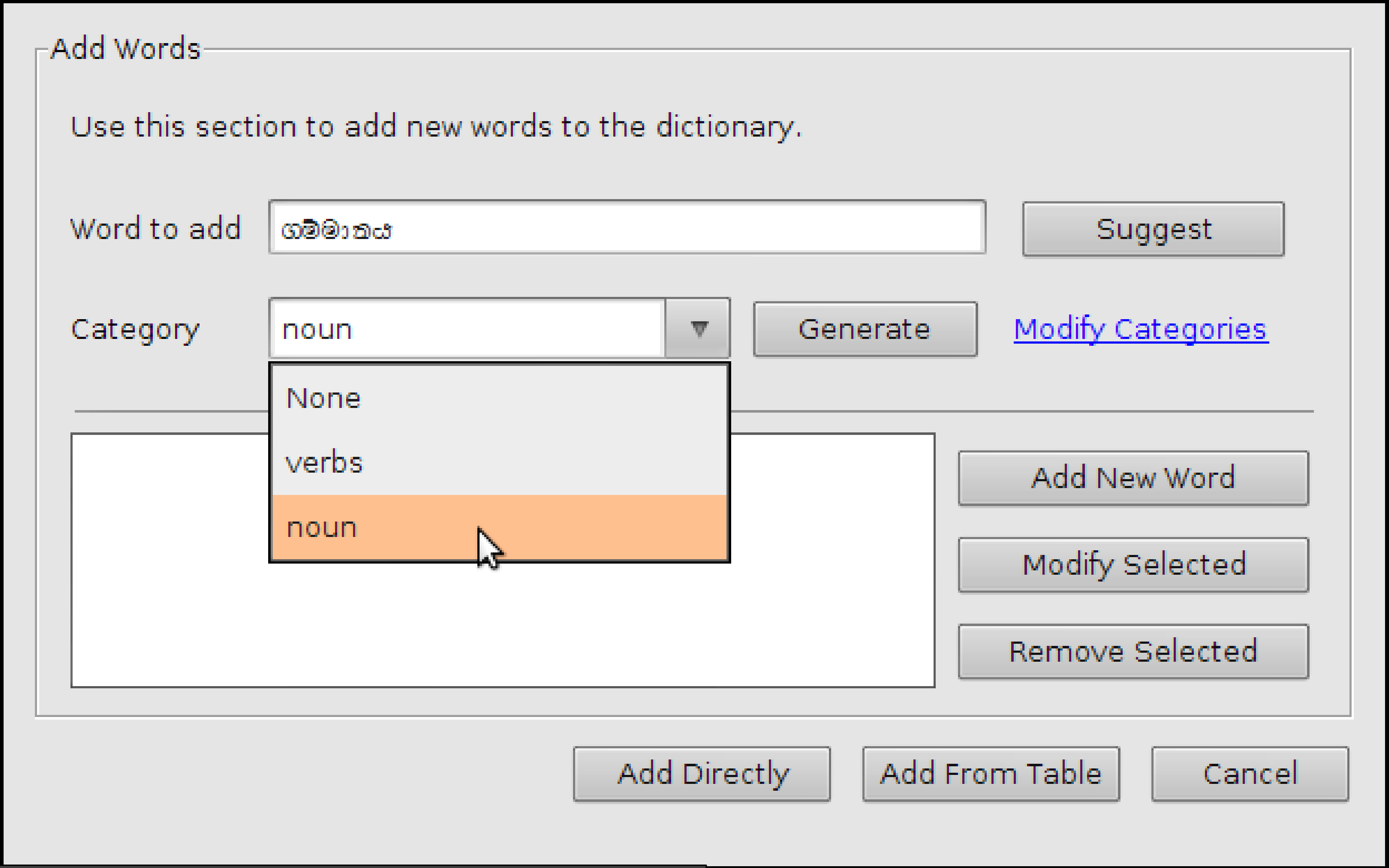
ගම්මානයක්

ගම්මානයකි

ගම් මානයක

ගම්මානය ක

4. Find Base words (Cont)

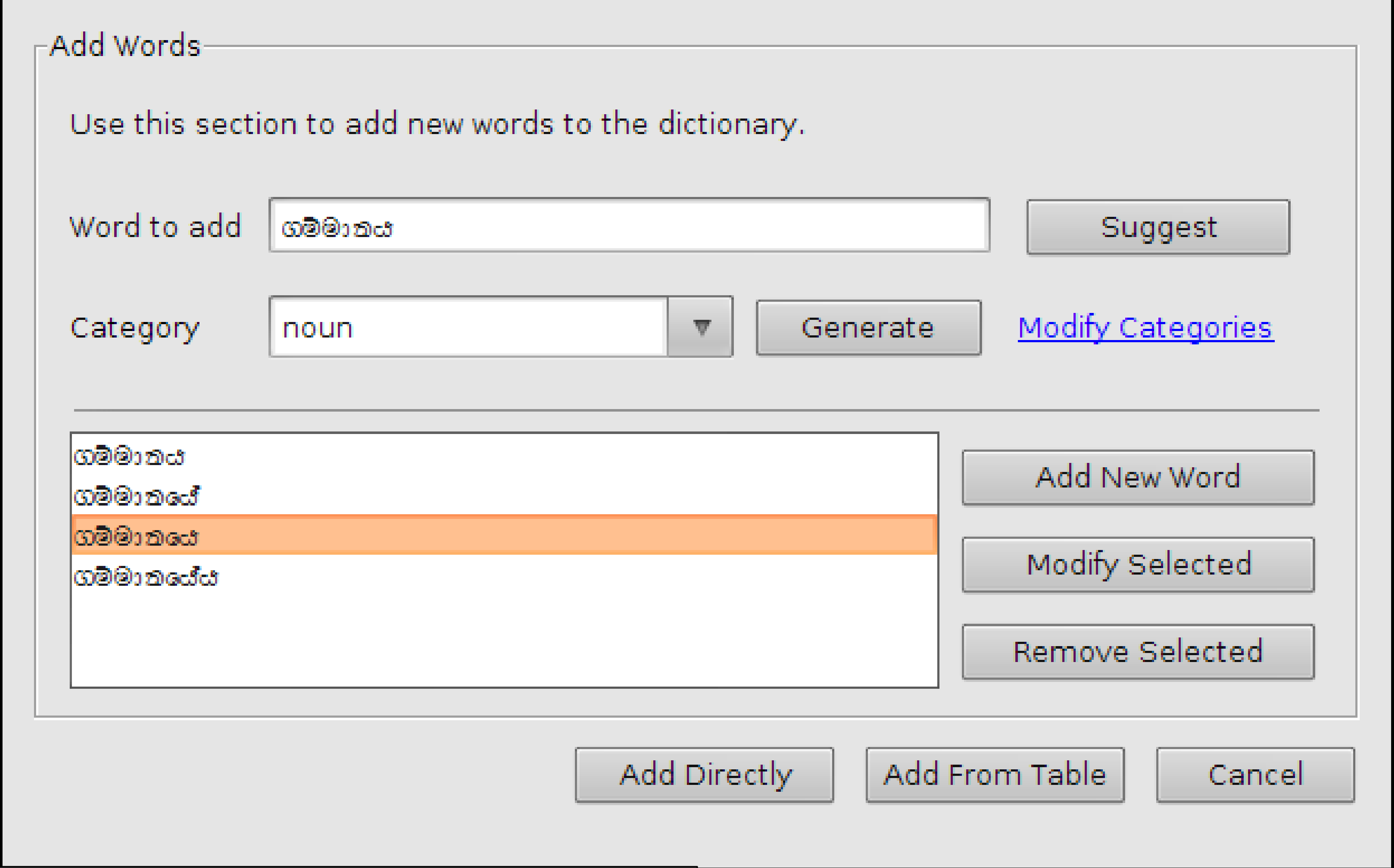


Use this section to add new words to the dictionary.

Word to add: ගම්මානය

Category: noun

5. Expanding words



Add Words

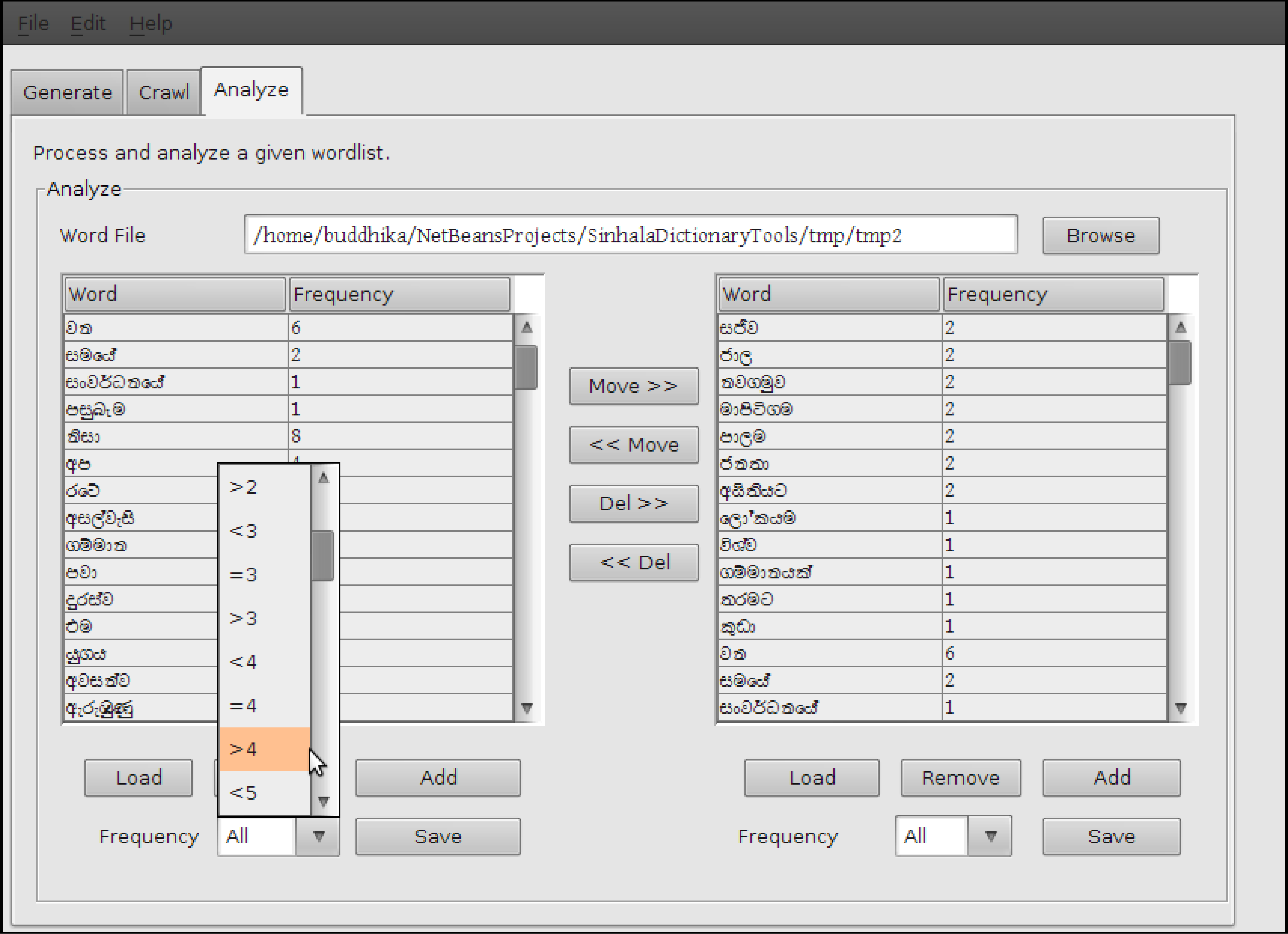
Use this section to add new words to the dictionary.

Word to add: ගම්මානය

Category: noun

ගම්මානය
ගම්මානයේ
ගම්මානයෙ
ගම්මානයේය

5. Expanding words (cont)



Process and analyze a given wordlist.

Analyze

Word File: /home/buddhika/NetBeansProjects/SinhalaDictionaryTools/tmp/tmp2

Word	Frequency
වත	6
සමයේ	2
සංවර්ධනයේ	1
පසුබැම	1
තිසා	8
අප	4
රටේ	
අසල්වැසි	
ගම්මාන	
පවා	
දුරස්ව	
එම	
යුගය	
අවසන්ව	
ඇරැඹුණු	

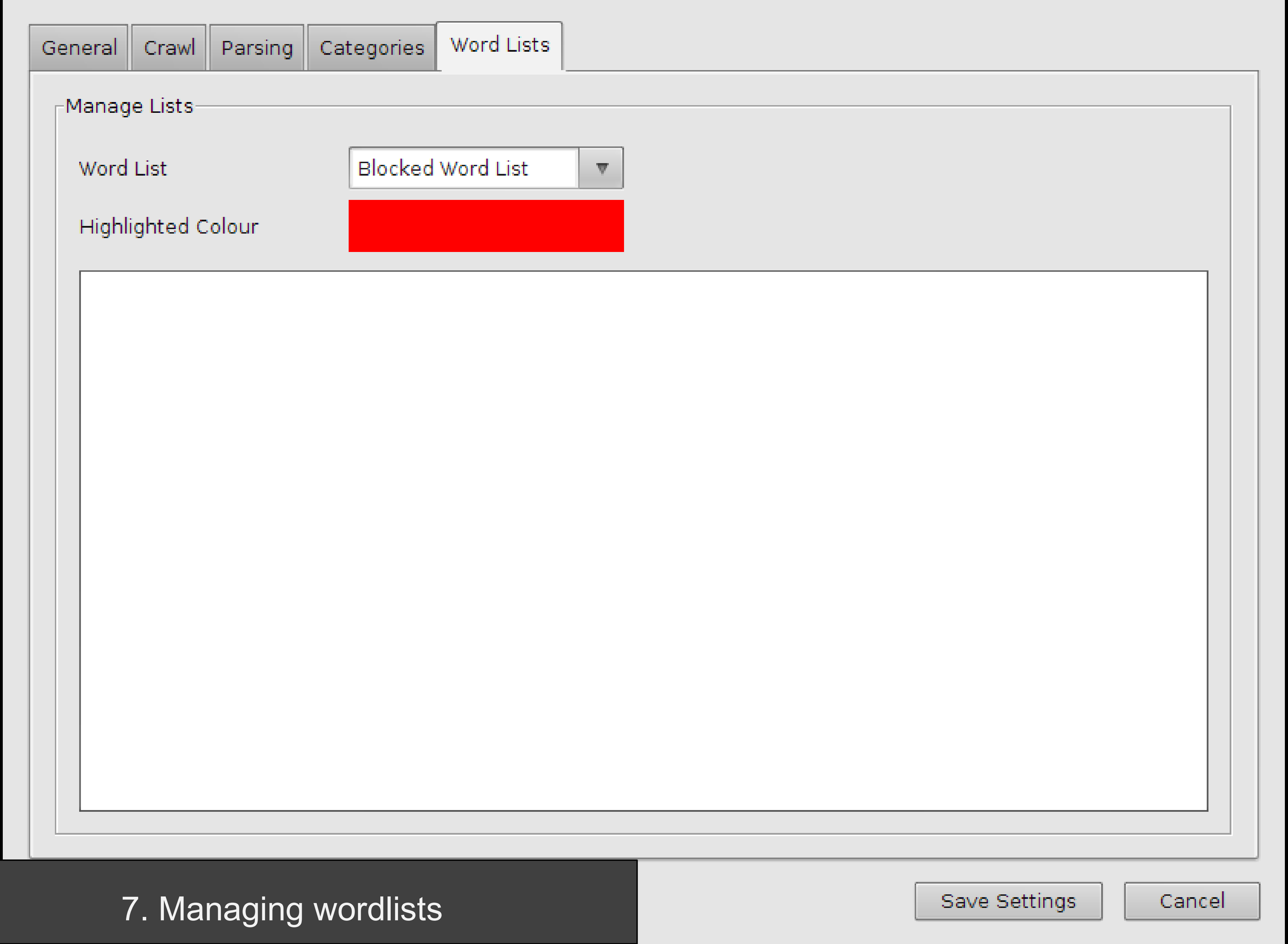
>2
<3
=3
>3
<4
=4
>4
<5

Frequency: All

Word	Frequency
සජීව	2
ජාල	2
තවගමුව	2
මාපිටිගම	2
පාලම	2
ජනතා	2
අයිතියට	2
ලො'කයම	1
විශ්ව	1
ගම්මානයක්	1
තරමට	1
කුඩා	1
වත	6
සමයේ	2
සංවර්ධනයේ	1

Frequency: All

6. Ensuring correctness through word harvesting and frequency matching



Word List: Blocked Word List

Highlighted Colour: 🟥

7. Managing wordlists