Stemmer Builder suite Manual

This is the manual for the Stemmer Builder suite. It is a suite of software programs to facilitate the experts to build stemmers (for some target language) without writing any line of program code. The suite is created by professor Nikitas N. Karanikolas (University of West Attica, Dept. of Informatics and Computer Enfineering). It has been used (till now) for the following languages: Albanian, Polish and Serbian.

Documentation

It is based on the ideas presented in the following papers:

B34
Nikitas N. Karanikolas, A methodology for building simple but robust stemmers without language knowledge: Overview, data model and ranking algorithm. CompSysTech'2013: 14th International Conference on Computer Systems and Technologies, June 2013, Ruse, Bulgaria. ACM ICPS, doi:10.1145/2516775.2516783.

A13
Nikitas N. Karanikolas. Supervised learning for building stemmers. Journal of Information Science, Vol. 41 (3), pp. 315-328, 2015, doi:10.1177/0165551515572528.

A10
Nikitas N. Karanikolas. A methodology for building simple but robust stemmers without language knowledge: Stemmer configuration. Procedia, Social and Behavioral Sciences, vol. 147, pp. 370-375, doi:10.1016/j.sbspro.2014.07.113.

See also the personal web page of author: http://users.uniwa.gr/nnk/papers/paper_index.htm

The suite uses:

The 2nd **Builder** (presented in A13) for forming (configuring) a trial stemmer

The **Stemmer Evaluator** (presented in B34) for evaluating a trial stemmer

A deprecated builder (1st Builder) is presented in A10. It is not any more used.

The suite has some other new facilities:

Language Manager

Experts Manager

Stem Editor

Code Builder

The whole suite is composed by:

- Language Manager
- Experts Manager
- Stem Editor

- 2nd Builder for forming (configuring) trial stemmers
- Stemmer Evaluator for evaluating trial stemmers
- Code Builder

and accompanied by

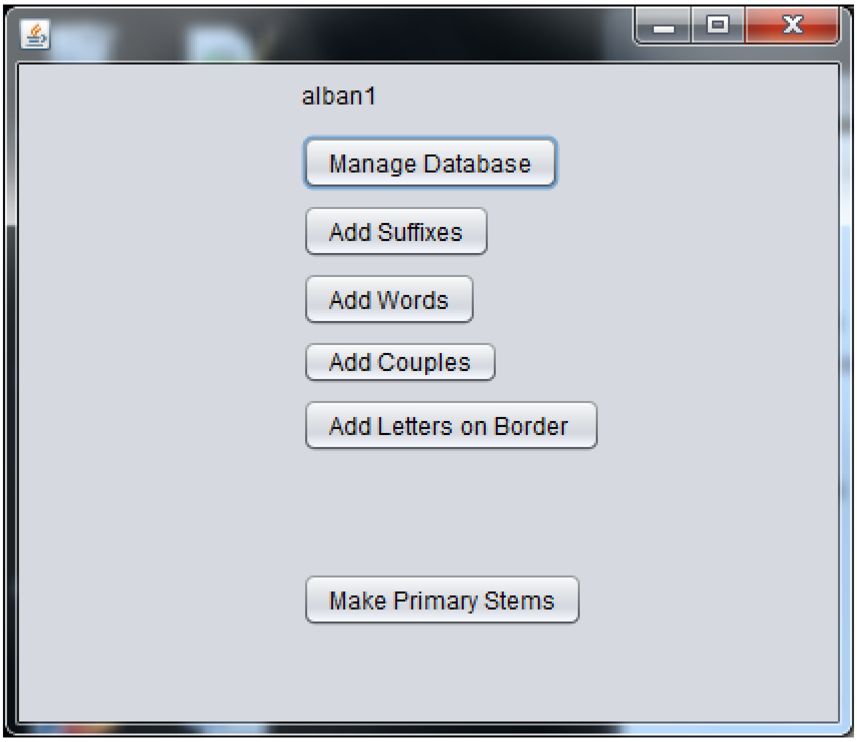
- Usage of java compiler (javac) to compile the source code and produce the executable stemmer
- Usage of java run time (java) for running the executable stemmer

Language Manager

Language manager is a configuration tool which is used to setup another language or another trial for some language. With language manager we can:

- Setup a new configuration for working with some language. This creates a new folder and a new database for holding (persisting) any data (details) for the language. The Language Manager is able to handle various alternatives for the same language, by using alternative names. For example, we can define two alternative setups for the Albanian language by using the names Alban1 and Alban2.
- Define the suffixes list which will be used for stemming
- Define the set of words that will be used for building primary stemmer, expressing arguments againts or for verifying the primary stemmers results and for evaluating stemmers (primary and other trial stemmers).
- Define the couples (digraph vowels, digraph consonants and diphongs) that the language might have.
- Define letters with special usage.
- Automate the process of primary stemmer building.

The main interface of Language manager is depicted in the following screenshot:



To invoke Language Manager:

C:\stemSuite\bin> **java –jar languageManager.jar**

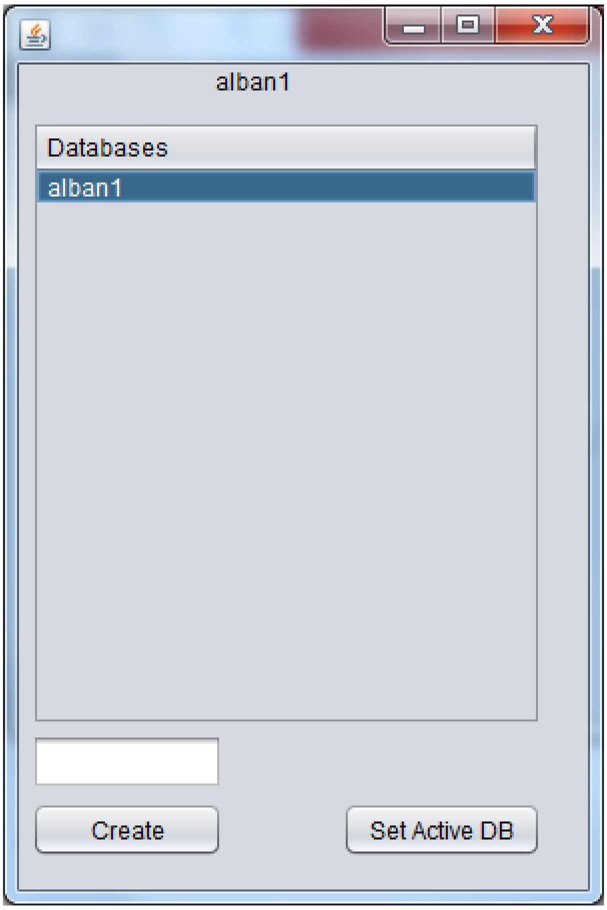
(assume that jar files are in folder bin under the basic stemSuite folder)

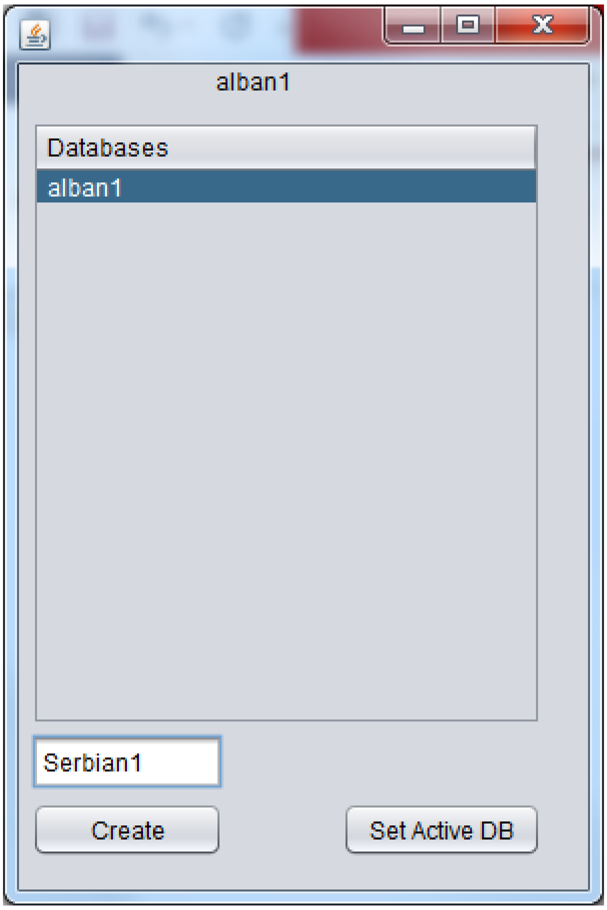
Language Manager / Manage Database

The “Manage Database” button brings in front the following dialog where the user can activate one of the existing database (language) setups or create a new database (a new language) setup.

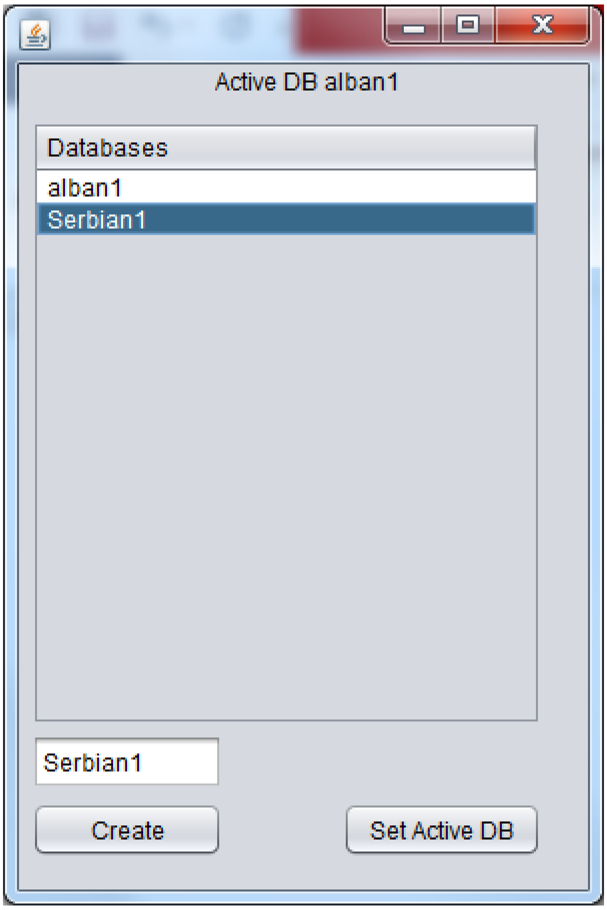
In order to create a language setup, fill the empty textbox with the name you prefer and press the button “create”.

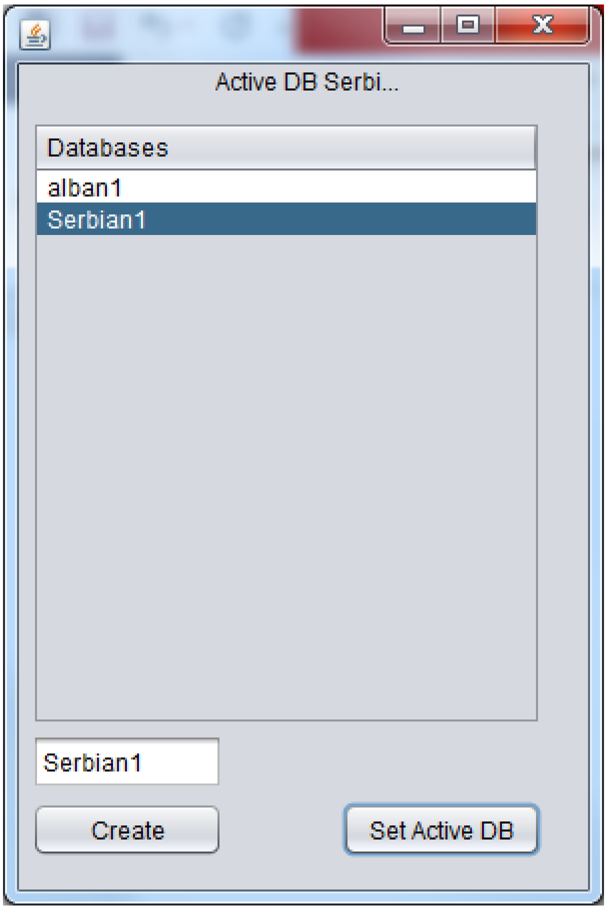
In order to select a language setup, select one of the listed in the Databases list box and press the button “Set Active Database”.


alban1
Databases
alban1
Create
Set Active DB


alban1
Databases
alban1
Create
Set Active DB

Serbian1





Language Manager / Add Suffixes

The “Add Suffixes” button of the main Language Manager dialog brings in front the following dialog. In this (the following) dialog the user can define which are the suffixes used for building stemmers. With the suffixes dialog, the user type a suffix in the empty textfield and presses the button “Add New Suffix”. The new Suffix is appended in the suffixes list box. The user repeat the sequence of steps, once per suffix. When the user finishes with the suffixes, he/she has to press the button “Save” to persist the list of suffixes.

The user can build the list in more than one usages of the dialog. In the first usage the user defines some suffixes and then use button “Save” to persist. In the following usages, the user has to load the existing suffixes (use button “load”), next define some more suffixes and next persist the updated list of suffixes (by pressing the button “Save”). After a number of dialog usages the list of suffixes is completed.

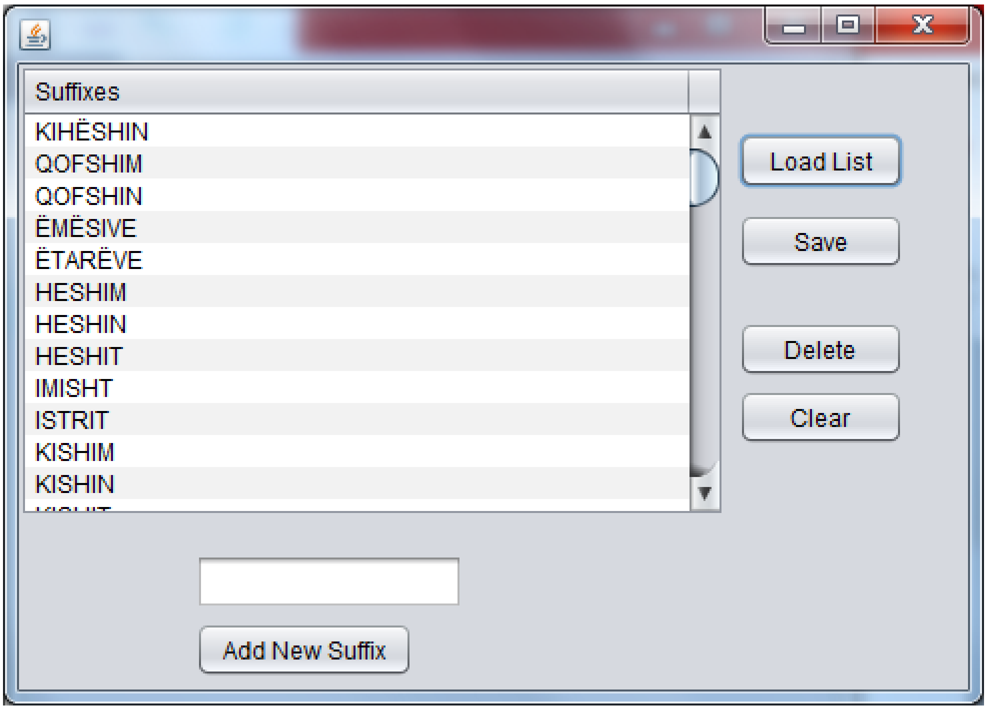
Button “Delete” can be used to remove the selected suffix. Button “Clear” removes all the entries in the suffix list.

The list of suffixes is persisted in a file with path:

C:\stemSuite\<language>\suffixlist.txt

For example:

C:\stemSuite\Alban1\suffixlist.txt



Suffixes
KIHËSHIN
QOFSHIM
QOFSHIN
ËMËSIVE
ËTARËVE
HESHIM
HESHIN
HESHIT
IMISHT
ISTRIT
KISHIM
KISHIN

Language Manager / Add Words

The "Add Words" button of the main Language Manager dialog brings in front the following dialog. In this (the following) dialog the user can define the set of words that will be used for building the primary stemmer's stems and for each other trial stemmer's stems. This set of words together with the results (stems) of the primary semmer's operation will be later presented to the experts who are responsible for expressing their arguments (complaints and confirmations) against the primary stemmer's stems. If the experts are not speakers of the taarget language, the original words should be inserted to this dialog together with translations to the language that experts speak. The user should provide the original word (left empty text box), its translation (right empty text box) and press the button "Add New Word". Repeating this procedure, will produce the list of words (with translations). This list is the resource needed for system's creation of primary stemmer's stems and also this list will permit experts to express their arguments.

The user can build the list in more than one usages of the dialog. In the first usage the user defines some words whith their translations and then use button "Save to File" to persist in file. In the following usages, the user has to load the existing suffixes (use button "Load from File"), next define some more couples (words with translations) and next persist the updated list of words (by pressing the button "Save to File"). After a number of dialog usages the list of words is completed.

The button "Save to DB" is nedded in order to persist the words (actualy couples of word – translation) into the database. This is a required step because this resource should exist into the database. The following is a small excerpt of the work behind the "Save to DB" button (we have replaced translations with '…'):

```
INSERT INTO words values (1, 'ABANDONOHET', '…');
INSERT INTO words values (2, 'ABAS', '…');
INSERT INTO words values (3, 'ABBAS', '…');
INSERT INTO words values (4, 'ABDI', '…');
INSERT INTO words values (5, 'ABSIDË', '…');
INSERT INTO words values (6, 'ABSOLUTISHT', '…');
INSERT INTO words values (7, 'ABUZIMI', '…');
INSERT INTO words values (8, 'ABUZIMIN', '…');
```

Button “Delete” can be used to remove the selected word – translation from the list. Button “Clear” removes all the entries from the word – translation list.

The list of word-translation couples is persisted in a file with path:

C:\stemSuite\<language>\words.txt

For example:

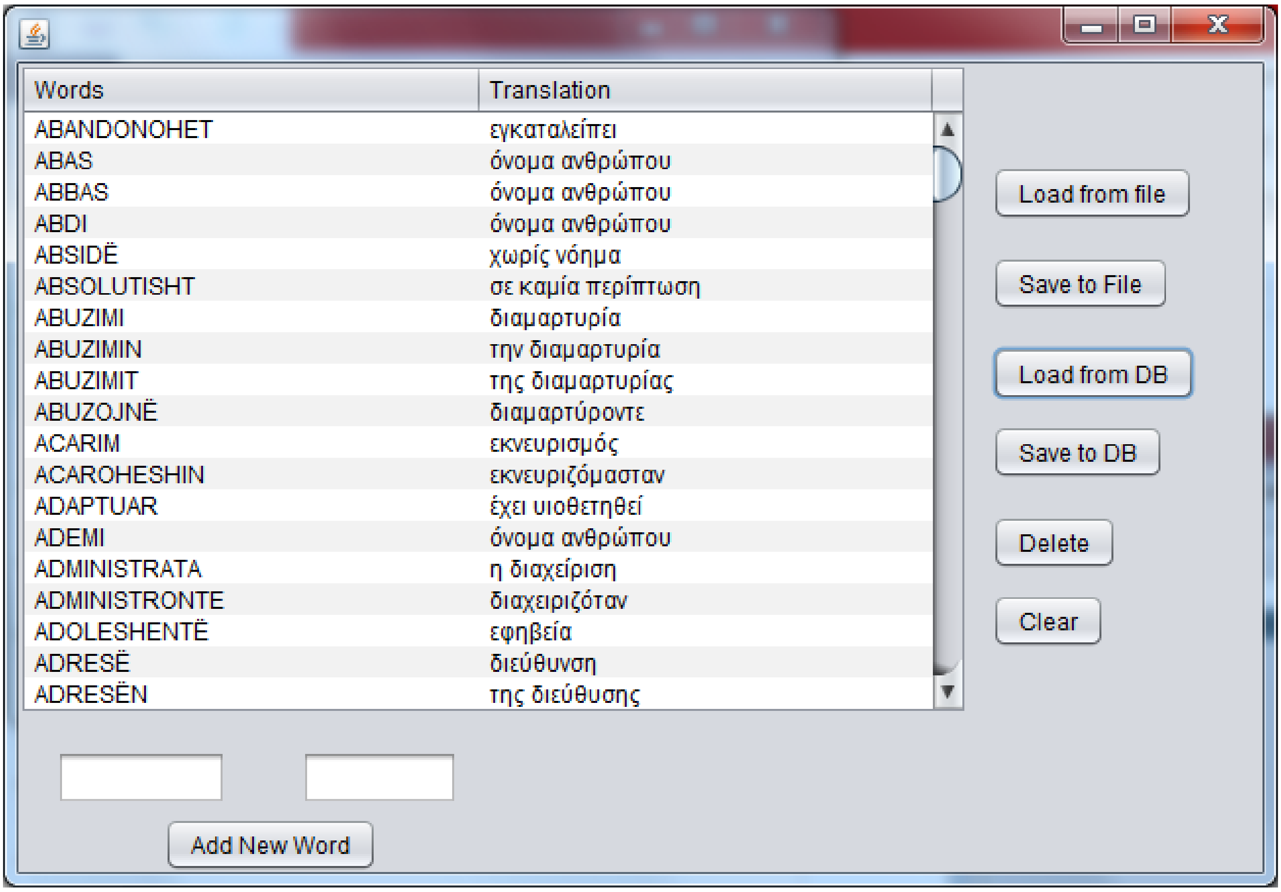
C:\stemSuite\Alban1\words.txt

The list of word-translation couples is persisted also in table words of a database named:

stemSuite_<language>

For example:

stemSuite_Alban1



Words	Translation
ABANDONOHET	εγκαταλείπει
ABAS	όνομα ανθρώπου
ABBAS	όνομα ανθρώπου
ABDI	όνομα ανθρώπου
ABSIDË	χωρίς νόημα
ABSOLUTISHT	σε καμία περίπτωση
ABUZIMI	διαμαρτυρία
ABUZIMIN	την διαμαρτυρία
ABUZIMIT	της διαμαρτυρίας
ABUZOJNË	διαμαρτύροντε
ACARIM	εκνευρισμός
ACAROHESHIN	εκνευριζόμασταν
ADAPTUAR	έχει υιοθετηθεί
ADEMI	όνομα ανθρώπου
ADMINISTRATA	η διαχείριση
ADMINISTRONTE	διαχειριζόταν
ADOLESHENTË	εφηβεία
ADRESË	διεύθυνση
ADRESËN	της διεύθυσης

Language Manager / Add Couples

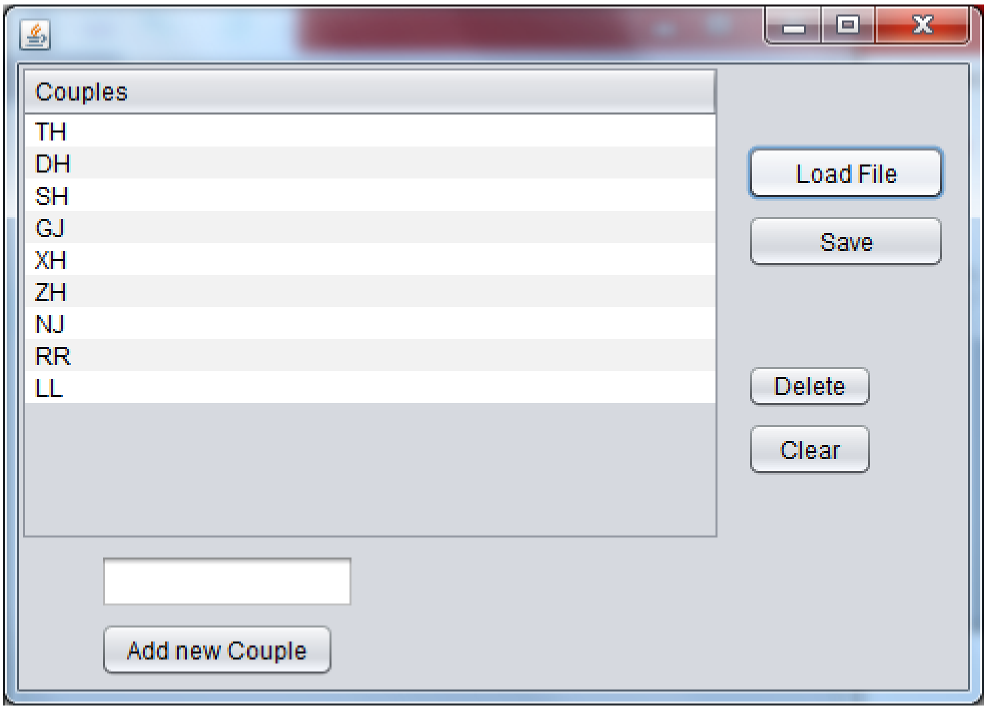
With the same methodology as the one used for the suffixes, the next dialog defines the couples of letters having a single phoneme (vowel digraphs, consonant digraphs and diphthongs).

The list of couples is persisted in a file with path:

C:\stemSuite\<language>\coupleslist.txt

For example:

C:\stemSuite\Alban1\coupleslist.txt



Couples
TH
DH
SH
GJ
XH
ZH
NJ
RR
LL

Language Manager / Add Letters On Border

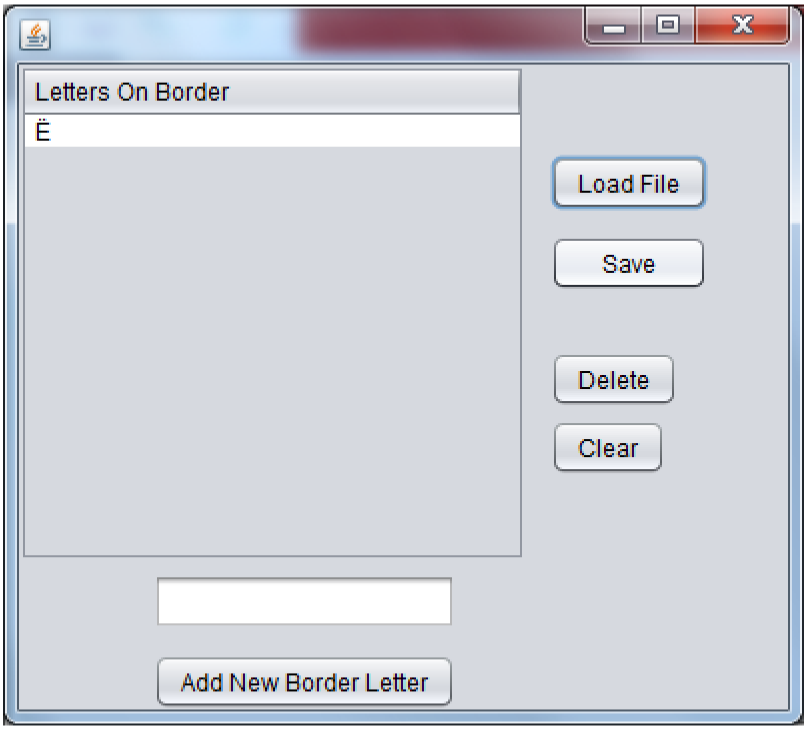
With the same methodology as the one used for the suffixes, the next dialog defines some letters (single graphs) that need special handling.

The list of special letters (letters on border) is persisted in a file with path:

C:\stemSuite\<language>\lettersOnBorderlist.txt

For example:

C:\stemSuite\Alban1\lettersOnBorderlist.txt



Letters On Border
Ë

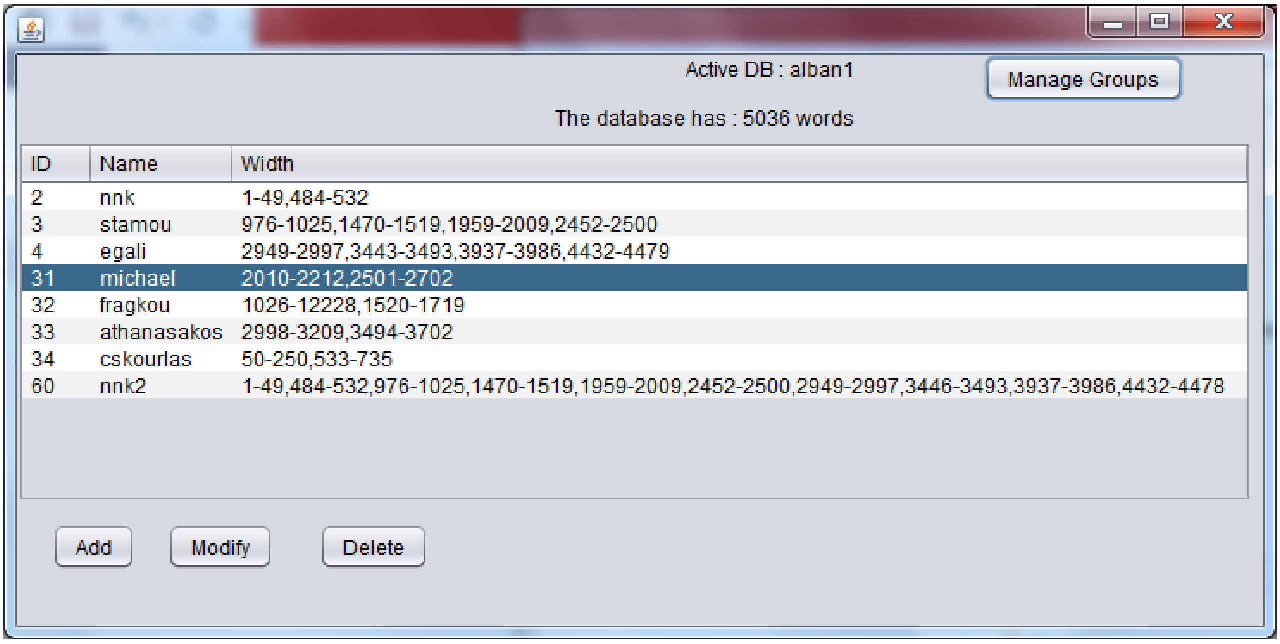
Language Manager / Make Primary Stems

The last button provided by the Language Manager dialog is the Make Primary Stems button. This button does not invoke any graphical interface. It works silently and creates the primary stemmer's stem which are peristed in the database. So this is the first automation of task used to be done manually. Instead of wtitting a computer program that read words and remove the longest matching suffix (among the defined for the language suffixes), then running this program to get stems for each defined (for the language) word and finaly convering the output into "insert into…" sql statements for inserting the results (the stems) into database, we simply press the button "Make Primary Stems". Following the example of words given previously, the last step behind "Make Primary Stems" button will be:

```
INSERT INTO sources values (1, 'Primary Stemmer\'s stems', 'STEMMER');
INSERT INTO stems values (1, 1, 'ABANDON');
INSERT INTO stems values (1, 2, 'ABAS');
INSERT INTO stems values (1, 3, 'ABB');
INSERT INTO stems values (1, 4, 'ABDI');
INSERT INTO stems values (1, 5, 'ABSID');
INSERT INTO stems values (1, 6, 'ABSOLUT');
INSERT INTO stems values (1, 7, 'ABUZ');
INSERT INTO stems values (1, 8, 'ABUZ');
```

Experts Manager

The experts manager is a program with grphical user interface that we need in order to declare the experts that are going to provide arguments (complaints and confirmations) against the primary stemmer's stems. For each Expert we declare his/hers urer name and password and also define the ranges of words that the expert will be able to see and express arguments. As you can see in the next screenshot, "michael" is one of the experts and he have the ability to express arguments for two ranges. Namely: he can express arguments for words having identifiers in range 2010-2212 and in range 2501-2702.



Active DB : alban1

The database has : 5036 words

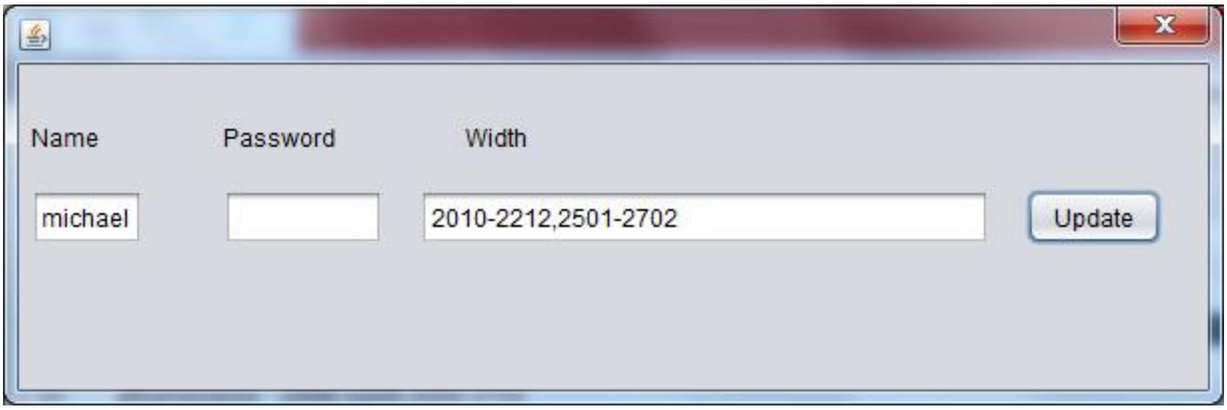
ID	Name	Width
2	nnk	1-49,484-532
3	stamou	976-1025,1470-1519,1959-2009,2452-2500
4	egali	2949-2997,3443-3493,3937-3986,4432-4479
31	michael	2010-2212,2501-2702
32	fragkou	1026-12228,1520-1719
33	athanasakos	2998-3209,3494-3702
34	cskourlas	50-250,533-735
60	nnk2	1-49,484-532,976-1025,1470-1519,1959-2009,2452-2500,2949-2997,3446-3493,3937-3986,4432-4478

To invoke Experts Manager:

C:\stemSuite\bin> **java –jar ExpertsManager.jar**

Experts Manager / Modify

The “modify” button from the Experts Manager dialog invokes the next GUI dialog for changing the password and the ranges of words that an expert can express arguments. The other two buttons (“add” and “delete”) available in the main Experts Manager dialog have obvious meanings (add a new Expert and delete an existing Expert, resppectively).



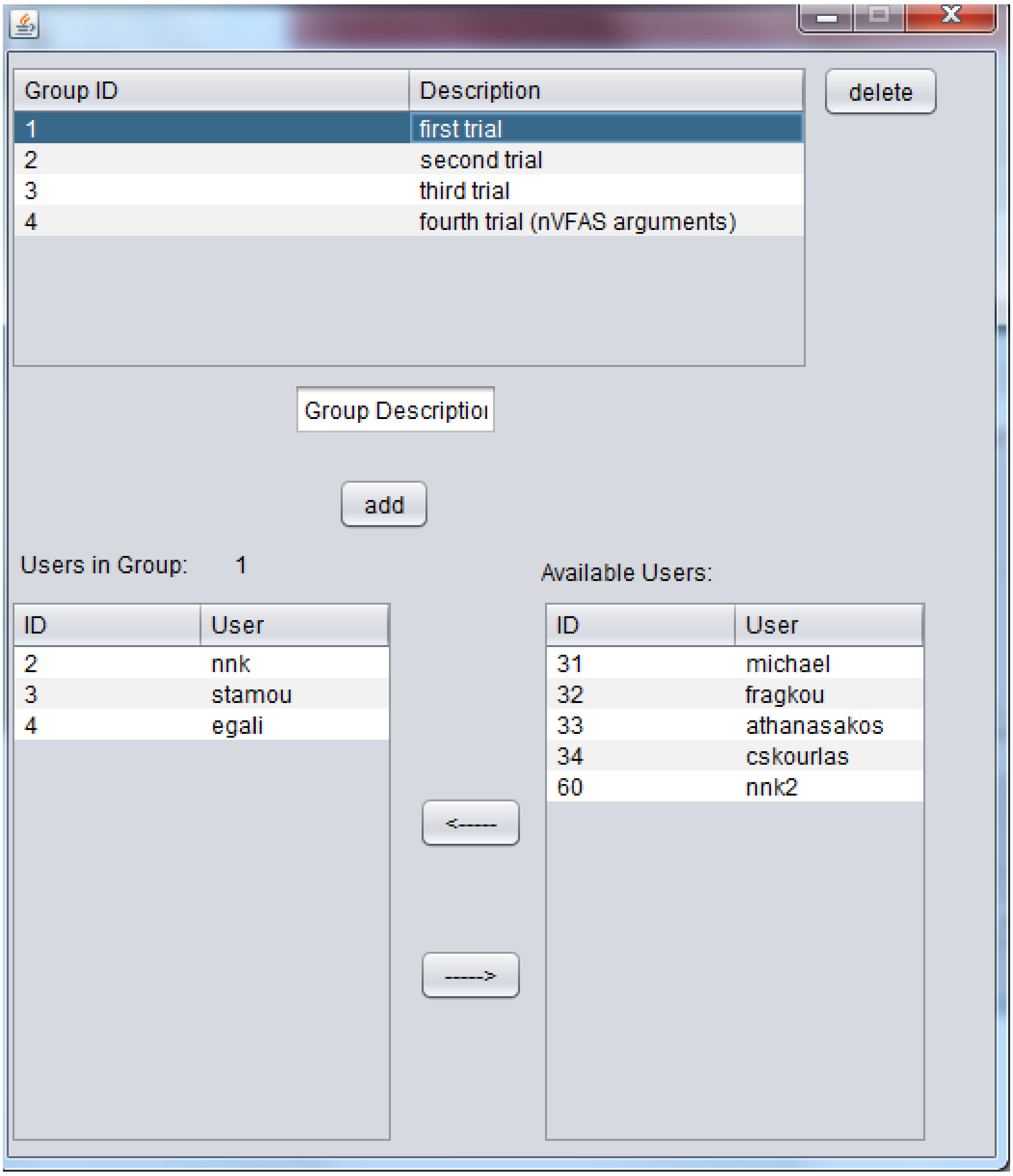
Name	Password	Width
michael		2010-2212,2501-2702

Experts Manager / Manage Groups

The Experts Manager offers also the “Manage Groups” button. This button invokes another dialog where the user can compose more than one experts in a set of experts named group.

In the following screenshot we can see that the user has selected the group named **“first trial”** and that this group has as its members the expers “stamou”, “nnk” and “egali”.

Groups can be used later while building new trial stemmers. The arguments of any expert alone or the set of arguments of all experts belonging to the same group can be used by the wizard to adapt the stemmer in order to conform (as much as possible) with the arguments.



Group ID	Description
1	first trial
2	second trial
3	third trial
4	fourth trial (nVFAS arguments)

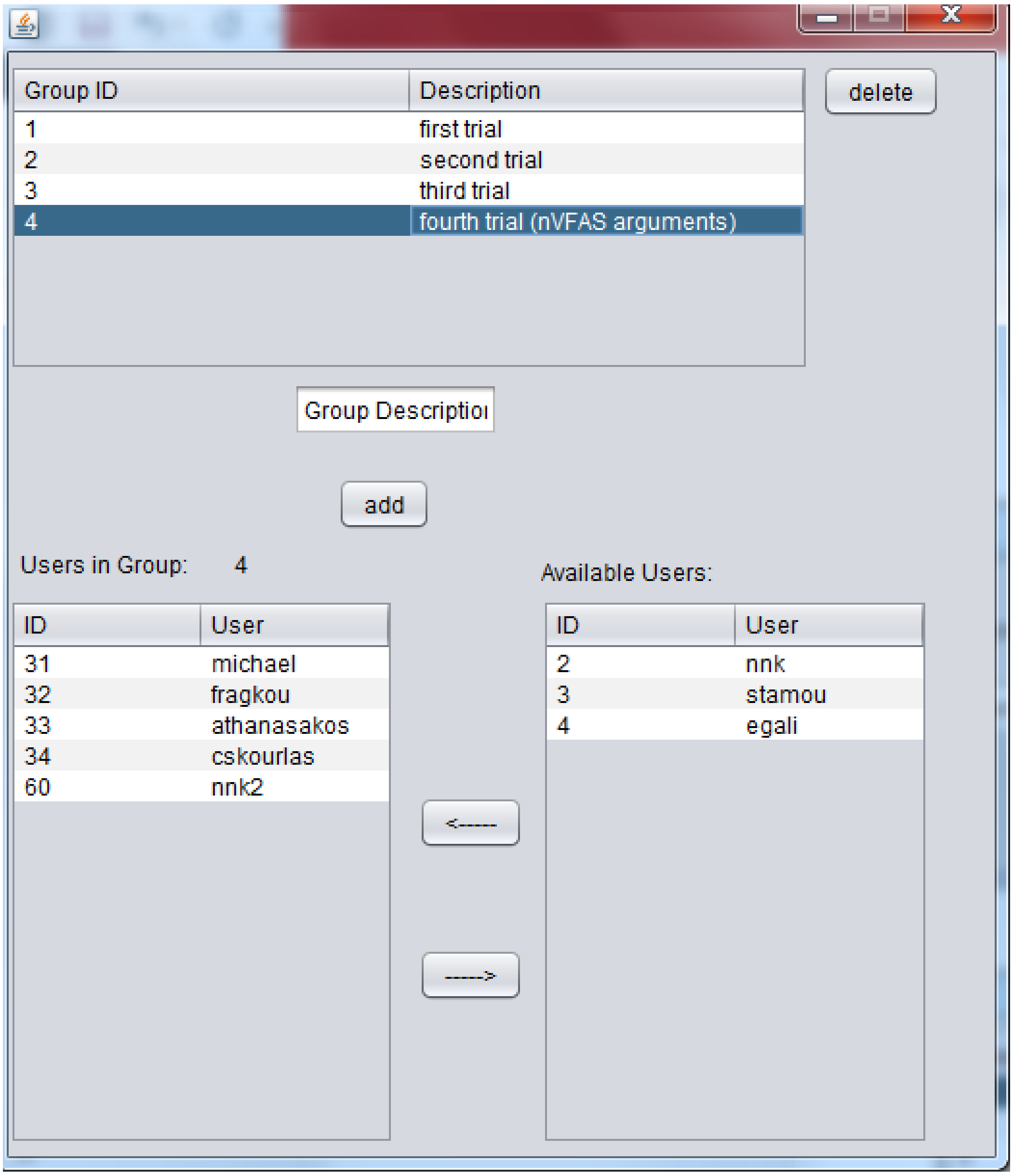
Users in Group: 1

ID	User
2	nnk
3	stamou
4	egali

Available Users:

ID	User
31	michael
32	fragkou
33	athanasakos
34	cskourlas
60	nnk2

In the following screenshot we can see another group named “**fourth trial (nVFAS arguments)**”. This group has as its members the expers “michael”, “fragkou”, “athanasakos”, “skourls” and “nnk2”. This group (according to the ranges we see in the main Experts Manager screen) has about 2100 words that can have an equivalent number of arguments.



Group ID	Description
1	first trial
2	second trial
3	third trial
4	fourth trial (nVFAS arguments)

Users in Group: 4

ID	User
31	michael
32	fragkou
33	athanasakos
34	cskourlas
60	nnk2

Available Users:

ID	User
2	nnk
3	stamou
4	egali

Stem Editor

The Stem Editor is another case of automation of the building stemmer process. Till now expert had to use excel (or other spreadsheet software) in order to declare complaints or verifications against/in favor to the results of the primary stemmer's stems. The spreadsheet files had ready made 3 columns: Word, Stem and Translation. The expert had to augment with an Argument column. The argument was (not about a single word, but) about a set of neighbour words and could be something like:

DS (different stem)

CS (common stem)

DS/CS (Different Stems with subsets of Common Stems)

As small excerpt of such an spreadsheet file follows:

id	word	stem	translation	argument	
18	ADRESË	ADRE	διεύθυνση	CS	DS
19	ADRESËN	ADRES	της διεύθυσης		
20	ADRESOI	ADRES	διευθύνει		
21	ADRESUAR	ADRES	διευθετημένος	CS	

Another example of argument expressed in a spreadsheet is the following:

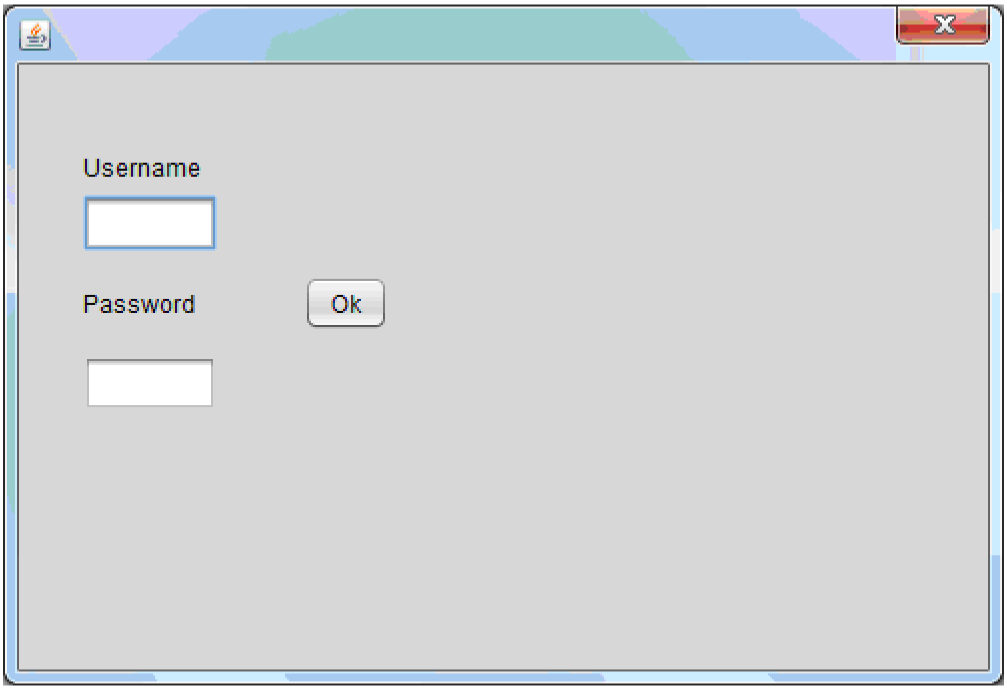
id	word	stem	Translation	argument	
994	FAKT	FAKT	πραγματικότητα (δεδομένο)	CS	DS
995	FAKTI	FAK	το λάθος	CS	
996	FAKTIN	FAKT	του λάθος		
997	FAKTIT	FAK	το λάθος		
998	FAKTOR	FAK	Παράγων	CS	
999	FAKTORË	FAKT	οι παράγοντες		
1000	FAKTORI	FAKT	ο παράγων		

Next, some programmer (or other expert) had to translate the arguments into sql statements. Now, (in order to "enhance the work-flow of Building Stemmers") we have build a new tool with graphical user interface that permit experts to express directly their arguments and the arguments are translated automatically to sql statements. This tool Let name this tool is named **Stem Editor** (but also **ExpressArguments**).

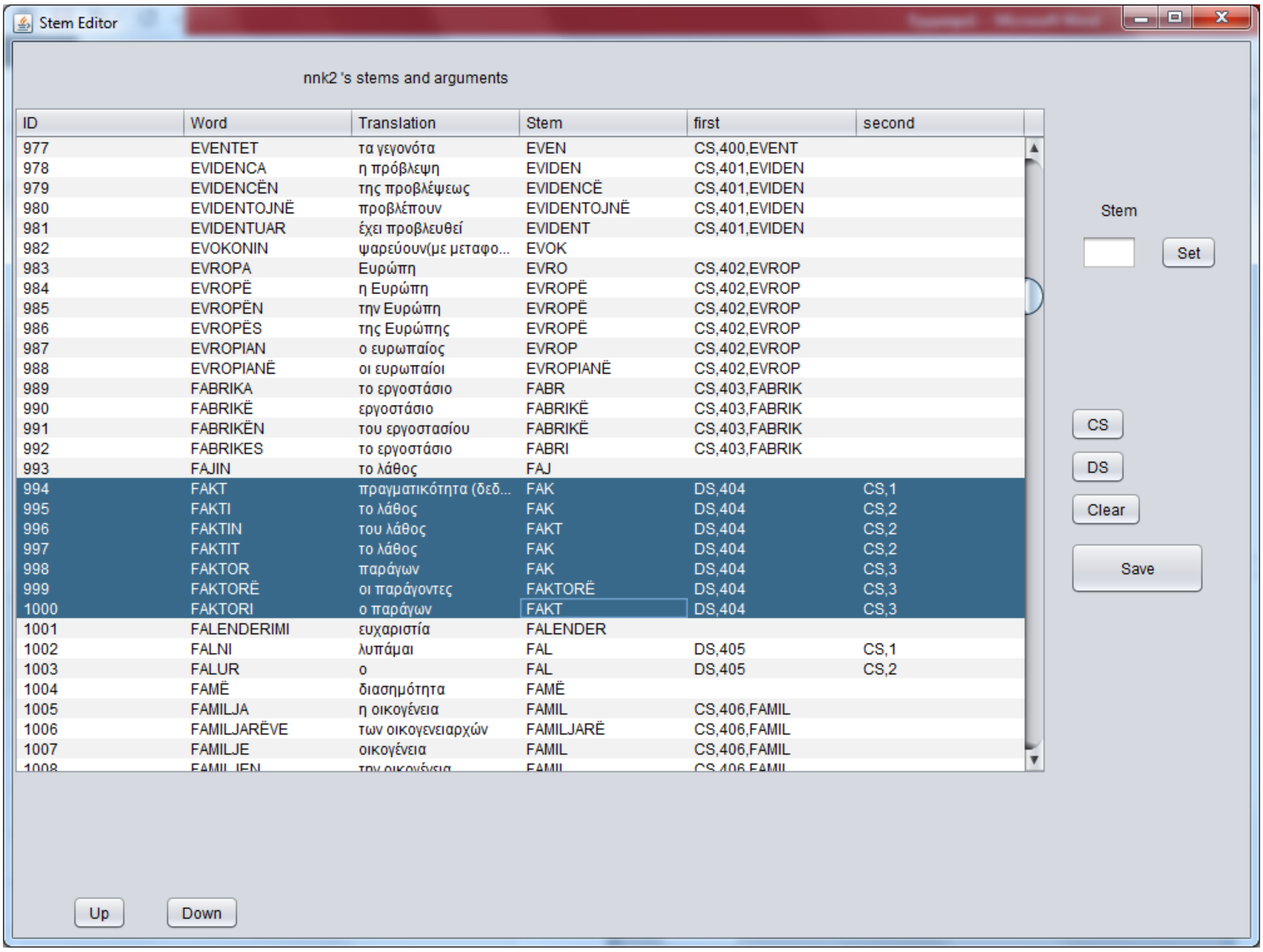
To invoke Stem Editor:

C:\stemSuite\bin> **java –jar StemEditorV4.jar**

Of course the user has first to login, using the next dialog:



After successful login, Stem Editor projects to the expert the words that the expert is authorized to see and express arguments. The following is a screenshot with the words that expert nnk2 is authorized to handle. In this screenshot, we can see that the expert has already expressed a complaint argument by defining that the set of words with ids 994..1000 should have different stems (DS) with 3 subsets of words having common stem (CS). In the same screenshot we can also see some other CS arguments. For example words 978..981 should have a common stem (CS) with suggested stem value of “EVIDEN”. The numbers (e.g. 404 and 401 in the mentioned examples) are the identifiers of the arguments and are of no interest for understanding the general idea.

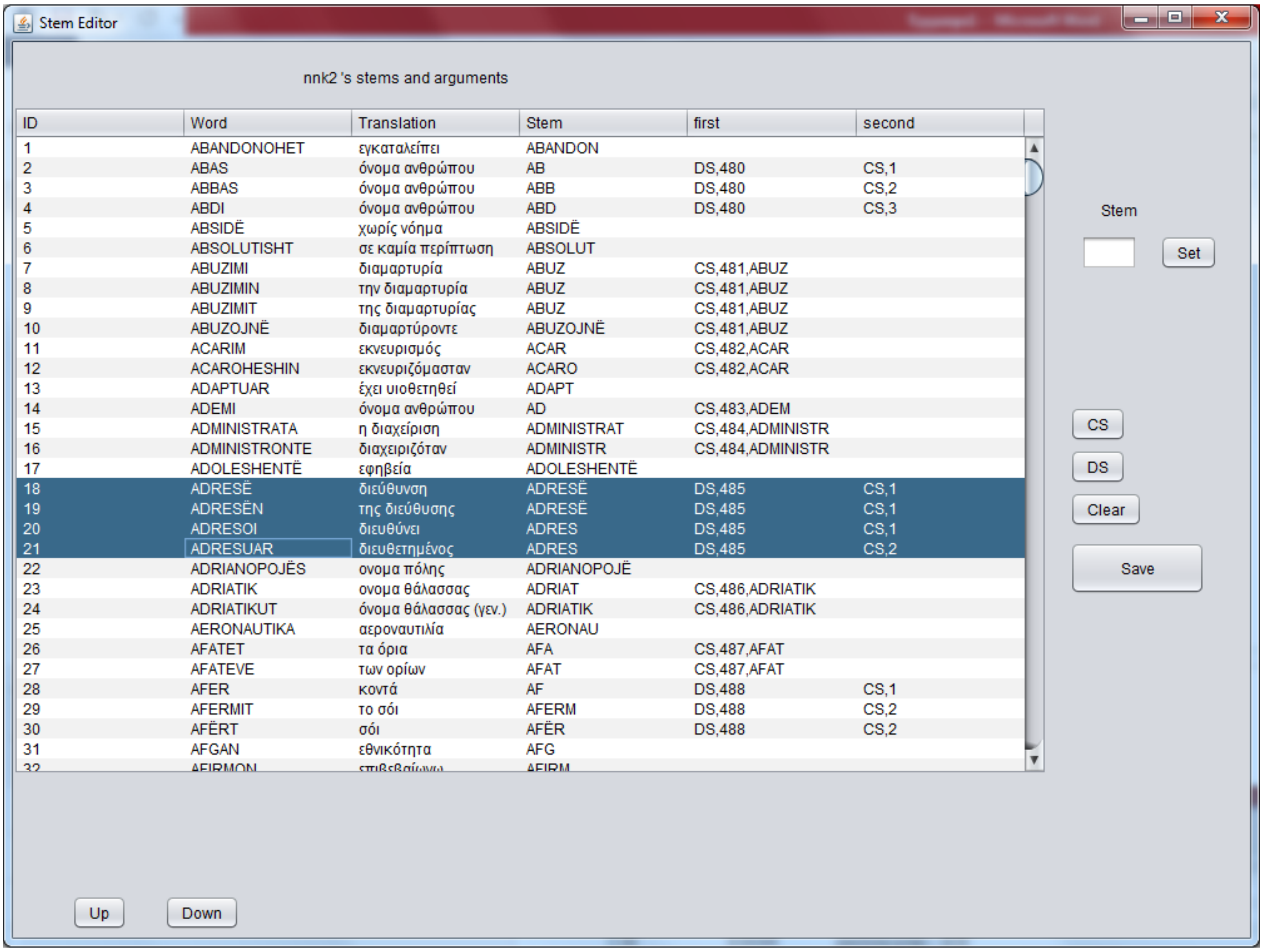


nnk2 's stems and arguments

ID	Word	Translation	Stem	first	second
977	EVENTET	τα γεγονότα	EVEN	CS,400,EVENT	
978	EVIDENCA	η πρόβλεψη	EVIDEN	CS,401,EVIDEN	
979	EVIDENCËN	της προβλέψεως	EVIDENCË	CS,401,EVIDEN	
980	EVIDENTOJNË	προβλέπουν	EVIDENTOJNË	CS,401,EVIDEN	
981	EVIDENTUAR	έχει προβλευθεί	EVIDENT	CS,401,EVIDEN	
982	EVOKONIN	ψαρεύουν(με μεταφο...	EVOK		
983	EVROPA	Ευρώπη	EVRO	CS,402,EVROP	
984	EVROPË	η Ευρώπη	EVROPË	CS,402,EVROP	
985	EVROPËN	την Ευρώπη	EVROPË	CS,402,EVROP	
986	EVROPËS	της Ευρώπης	EVROPË	CS,402,EVROP	
987	EVROPIAN	ο ευρωπαίος	EVROP	CS,402,EVROP	
988	EVROPIANË	οι ευρωπαίοι	EVROPIANË	CS,402,EVROP	
989	FABRIKA	το εργοστάσιο	FABR	CS,403,FABRIK	
990	FABRIKË	εργοστάσιο	FABRIKË	CS,403,FABRIK	
991	FABRIKËN	του εργοστασίου	FABRIKË	CS,403,FABRIK	
992	FABRIKES	το εργοστάσιο	FABRI	CS,403,FABRIK	
993	FAJIN	το λάθος	FAJ		
994	FAKT	πραγματικότητα (δεδ...	FAK	DS,404	CS,1
995	FAKTI	το λάθος	FAK	DS,404	CS,2
996	FAKTIN	του λάθος	FAKT	DS,404	CS,2
997	FAKTIT	το λάθος	FAK	DS,404	CS,2
998	FAKTOR	παράγων	FAK	DS,404	CS,3
999	FAKTORË	οι παράγοντες	FAKTORË	DS,404	CS,3
1000	FAKTORI	ο παράγων	FAKT	DS,404	CS,3
1001	FALENDERIMI	ευχαριστία	FALENDER		
1002	FALNI	λυπάμαι	FAL	DS,405	CS,1
1003	FALUR	ο	FAL	DS,405	CS,2
1004	FAMË	διασημότητα	FAMË		
1005	FAMILJA	η οικογένεια	FAMIL	CS,406,FAMIL	
1006	FAMILJARËVE	των οικογενειαρχών	FAMILJARË	CS,406,FAMIL	
1007	FAMILJE	οικογένεια	FAMIL	CS,406,FAMIL	
1008	FAMILJEN	την οικογένεια	FAMIL	CS,406,FAMIL	

Another example of already expressed DS/CS argument

Next screenshot depicts a DS/CS argument about words 18..21. It has two CS subsets. The first CS subset is about words with ids 18..20 while the second subset is only about the word with id 21. Number 485 is the internal identifier of the whole (4 word set) argument.

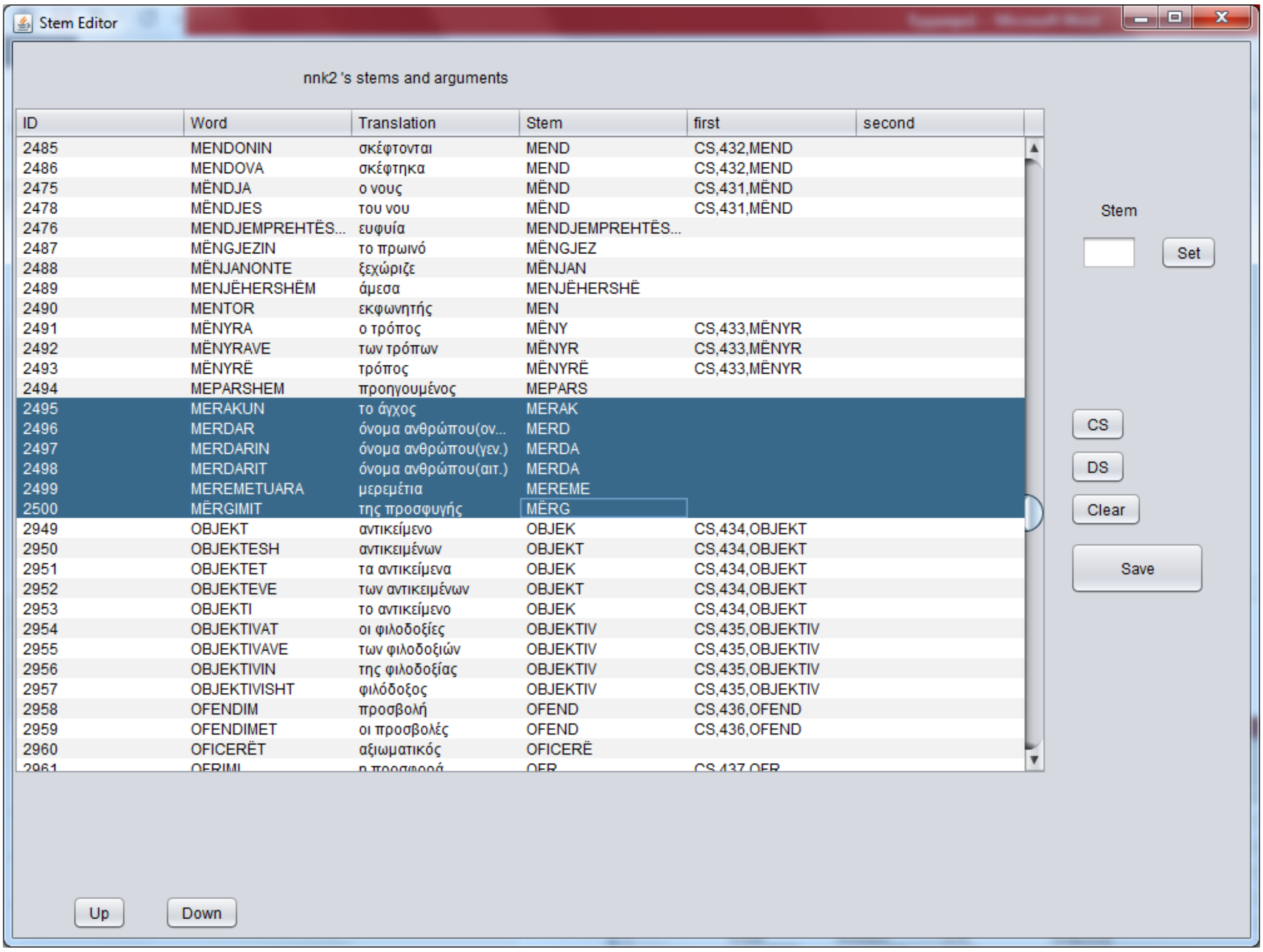


nnk2 's stems and arguments

ID	Word	Translation	Stem	first	second
1	ABANDONOHET	εγκαταλείπει	ABANDON		
2	ABAS	όνομα ανθρώπου	AB	DS,480	CS,1
3	ABBAS	όνομα ανθρώπου	ABB	DS,480	CS,2
4	ABDI	όνομα ανθρώπου	ABD	DS,480	CS,3
5	ABSIDË	χωρίς νόημα	ABSIDË		
6	ABSOLUTISHT	σε καμία περίπτωση	ABSOLUT		
7	ABUZIMI	διαμαρτυρία	ABUZ	CS,481,ABUZ	
8	ABUZIMIN	την διαμαρτυρία	ABUZ	CS,481,ABUZ	
9	ABUZIMIT	της διαμαρτυρίας	ABUZ	CS,481,ABUZ	
10	ABUZOJNË	διαμαρτύροντε	ABUZOJNË	CS,481,ABUZ	
11	ACARIM	εκνευρισμός	ACAR	CS,482,ACAR	
12	ACAROHESHIN	εκνευριζόμασταν	ACARO	CS,482,ACAR	
13	ADAPTUAR	έχει υιοθετηθεί	ADAPT		
14	ADEMI	όνομα ανθρώπου	AD	CS,483,ADEM	
15	ADMINISTRATA	η διαχείριση	ADMINISTRAT	CS,484,ADMINISTR	
16	ADMINISTRONTE	διαχειριζόταν	ADMINISTR	CS,484,ADMINISTR	
17	ADOLESHENTË	εφηβεία	ADOLESHENTË		
18	ADRESË	διεύθυνση	ADRESË	DS,485	CS,1
19	ADRESËN	της διεύθυσης	ADRESË	DS,485	CS,1
20	ADRESOI	διευθύνει	ADRES	DS,485	CS,1
21	ADRESUAR	διευθετημένος	ADRES	DS,485	CS,2
22	ADRIANOPOJËS	ονομα πόλης	ADRIANOPOJË		
23	ADRIATIK	ονομα θάλασσας	ADRIAT	CS,486,ADRIATIK	
24	ADRIATIKUT	όνομα θάλασσας (γεν.)	ADRIATIK	CS,486,ADRIATIK	
25	AERONAUTIKA	αεροναυτιλία	AERONAU		
26	AFATET	τα όρια	AFA	CS,487,AFAT	
27	AFATEVE	των ορίων	AFAT	CS,487,AFAT	
28	AFER	κοντά	AF	DS,488	CS,1
29	AFERMIT	το σόι	AFERM	DS,488	CS,2
30	AFËRT	σόι	AFËR	DS,488	CS,2
31	AFGAN	εθνικότητα	AFG		
32	AFIRMON	επιβεβαίωνω	AFIRM		

Stem Editor – step by step definition of a DS/CS argument

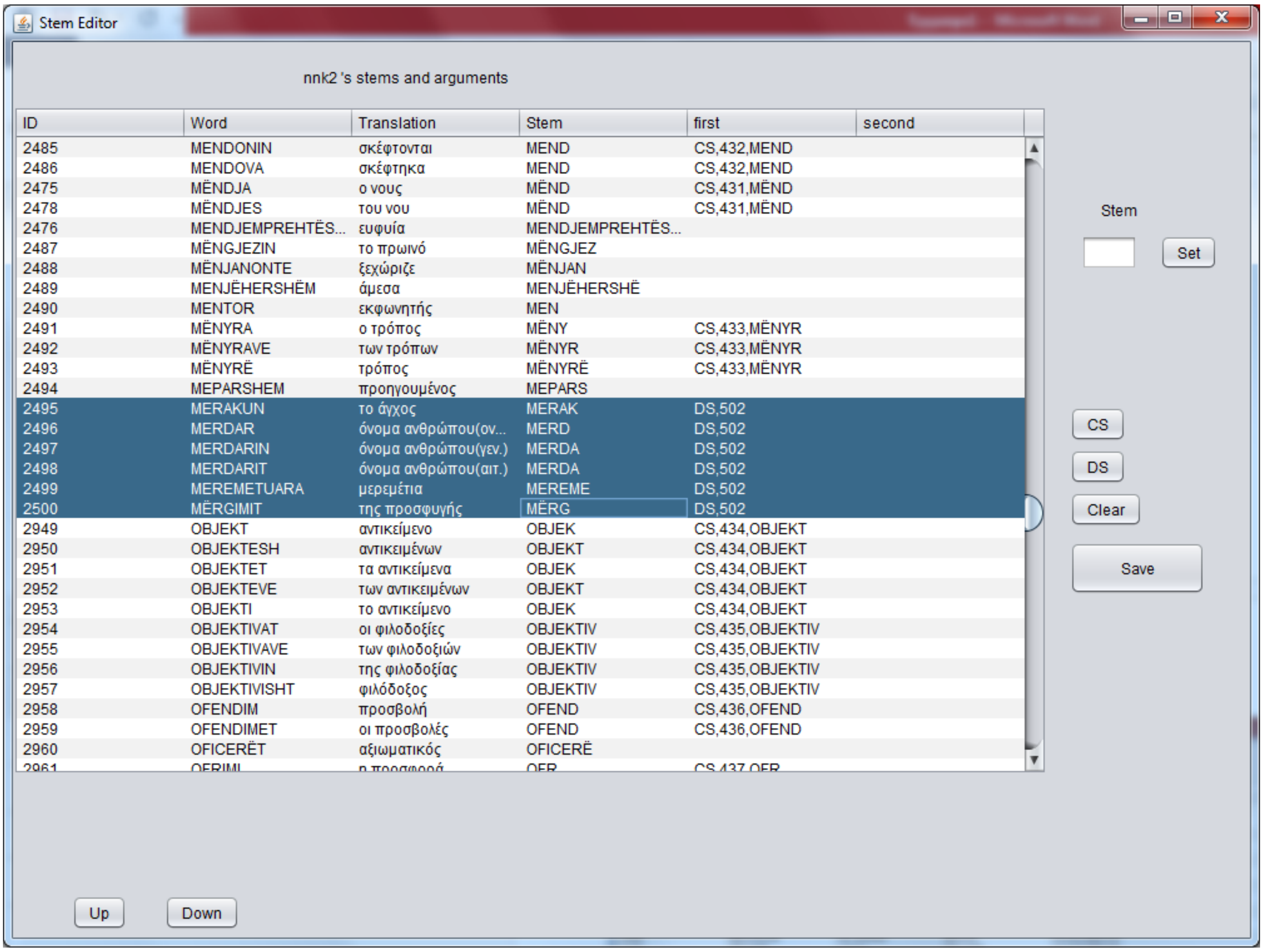
In next screenshot we can see that there is no argument about words 2495..2500. We will declare a DS/CS argument (complaint) about these words.



nnk2 's stems and arguments

ID	Word	Translation	Stem	first	second
2485	MENDONIN	σκέφτονται	MEND	CS,432,MEND	
2486	MENDOVA	σκέφτηκα	MEND	CS,432,MEND	
2475	MËNDJA	ο νους	MËND	CS,431,MËND	
2478	MËNDJES	του νου	MËND	CS,431,MËND	
2476	MENDJEMPREHTËS...	ευφυία	MENDJEMPREHTËS...		
2487	MËNGJEZIN	το πρωινό	MËNGJEZ		
2488	MËNJANONTE	ξεχώριζε	MËNJAN		
2489	MENJËHERSHËM	άμεσα	MENJËHERSHË		
2490	MENTOR	εκφωνητής	MEN		
2491	MËNYRA	ο τρόπος	MËNY	CS,433,MËNYR	
2492	MËNYRAVE	των τρόπων	MËNYR	CS,433,MËNYR	
2493	MËNYRË	τρόπος	MËNYRË	CS,433,MËNYR	
2494	MEPARSHEM	προηγουμένος	MEPARS		
2495	MERAKUN	το άγχος	MERAK		
2496	MERDAR	όνομα ανθρώπου(ον...	MERD		
2497	MERDARIN	όνομα ανθρώπου(γεν.)	MERDA		
2498	MERDARIT	όνομα ανθρώπου(αιτ.)	MERDA		
2499	MEREMETUARA	μερεμέτια	MEREME		
2500	MËRGIMIT	της προσφυγής	MËRG		
2949	OBJEKT	αντικείμενο	OBJEK	CS,434,OBJEKT	
2950	OBJEKTESH	αντικειμένων	OBJEKT	CS,434,OBJEKT	
2951	OBJEKTET	τα αντικείμενα	OBJEK	CS,434,OBJEKT	
2952	OBJEKTEVE	των αντικειμένων	OBJEKT	CS,434,OBJEKT	
2953	OBJEKTI	το αντικείμενο	OBJEK	CS,434,OBJEKT	
2954	OBJEKTIVAT	οι φιλοδοξίες	OBJEKTIV	CS,435,OBJEKTIV	
2955	OBJEKTIVAVE	των φιλοδοξιών	OBJEKTIV	CS,435,OBJEKTIV	
2956	OBJEKTIVIN	της φιλοδοξίας	OBJEKTIV	CS,435,OBJEKTIV	
2957	OBJEKTIVISHT	φιλόδοξος	OBJEKTIV	CS,435,OBJEKTIV	
2958	OFENDIM	προσβολή	OFEND	CS,436,OFEND	
2959	OFENDIMET	οι προσβολές	OFEND	CS,436,OFEND	
2960	OFICERËT	αξιωματικός	OFICERË		
2961	OFERIMI	η προσφορά	OFER	CS,437,OFER	

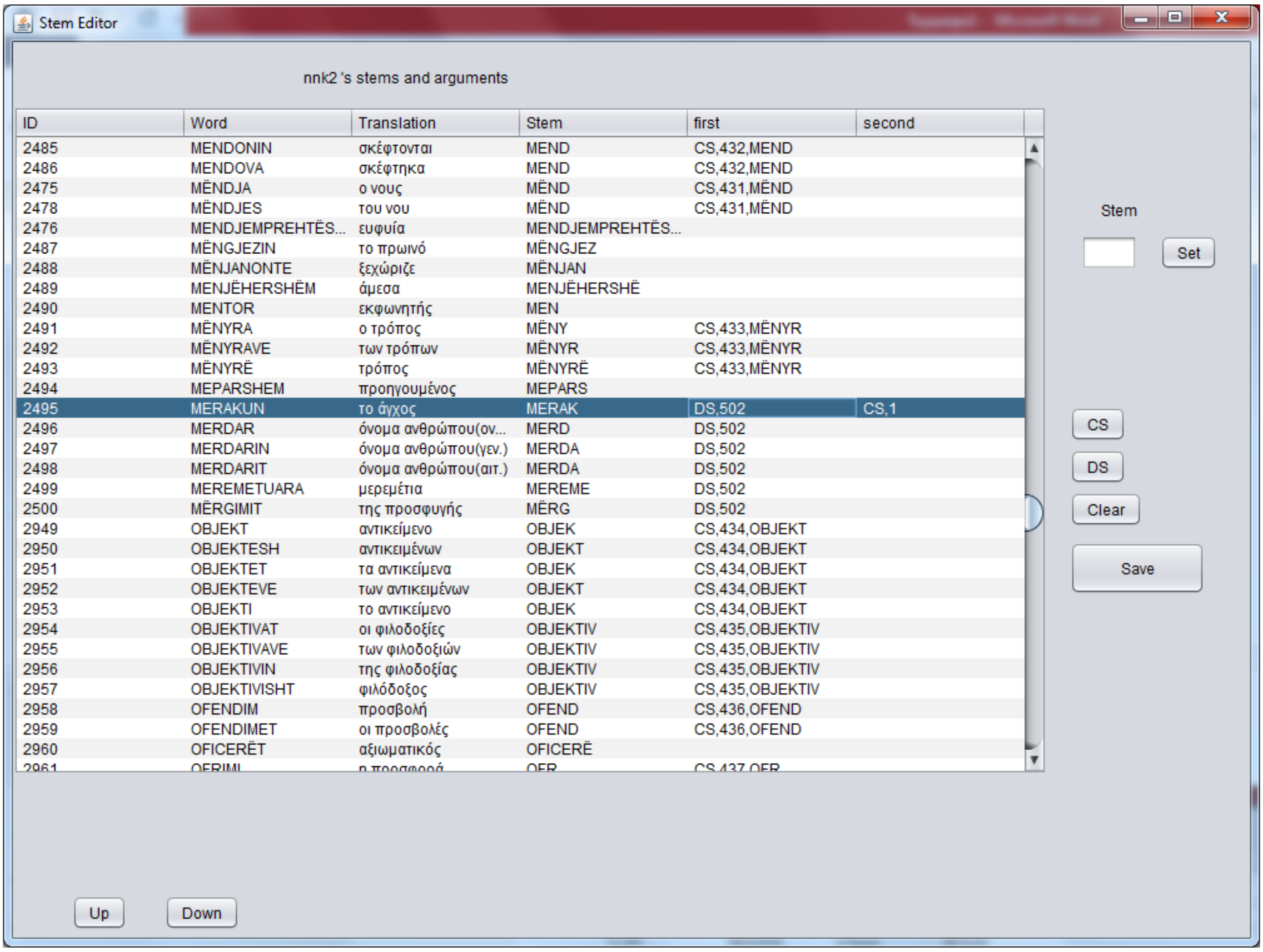
First, in the next screenshot, we declare the DS (different stem) argument. To do so, we have to select the words (with ID ranging 2495 to 2500) and press the button DS. The result is depicted in the following screenshot.



nnk2 's stems and arguments

ID	Word	Translation	Stem	first	second
2485	MENDONIN	σκέφτονται	MEND	CS,432,MEND	
2486	MENDOVA	σκέφτηκα	MEND	CS,432,MEND	
2475	MËNDJA	ο νους	MËND	CS,431,MËND	
2478	MËNDJES	του νου	MËND	CS,431,MËND	
2476	MENDJEMPREHTËS...	ευφυία	MENDJEMPREHTËS...		
2487	MËNGJEZIN	το πρωινό	MËNGJEZ		
2488	MËNJANONTE	ξεχώριζε	MËNJAN		
2489	MENJËHERSHËM	άμεσα	MENJËHERSHË		
2490	MENTOR	εκφωνητής	MEN		
2491	MËNYRA	ο τρόπος	MËNY	CS,433,MËNYR	
2492	MËNYRAVE	των τρόπων	MËNYR	CS,433,MËNYR	
2493	MËNYRË	τρόπος	MËNYRË	CS,433,MËNYR	
2494	MEPARSHEM	προηγουμένος	MEPARS		
2495	MERAKUN	το άγχος	MERAK	DS,502	
2496	MERDAR	όνομα ανθρώπου(ον...	MERD	DS,502	
2497	MERDARIN	όνομα ανθρώπου(γεν.)	MERDA	DS,502	
2498	MERDARIT	όνομα ανθρώπου(αιτ.)	MERDA	DS,502	
2499	MEREMETUARA	μερεμέτια	MEREME	DS,502	
2500	MËRGIMIT	της προσφυγής	MËRG	DS,502	
2949	OBJEKT	αντικείμενο	OBJEK	CS,434,OBJEKT	
2950	OBJEKTESH	αντικειμένων	OBJEKT	CS,434,OBJEKT	
2951	OBJEKTET	τα αντικείμενα	OBJEK	CS,434,OBJEKT	
2952	OBJEKTEVE	των αντικειμένων	OBJEKT	CS,434,OBJEKT	
2953	OBJEKTI	το αντικείμενο	OBJEK	CS,434,OBJEKT	
2954	OBJEKTIVAT	οι φιλοδοξίες	OBJEKTIV	CS,435,OBJEKTIV	
2955	OBJEKTIVAVE	των φιλοδοξιών	OBJEKTIV	CS,435,OBJEKTIV	
2956	OBJEKTIVIN	της φιλοδοξίας	OBJEKTIV	CS,435,OBJEKTIV	
2957	OBJEKTIVISHT	φιλόδοξος	OBJEKTIV	CS,435,OBJEKTIV	
2958	OFENDIM	προσβολή	OFEND	CS,436,OFEND	
2959	OFENDIMET	οι προσβολές	OFEND	CS,436,OFEND	
2960	OFICERËT	αξιωματικός	OFICERË		
2961	OFRIMI	η προσφορά	OFR	CS,437,OFR	

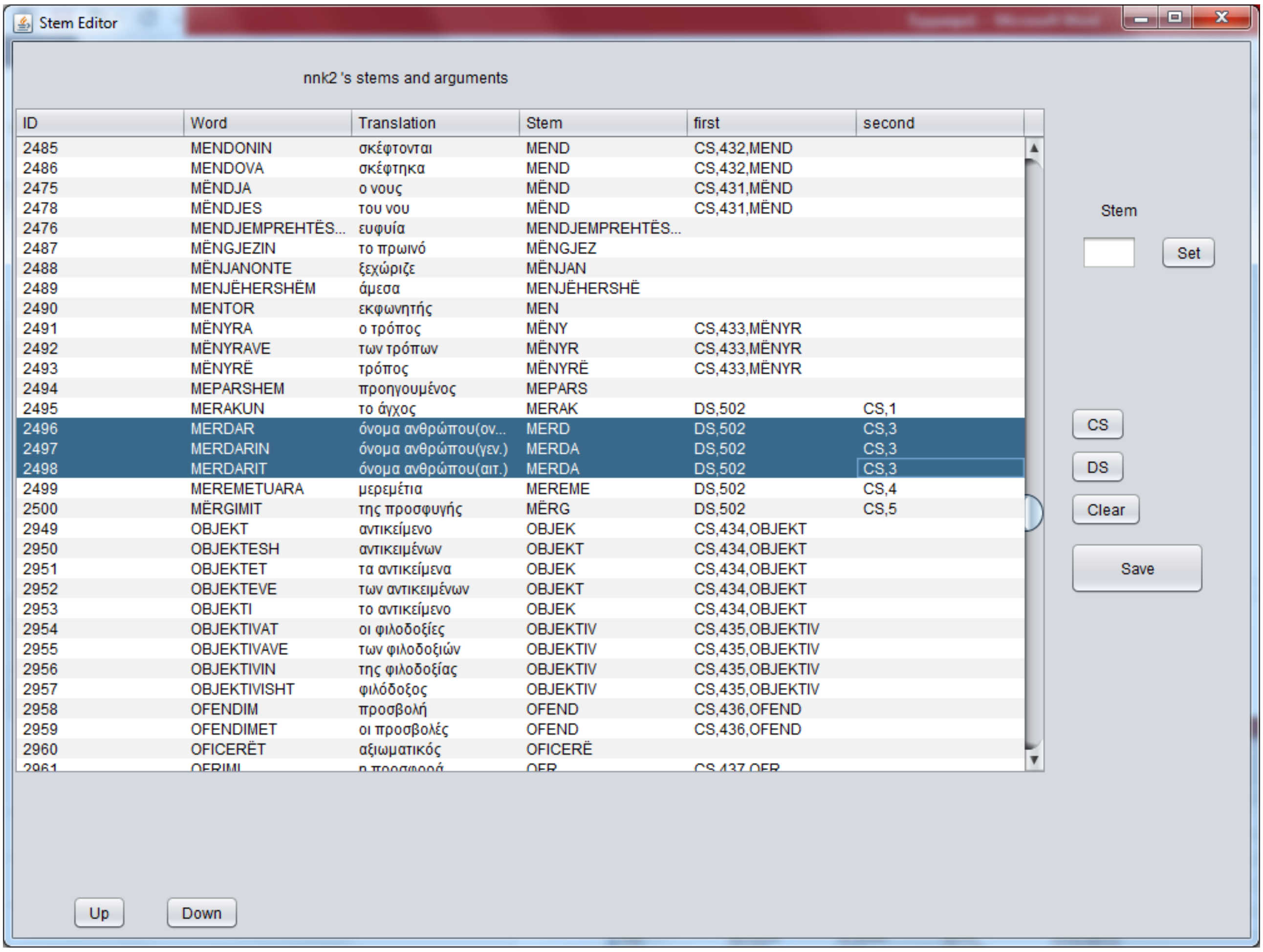
In order to declare (define) one of the suggested CS subset, we select the words of the subset and press the button CS. In the following screenshot we can see the first CS subset having only one item (the word with id 2495)



nnk2 's stems and arguments

ID	Word	Translation	Stem	first	second
2485	MENDONIN	σκέφτονται	MEND	CS,432,MEND	
2486	MENDOVA	σκέφτηκα	MEND	CS,432,MEND	
2475	MËNDJA	ο νους	MËND	CS,431,MËND	
2478	MËNDJES	του νου	MËND	CS,431,MËND	
2476	MENDJEMPREHTËS...	ευφυία	MENDJEMPREHTËS...		
2487	MËNGJEZIN	το πρωινό	MËNGJEZ		
2488	MËNJANONTE	ξεχώριζε	MËNJAN		
2489	MENJËHERSHËM	άμεσα	MENJËHERSHË		
2490	MENTOR	εκφωνητής	MEN		
2491	MËNYRA	ο τρόπος	MËNY	CS,433,MËNYR	
2492	MËNYRAVE	των τρόπων	MËNYR	CS,433,MËNYR	
2493	MËNYRË	τρόπος	MËNYRË	CS,433,MËNYR	
2494	MEPARSHEM	προηγουμένος	MEPARS		
2495	MERAKUN	το άγχος	MERAK	DS,502	CS,1
2496	MERDAR	όνομα ανθρώπου(ον...	MERD	DS,502	
2497	MERDARIN	όνομα ανθρώπου(γεν.)	MERDA	DS,502	
2498	MERDARIT	όνομα ανθρώπου(αιτ.)	MERDA	DS,502	
2499	MEREMETUARA	μερεμέτια	MEREME	DS,502	
2500	MËRGIMIT	της προσφυγής	MËRG	DS,502	
2949	OBJEKT	αντικείμενο	OBJEK	CS,434,OBJEKT	
2950	OBJEKTESH	αντικειμένων	OBJEKT	CS,434,OBJEKT	
2951	OBJEKTET	τα αντικείμενα	OBJEK	CS,434,OBJEKT	
2952	OBJEKTEVE	των αντικειμένων	OBJEKT	CS,434,OBJEKT	
2953	OBJEKTI	το αντικείμενο	OBJEK	CS,434,OBJEKT	
2954	OBJEKTIVAT	οι φιλοδοξίες	OBJEKTIV	CS,435,OBJEKTIV	
2955	OBJEKTIVAVE	των φιλοδοξιών	OBJEKTIV	CS,435,OBJEKTIV	
2956	OBJEKTIVIN	της φιλοδοξίας	OBJEKTIV	CS,435,OBJEKTIV	
2957	OBJEKTIVISHT	φιλόδοξος	OBJEKTIV	CS,435,OBJEKTIV	
2958	OFENDIM	προσβολή	OFEND	CS,436,OFEND	
2959	OFENDIMET	οι προσβολές	OFEND	CS,436,OFEND	
2960	OFICERËT	αξιωματικός	OFICERË		
2961	OFRIMI	η προσφορά	OFR	CS,437,OFR	

In the next screenshot we can see how we define the 2nd, 3rd and 4th CS subset under (inside) the DS argument. Each time we select some words inside a DS argument and press the CS button, these (the selected) words are forming a CS subset. If by mistake you press the CS button twice for the same subset, you don't have to worry. The subset get an increased number but this is not a problem because subsets are separated between each other. In the next screenshot we can see that there are 4 CD subsets inside the DS and the subsets are numbered 1,3,4 and 5.



nnk2 's stems and arguments

ID	Word	Translation	Stem	first	second
2485	MENDONIN	σκέφτονται	MEND	CS,432,MEND	
2486	MENDOVA	σκέφτηκα	MEND	CS,432,MEND	
2475	MËNDJA	ο νους	MËND	CS,431,MËND	
2478	MËNDJES	του νου	MËND	CS,431,MËND	
2476	MENDJEMPREHTËS...	ευφυία	MENDJEMPREHTËS...		
2487	MËNGJEZIN	το πρωινό	MËNGJEZ		
2488	MËNJANONTE	ξεχώριζε	MËNJAN		
2489	MENJËHERSHËM	άμεσα	MENJËHERSHË		
2490	MENTOR	εκφωνητής	MEN		
2491	MËNYRA	ο τρόπος	MËNY	CS,433,MËNYR	
2492	MËNYRAVE	των τρόπων	MËNYR	CS,433,MËNYR	
2493	MËNYRË	τρόπος	MËNYRË	CS,433,MËNYR	
2494	MEPARSHEM	προηγουμένος	MEPARS		
2495	MERAKUN	το άγχος	MERAK	DS,502	CS,1
2496	MERDAR	όνομα ανθρώπου(ον...	MERD	DS,502	CS,3
2497	MERDARIN	όνομα ανθρώπου(γεν.)	MERDA	DS,502	CS,3
2498	MERDARIT	όνομα ανθρώπου(αιτ.)	MERDA	DS,502	CS,3
2499	MEREMETUARA	μερεμέτια	MEREME	DS,502	CS,4
2500	MËRGIMIT	της προσφυγής	MËRG	DS,502	CS,5
2949	OBJEKT	αντικείμενο	OBJEK	CS,434,OBJEKT	
2950	OBJEKTESH	αντικειμένων	OBJEKT	CS,434,OBJEKT	
2951	OBJEKTET	τα αντικείμενα	OBJEK	CS,434,OBJEKT	
2952	OBJEKTEVE	των αντικειμένων	OBJEKT	CS,434,OBJEKT	
2953	OBJEKTI	το αντικείμενο	OBJEK	CS,434,OBJEKT	
2954	OBJEKTIVAT	οι φιλοδοξίες	OBJEKTIV	CS,435,OBJEKTIV	
2955	OBJEKTIVAVE	των φιλοδοξιών	OBJEKTIV	CS,435,OBJEKTIV	
2956	OBJEKTIVIN	της φιλοδοξίας	OBJEKTIV	CS,435,OBJEKTIV	
2957	OBJEKTIVISHT	φιλόδοξος	OBJEKTIV	CS,435,OBJEKTIV	
2958	OFENDIM	προσβολή	OFEND	CS,436,OFEND	
2959	OFENDIMET	οι προσβολές	OFEND	CS,436,OFEND	
2960	OFICERËT	αξιωματικός	OFICERË		
2961	OFRIMI	η προσφορά	OFR	CS,437,OFR	

Forming (configuring) a trial stemmer

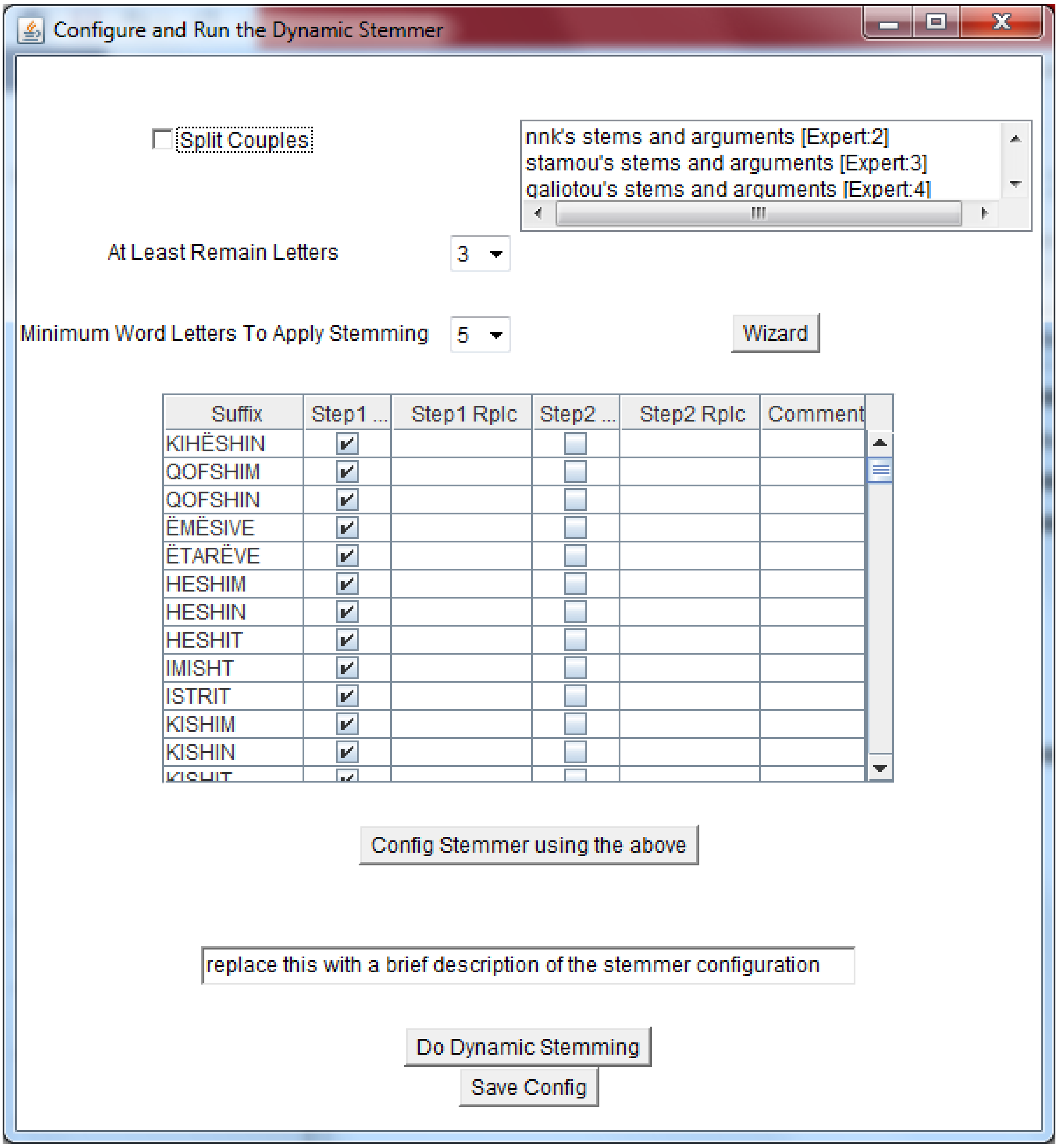
The “trial stemmer builder” (simply “builder”) is a software application with Graphical User Interface. It is used for configuring alternative trial stemmers. All variations of trial stemmers follow the same function which is a two step removal of suffixes. The variation of trial stemmers is based in some configuration parameters but mainly the variation is the result of enabling or disabling a number of suffixes in each (of the two removal) steps. Because both the “stemmer builder” and the “stemmer evaluator” (discussed later) are classes of the same project, the invocation command should clarify which class (“trial stemmer builder” or “trial stemmer evaluator”) we invoke.

Because the project name is StemmerEvaluatorV3, in order to invoke the “trial stemmer builder” we have to issue the command:

```
C:\stemSuite\bin> java -classpath StemmerEvaluatorV3.jar Matching.Stemmer2UI
```

(the package name is Matching and the internal class name of “trial stemmer builder” is Stemmer2UI)

After issuing the above command we will see something similar with the following screenshot:


Configure and Run the Dynamic Stemmer
Wizard
Config Stemmer using the above
Do Dynamic Stemming
Save Config

□ Split Couples

nnk's stems and arguments [Expert:2]
stamou's stems and arguments [Expert:3]
galiotou's stems and arguments [Expert:4]

At Least Remain Letters 3

Minimum Word Letters To Apply Stemming 5

Suffix	Step1 ...	Step1 Rplc	Step2 ...	Step2 Rplc	Comment
KIHËSHIN	☑		□		
QOFSHIM	☑		□		
QOFSHIN	☑		□		
ËMËSIVE	☑		□		
ËTARËVE	☑		□		
HESHIM	☑		□		
HESHIN	☑		□		
HESHIT	☑		□		
IMISHT	☑		□		
ISTRIT	☑		□		
KISHIM	☑		□		
KISHIN	☑		□		
KISHIT	☑		□		

replace this with a brief description of the stemmer configuration

In order to configure a trial semmer the user has to:

- enable/disable the “split couples” (SC) parameter,
- define the value of parameter “At least remain letters” (RL),
- define the value of parameter “Minimum word length to apply stemming” (MWL),
- enable/disable the available suffixes for the first and for the second removal step.

The last step can be done by the provided wizard. In order to run the wizard, the user has to select (from the list above button “Wizard”) an expert or a group of experts and next press the button “Wizard”. The wizard will automatically enable/disable suffixes in order to make the trial stemmer’s result to be compliant (as much as possible) with the selected expert or group of experts. After the configuration the user has to save the results.

Saving the results is an easy process. The user has to follow (in order presented):

- press the button “Config stemmer using the above”,
- type a name for the trial stemmer in the text box,
- press the button “Do dynamic stemming” (it takes some minutes because it updates the db),
- press the button “Save config”.

The first one (“Config stemmer using the above”) put the values from the interface items into internal program variable. The third one (“Do dynamic stemming”) applies the new trial stemmer in each word and save the stemmer results in the database. This is the reason that the third step takes some minutes. The last step (“Save config”) saves the configuration in a text file in order to be used later for automatic code creation (by code builder). Consider that the name of trial stemmer can have letters, digits and spaces and should start with letter. This is because this name will become the name of the class (java source code) that the code builder will produce.

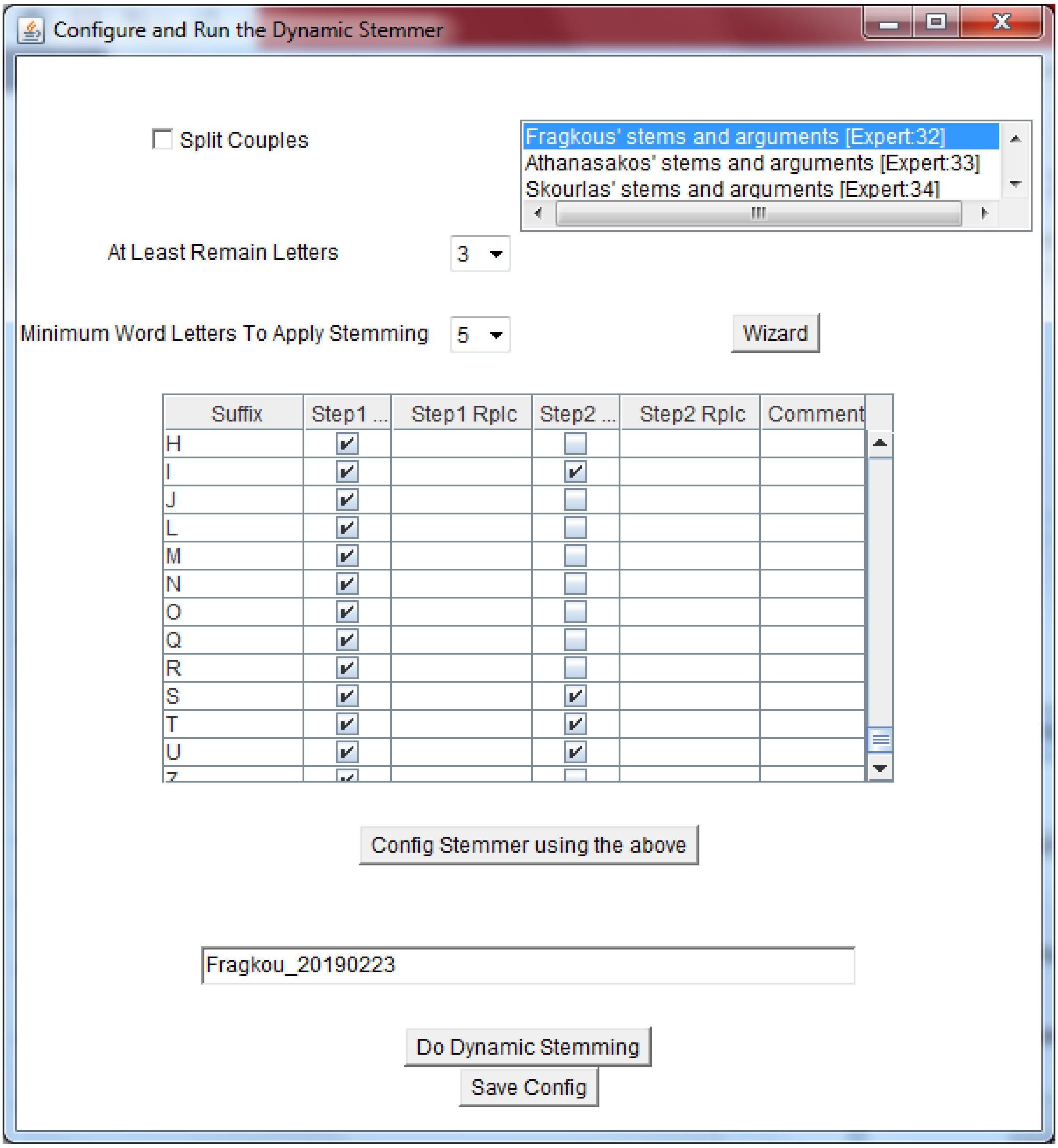
Example: Forming (configuring) a trial stemmer to be compliant with some expert’s arguments

In the following screenshot, we are configuring the trial stemmer. As you can see the configuration is done by disabling SC, set RL:3, set MWL:5 and running the wizard to be compliant with “Fragkou’s stems and arguments”). Next we have to:

- Press “Config Stemmer using the above”,
- Name the stemmer: Fragkou_20190223,
- Press “Do Dynamic Stemming”,
- Press “Save Config”.

The third step enters stems produced by this trial stemmer into table “stems” of the database.

The fourth step creates a file with name “C:\stemmerSuite\Alban1\Fragkou_20190223.txt” (assume that the selected language is Alban1).


Configure and Run the Dynamic Stemmer
Wizard
Config Stemmer using the above
Do Dynamic Stemming
Save Config

□ Split Couples

Fragkous' stems and arguments [Expert:32]
Athanasakos' stems and arguments [Expert:33]
Skourlas' stems and arguments [Expert:34]

At Least Remain Letters 3

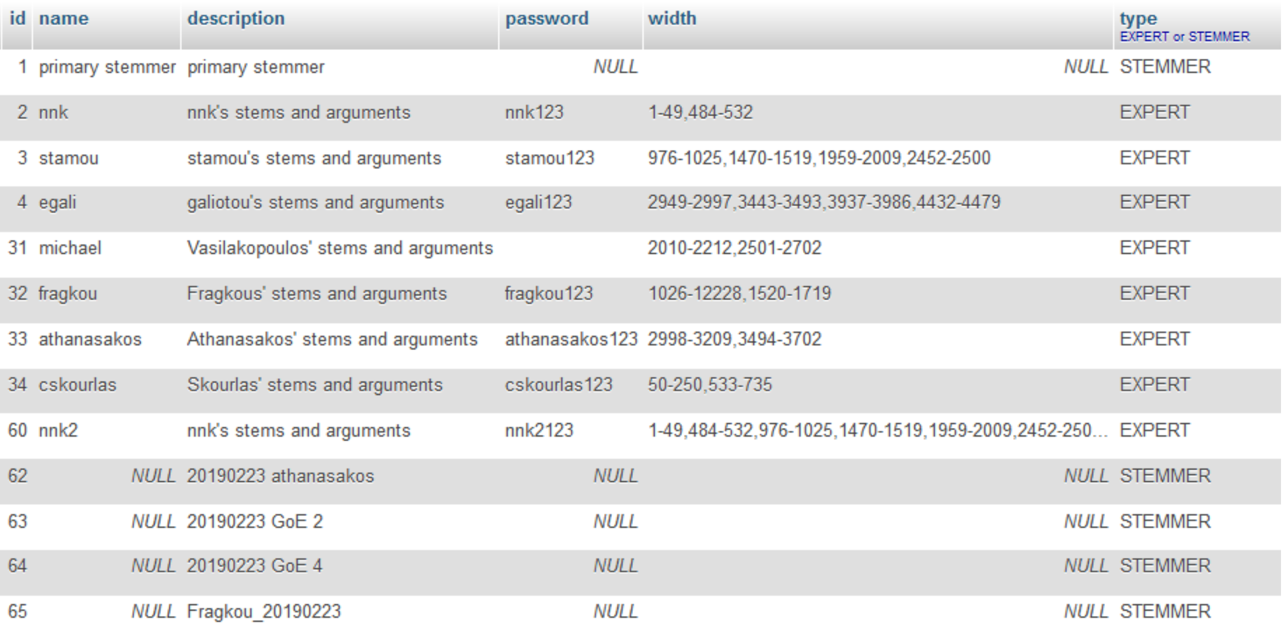
Minimum Word Letters To Apply Stemming 5

Suffix	Step1 ...	Step1 Rplc	Step2 ...	Step2 Rplc	Comment
H	☑		□		
I	☑		☑		
J	☑		□		
L	☑		□		
M	☑		□		
N	☑		□		
O	☑		□		
Q	☑		□		
R	☑		□		
S	☑		☑		
T	☑		☑		
U	☑		☑		
Z	☑				

Fragkou_20190223

The first next picture present the content of table sources after clicking the button “Do Dynamic Stemming”. The Last row of this table (sources) is about the newly created trial stemmer.

The second next picture has the number of rows in each table in the database. We can see that table “stems” has 25180 rows and table “words” has 5036 rows. This is because in the depicted language (Alban1) we have a set of 5036 distinct words and the present example of trial stemmer were the fifth (actually the primary stemmer and four trial ones, 25180 = 5 x 5036).



id	name	description	password	width	type EXPERT or STEMMER
1	primary stemmer	primary stemmer	NULL	NULL	STEMMER
2	nnk	nnk's stems and arguments	nnk123	1-49,484-532	EXPERT
3	stamou	stamou's stems and arguments	stamou123	976-1025,1470-1519,1959-2009,2452-2500	EXPERT
4	egali	galiotou's stems and arguments	egali123	2949-2997,3443-3493,3937-3986,4432-4479	EXPERT
31	michael	Vasilakopoulos' stems and arguments		2010-2212,2501-2702	EXPERT
32	fragkou	Fragkous' stems and arguments	fragkou123	1026-12228,1520-1719	EXPERT
33	athanasakos	Athanasakos' stems and arguments	athanasakos123	2998-3209,3494-3702	EXPERT
34	cskourlas	Skourlas' stems and arguments	cskourlas123	50-250,533-735	EXPERT
60	nnk2	nnk's stems and arguments	nnk2123	1-49,484-532,976-1025,1470-1519,1959-2009,2452-250...	EXPERT
62	NULL	20190223 athanasakos	NULL	NULL	STEMMER
63	NULL	20190223 GoE 2	NULL	NULL	STEMMER
64	NULL	20190223 GoE 4	NULL	NULL	STEMMER
65	NULL	Fragkou_20190223	NULL	NULL	STEMMER

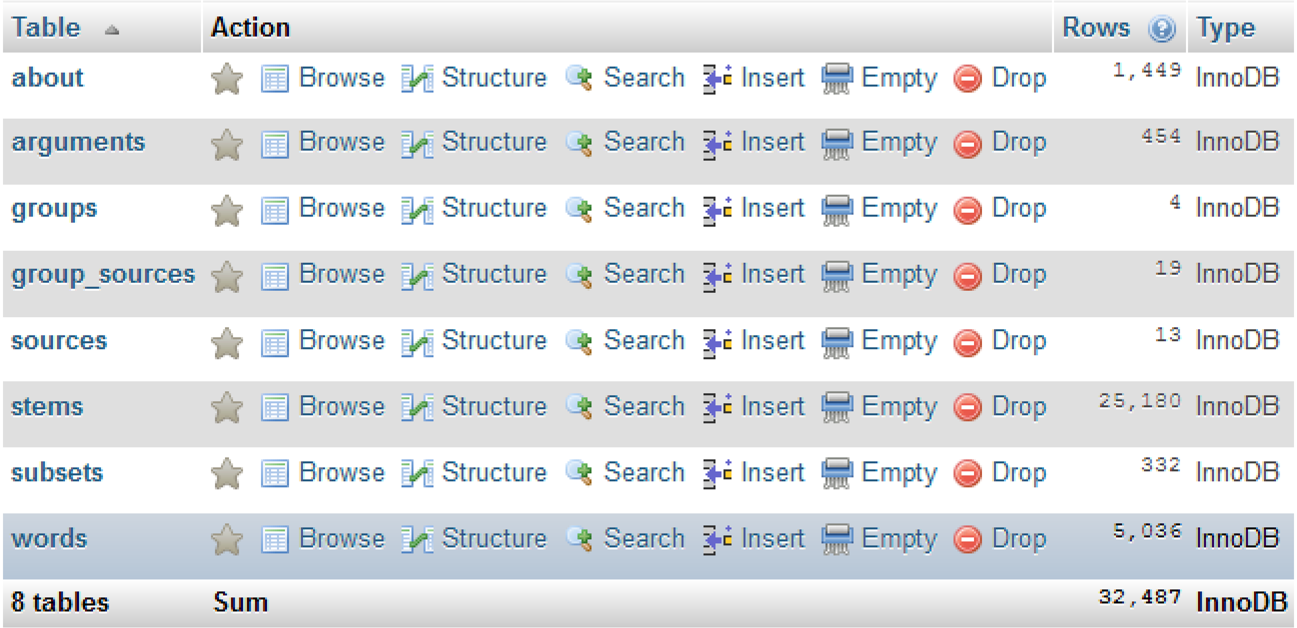
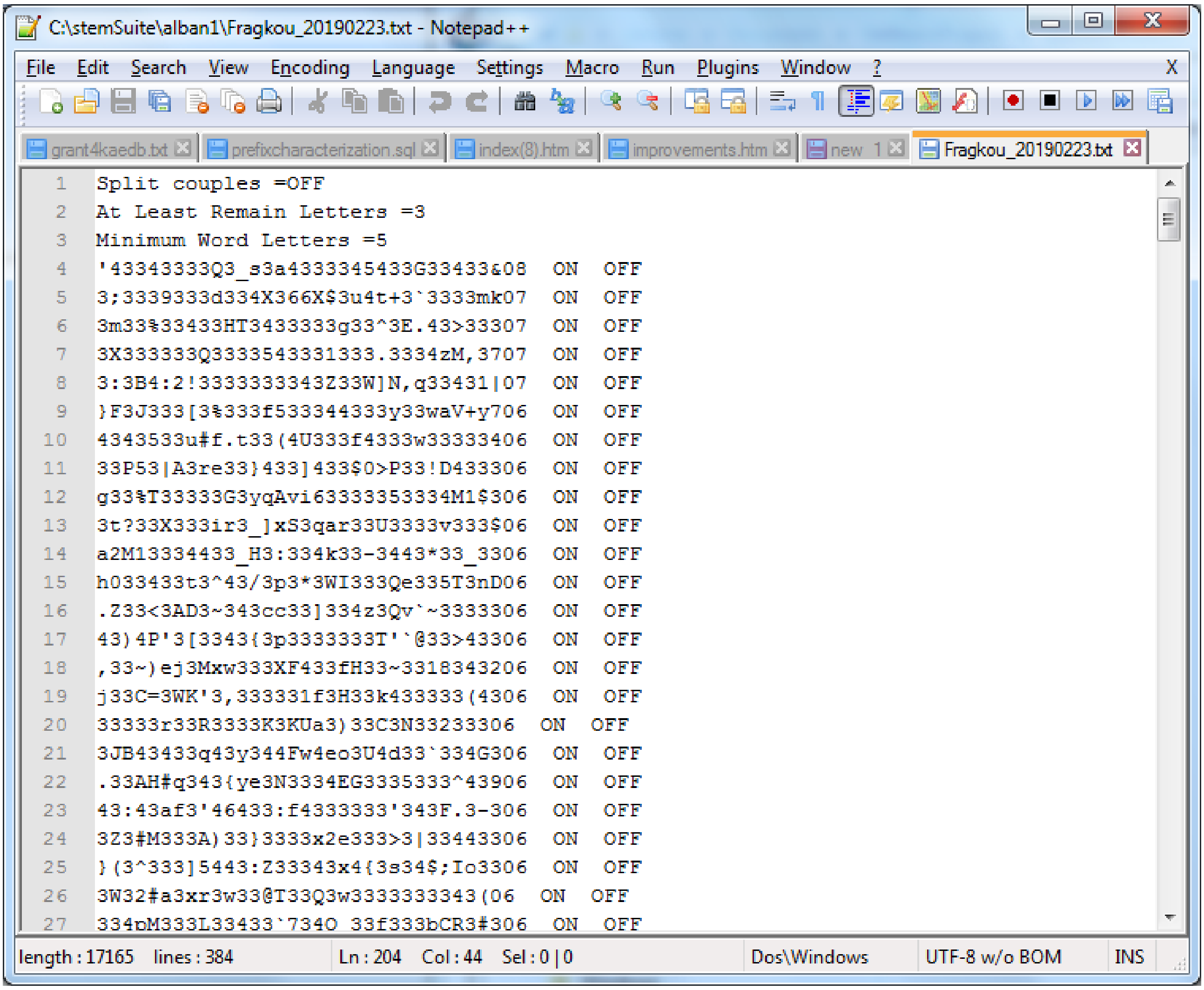


Table	Action	Rows	Type
about	Browse Structure Search Insert Empty Drop	1,449	InnoDB
arguments	Browse Structure Search Insert Empty Drop	454	InnoDB
groups	Browse Structure Search Insert Empty Drop	4	InnoDB
group_sources	Browse Structure Search Insert Empty Drop	19	InnoDB
sources	Browse Structure Search Insert Empty Drop	13	InnoDB
stems	Browse Structure Search Insert Empty Drop	25,180	InnoDB
subsets	Browse Structure Search Insert Empty Drop	332	InnoDB
words	Browse Structure Search Insert Empty Drop	5,036	InnoDB
8 tables	Sum	32,487	InnoDB

The configuration of trial stemmer is saved with SHA256 encrypted suffixes

As we alredy have said, the “Save config” step creates a file with the given name inside the folder “C:\stemmerSuite\Alban1\” (the subfolder – Alban1 – can be the currently selected language). In the last example we have created the ”Fragkou_20190223.txt” configuration file. The configuration file contains the 3 basic configuration parameters (SC, RL, MWL) and encrypted versions of suffixes flagged with ON|OFF for each step. The following screenshot contains an excerpt of the configuration file created in the last example:


C:\stemSuite\alban1\Fragkou_20190223.txt - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
grant4kaedb.txt
prefixcharacterization.sql
index(8).htm
improvements.htm
new 1
Fragkou_20190223.txt
length : 17165 lines : 384
Ln : 204 Col : 44 Sel : 0 | 0
Dos\Windows
UTF-8 w/o BOM
INS

```
Split couples =OFF
At Least Remain Letters =3
Minimum Word Letters =5
'43343333Q3_s3a4333345433G33433&08  ON  OFF
3;3339333d334X366X$3u4t+3`3333mk07  ON  OFF
3m33%33433HT3433333g33^3E.43>33307  ON  OFF
3X333333Q33335433331333.3334zM,3707  ON  OFF
3:3B4:2!333333333433Z33W]N,q33431|07  ON  OFF
}F3J333[3%333f533344333y33waV+y706  ON  OFF
4343533u#f.t33(4U333f4333w33333406  ON  OFF
33P53|A3re33}433]433$0>P33!D433306  ON  OFF
g33%T33333G3yqAvi6333335333 4M1$306  ON  OFF
3t?33X333ir3_]xS3qar33U3333v333$06  ON  OFF
a2M13334433_H3:334k33-3443*33_3306  ON  OFF
h033433t3^43/3p3*3WI333Qe335T3nD06  ON  OFF
.Z33<3AD3~343cc33]334z3Qv`~3333306  ON  OFF
43)4P'3[3343{3p3333333T'`@33>43306  ON  OFF
,33~)ej3Mxw333XF433fH33~3318343206  ON  OFF
j33C=3WK'3,333331f3H33k433333(4306  ON  OFF
33333r33R3333K3KUa3)33C3N33233306  ON  OFF
3JB43433q43y344Fw4eo3U4d33`334G306  ON  OFF
.33AH#q343{ye3N3334EG3335333^43906  ON  OFF
43:43af3'46433:f4333333'343F.3-306  ON  OFF
3Z3#M333A)33}3333x2e333>3|33443306  ON  OFF
}(3^333]5443:Z33343x4{3s34$;Io3306  ON  OFF
3W32#a3xr3w33@T33Q3w3333333343(06  ON  OFF
334pM333L33433`734O 33f333bCR3#306  ON  OFF
```

Evaluating a trial stemmer

We can form (configure) more than one trial stemmers. Next we have to decide which one is the best one for production (to be used in some search engine or in some summarizer or in some text classifier and so on). For this reason we have implemented the Stemmer Evaluator.

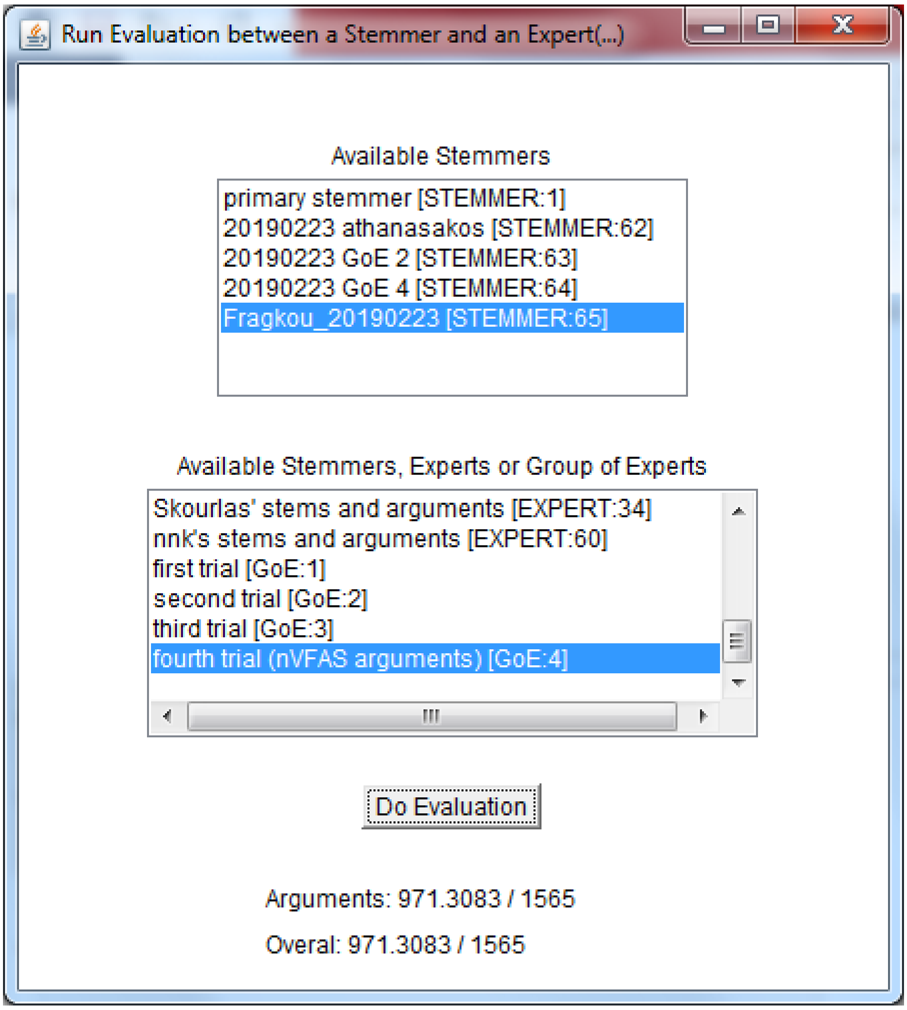
To invoke the “stemmer evaluator” we have to issue the command:

C:\stemSuite\bin> **java -classpath StemmerEvaluatorV3.jar Matching. EvaluatorUI**

(the project name is StemmerEvaluatorV3, the package name is Matching and the internal class name of “stemmer evaluator” is EvaluatorUI)

Example: evaluating a trial Stemmer

We will describe the evaluation of the stemmer produced previously. We will measure how much compliant is the stemmer configured to comply with Fragkou’s arguments (Fragkou_20190223) against the total set of arguments (nnk, Vasilakopoulos, Athanasakos, Fragkou, Skourlas – shortly nVFAS). The next screenshot is the stemmer evaluator in action:


Run Evaluation between a Stemmer and an Expert(...)
Available Stemmers

primary stemmer [STEMMER:1]
20190223 athanasakos [STEMMER:62]
20190223 GoE 2 [STEMMER:63]
20190223 GoE 4 [STEMMER:64]
Fragkou_20190223 [STEMMER:65]

Available Stemmers, Experts or Group of Experts

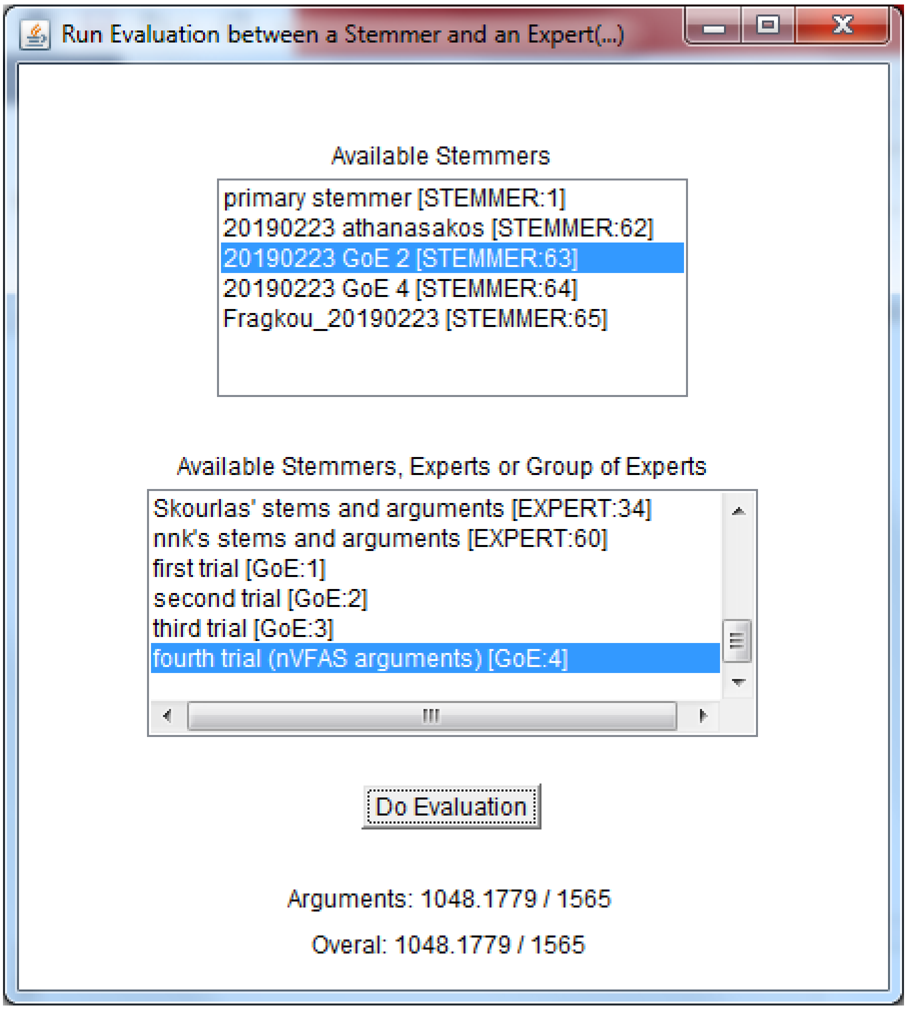
Skourlas' stems and arguments [EXPERT:34]
nnk's stems and arguments [EXPERT:60]
first trial [GoE:1]
second trial [GoE:2]
third trial [GoE:3]
fourth trial (nVFAS arguments) [GoE:4]

Arguments: 971.3083 / 1565

Overal: 971.3083 / 1565

The result is 971,31 / 1565.

Next screenshot is another trial stemmer's evaluation. We are measuring how much compliant is the stemmer configured to comply with the arguments of a group of experts (configuration file 20190223_GoE_2.txt) against the total set of arguments (nnk, Vasilakopoulos, Athanasakos, Fragkou, Skourlas – shortly nVFAS). It is obvious that this stemmer (GoE_2) is a better one than the previous.


Run Evaluation between a Stemmer and an Expert(...)
Available Stemmers
primary stemmer [STEMMER:1]
20190223 athanasakos [STEMMER:62]
20190223 GoE 2 [STEMMER:63]
20190223 GoE 4 [STEMMER:64]
Fragkou_20190223 [STEMMER:65]
Available Stemmers, Experts or Group of Experts
Skourlas' stems and arguments [EXPERT:34]
nnk's stems and arguments [EXPERT:60]
first trial [GoE:1]
second trial [GoE:2]
third trial [GoE:3]
fourth trial (nVFAS arguments) [GoE:4]
Do Evaluation

Arguments: 1048.1779 / 1565

Overal: 1048.1779 / 1565

Code Builder

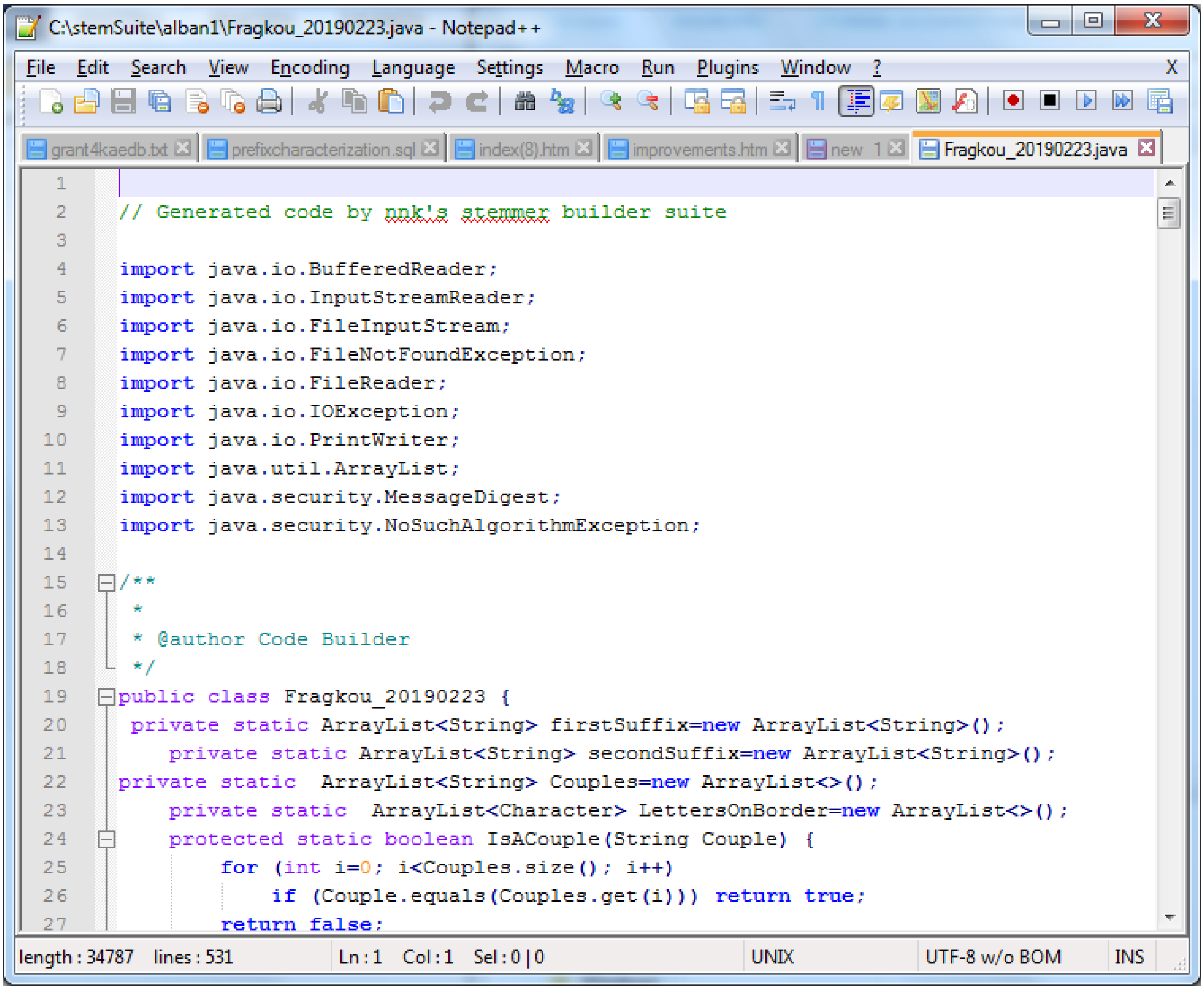
Code Builder gets a command line argument (which is the configuration file without the extension .txt) and produces java code. The configuration file should be inside the selected language subfolder, under the C:\stemSuite\ basic folder.

As an example we will assume that we prefer the stemmer build according the Fragkou's arguments. From a command prompt and having access to the code builder (CodeBuilder.jar) we have to issue the command:

C:\stemSuite\bin> **java -jar CodeBuilder.jar Fragkou_20190223**

If we assume that the selected language is Alban1, the command reads the configuration file C:\stemSuite\alban1\Fragkou_20190223.txt and write/create the java source code C:\stemSuite\alban1\Fragkou_20190223.java.

An excerpt of the produced java code (Fragkou_20190223.java) is presented in the next screenshot:


C:\stemSuite\alban1\Fragkou_20190223.java - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
grant4kaedb.txt
prefixcharacterization.sql
index(8).htm
improvements.htm
new 1
Fragkou_20190223.java
length : 34787 lines : 531
Ln : 1 Col : 1 Sel : 0 | 0
UNIX
UTF-8 w/o BOM
INS

```

// Generated code by nnk's stemmer builder suite

import java.io.BufferedReader;
import java.io.InputStreamReader;
import java.io.FileInputStream;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.IOException;
import java.io.PrintWriter;
import java.util.ArrayList;
import java.security.MessageDigest;
import java.security.NoSuchAlgorithmException;

/**
 *
 * @author Code Builder
 */
public class Fragkou_20190223 {
 private static ArrayList<String> firstSuffix=new ArrayList<String>();
    private static ArrayList<String> secondSuffix=new ArrayList<String>();
private static  ArrayList<String> Couples=new ArrayList<>();
    private static  ArrayList<Character> LettersOnBorder=new ArrayList<>();
    protected static boolean IsACouple(String Couple) {
        for (int i=0; i<Couples.size(); i++)
            if (Couple.equals(Couples.get(i))) return true;
        return false;
```

Compiling the source code

The source code can now be turned to an executable program. To do so open a command prompt, change directory to the language of interest and issue the compilation command, as following:

C:\> **cd stemsuite**

C:\stemsuite> **cd alban1**

C:\stemsuite\alban1> **javac –encoding UTF-8 Fragkou_2019023.java**

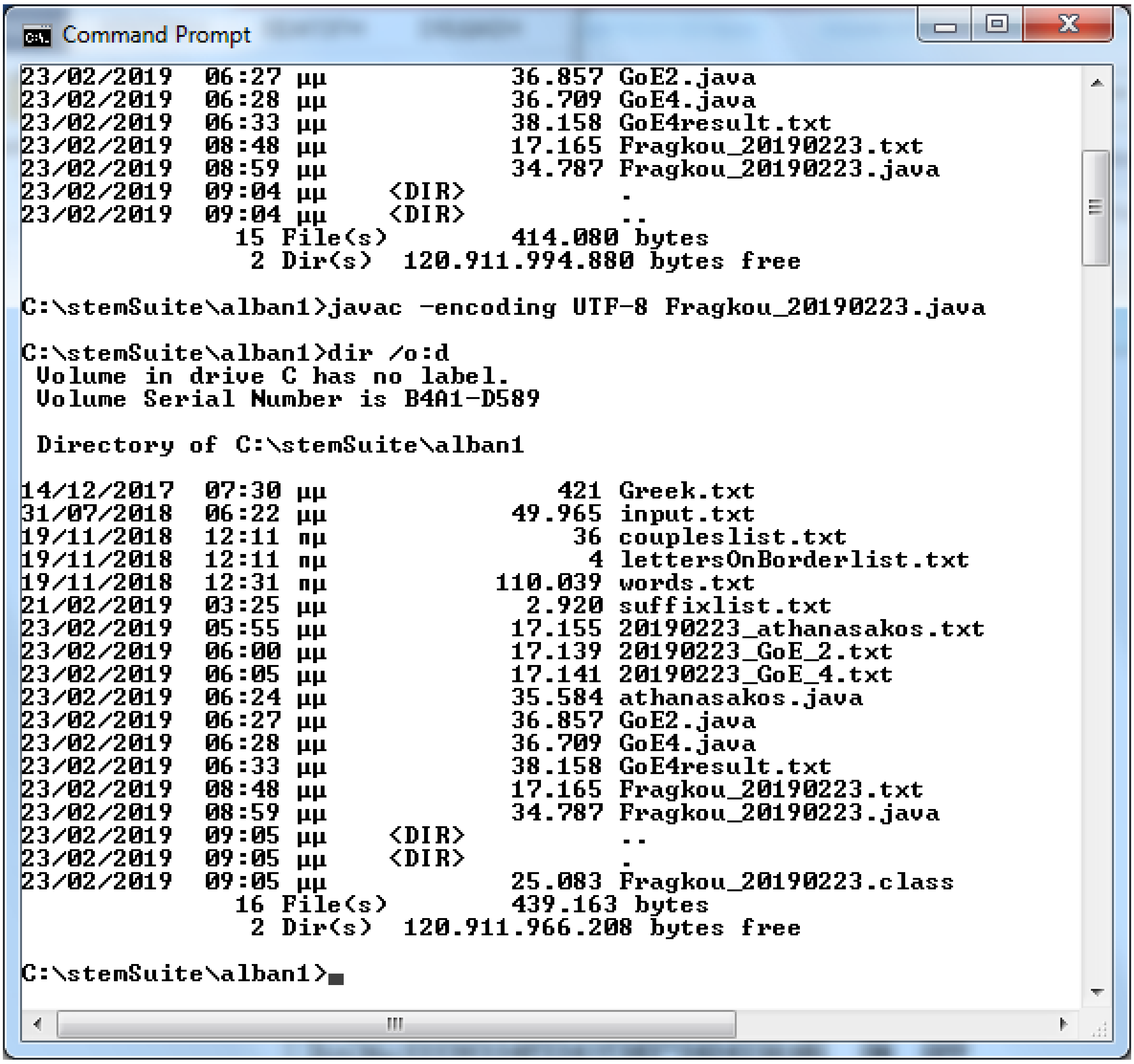
If no mistakes, our folder (for the selected language) will contain the config, the source and the executable files. For the example we have followed, we are expecting to see:

The Config file: Fragkou_20190223.txt

The Source code file: Fragkou_20190223.java

The Executable file: Fragkou_20190223.class

Next screenshot contains the compilation command and a listing (dir) command which displays all the expected to see files:



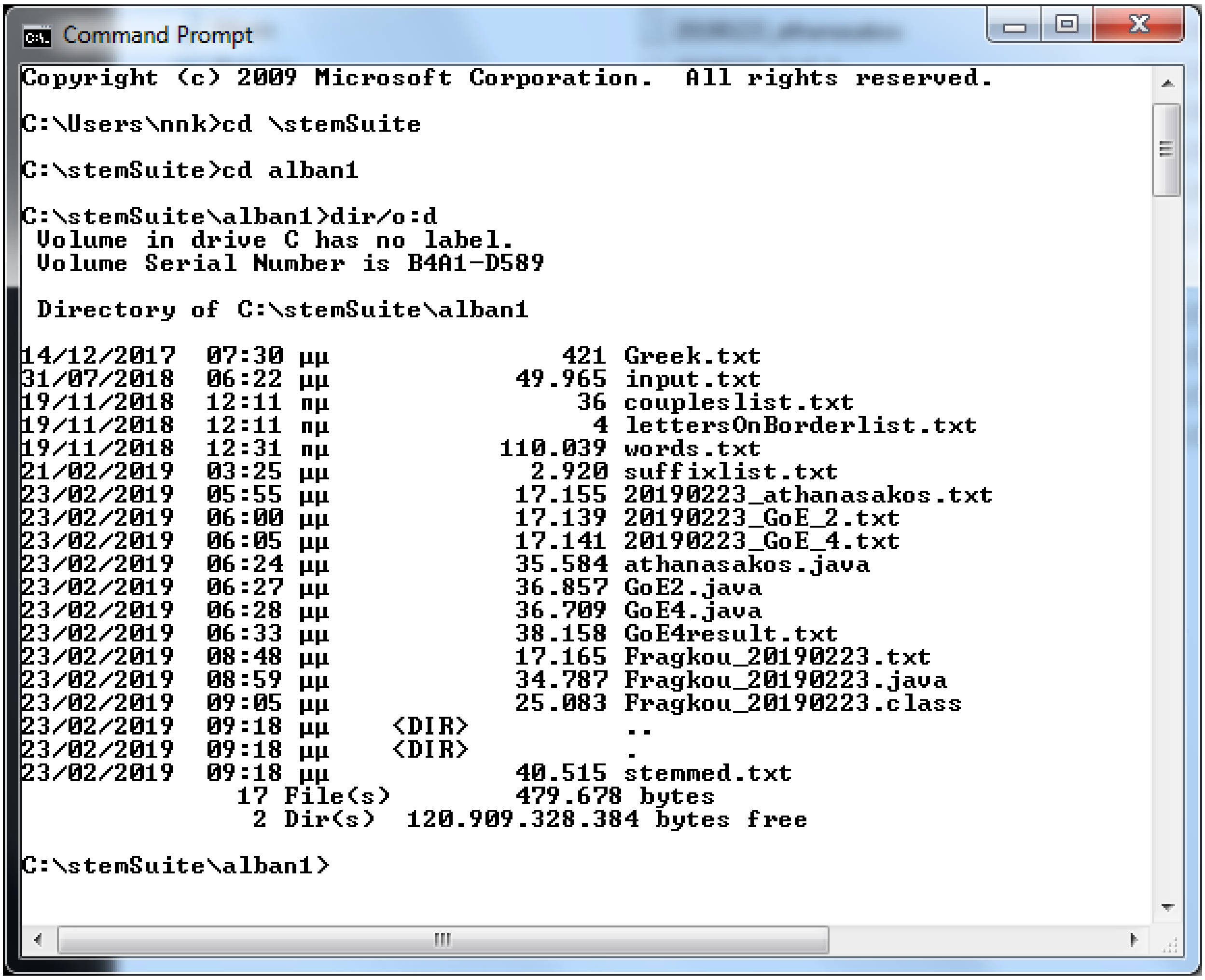
```
23/02/2019  06:27 μμ            36.857 GoE2.java
23/02/2019  06:28 μμ            36.709 GoE4.java
23/02/2019  06:33 μμ            38.158 GoE4result.txt
23/02/2019  08:48 μμ            17.165 Fragkou_20190223.txt
23/02/2019  08:59 μμ            34.787 Fragkou_20190223.java
23/02/2019  09:04 μμ    <DIR>          .
23/02/2019  09:04 μμ    <DIR>          ..
              15 File(s)        414.080 bytes
               2 Dir(s)  120.911.994.880 bytes free

C:\stemSuite\alban1>javac -encoding UTF-8 Fragkou_20190223.java

C:\stemSuite\alban1>dir /o:d
 Volume in drive C has no label.
 Volume Serial Number is B4A1-D589

 Directory of C:\stemSuite\alban1

14/12/2017  07:30 μμ               421 Greek.txt
31/07/2018  06:22 μμ            49.965 input.txt
19/11/2018  12:11 πμ                36 coupleslist.txt
19/11/2018  12:11 πμ                 4 lettersOnBorderlist.txt
19/11/2018  12:31 πμ           110.039 words.txt
21/02/2019  03:25 μμ             2.920 suffixlist.txt
23/02/2019  05:55 μμ            17.155 20190223_athanasakos.txt
23/02/2019  06:00 μμ            17.139 20190223_GoE_2.txt
23/02/2019  06:05 μμ            17.141 20190223_GoE_4.txt
23/02/2019  06:24 μμ            35.584 athanasakos.java
23/02/2019  06:27 μμ            36.857 GoE2.java
23/02/2019  06:28 μμ            36.709 GoE4.java
23/02/2019  06:33 μμ            38.158 GoE4result.txt
23/02/2019  08:48 μμ            17.165 Fragkou_20190223.txt
23/02/2019  08:59 μμ            34.787 Fragkou_20190223.java
23/02/2019  09:05 μμ    <DIR>          ..
23/02/2019  09:05 μμ    <DIR>          .
23/02/2019  09:05 μμ            25.083 Fragkou_20190223.class
              16 File(s)        439.163 bytes
               2 Dir(s)  120.911.966.208 bytes free

C:\stemSuite\alban1>
```

Running the executable stemmer

The executable stemmer can now be used to stem the words of any document in the language of interest. If we assume that the input text is “input.txt” and the result (the stems) we want to be saved in a file named “stemmed.txt“, we have to issue the following command:

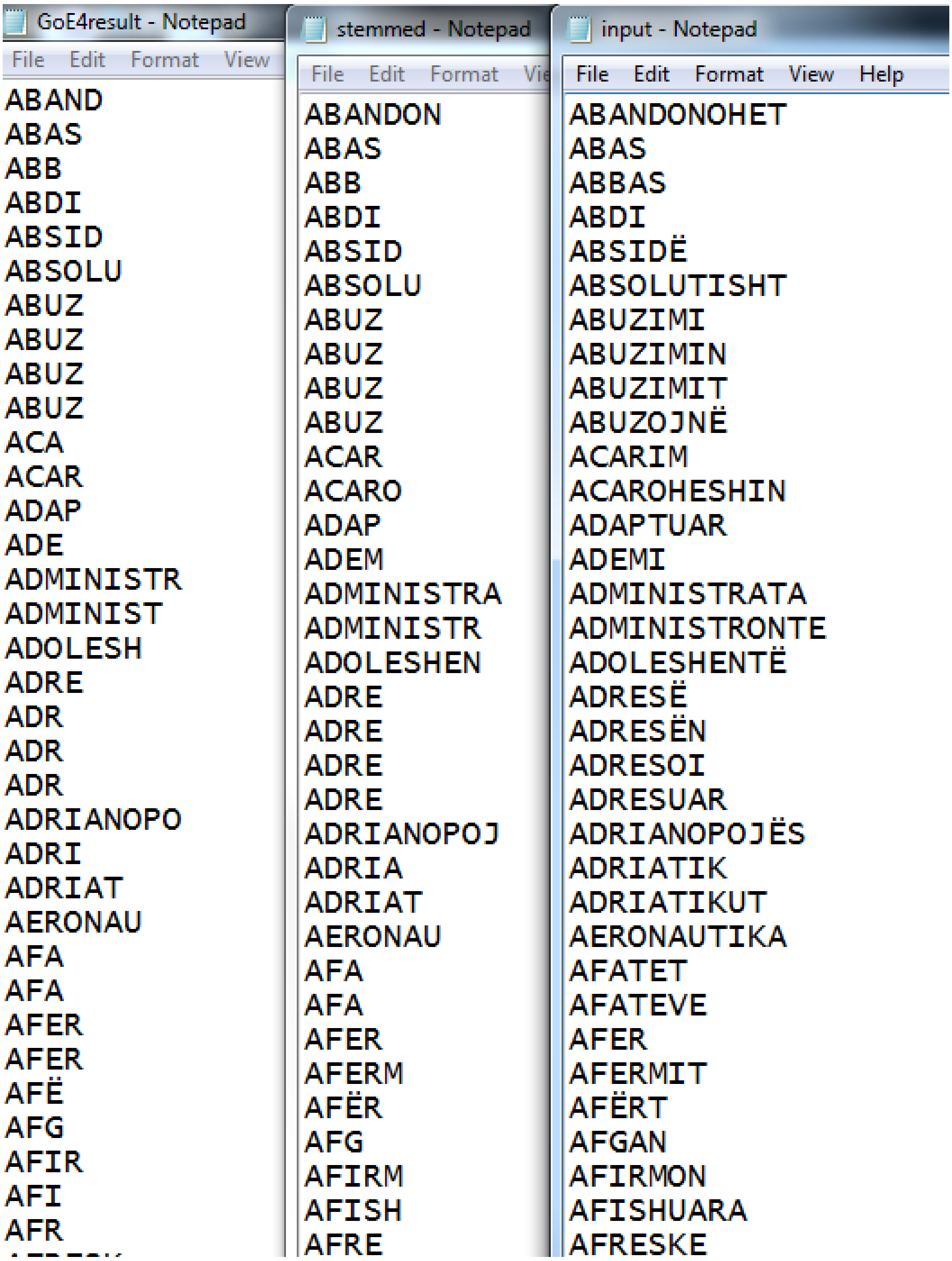
C:\stemsuite\alban1> **java Fragkou_20190223 input.txt stemmed.txt**

Next screenshot is a listing (dir) which contains also the results (stemmed.txt):



```
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\nnk>cd \stemSuite

C:\stemSuite>cd alban1

C:\stemSuite\alban1>dir/o:d
 Volume in drive C has no label.
 Volume Serial Number is B4A1-D589

 Directory of C:\stemSuite\alban1

14/12/2017  07:30 μμ               421 Greek.txt
31/07/2018  06:22 μμ            49.965 input.txt
19/11/2018  12:11 πμ                36 coupleslist.txt
19/11/2018  12:11 πμ                 4 lettersOnBorderlist.txt
19/11/2018  12:31 πμ           110.039 words.txt
21/02/2019  03:25 μμ             2.920 suffixlist.txt
23/02/2019  05:55 μμ            17.155 20190223_athanasakos.txt
23/02/2019  06:00 μμ            17.139 20190223_GoE_2.txt
23/02/2019  06:05 μμ            17.141 20190223_GoE_4.txt
23/02/2019  06:24 μμ            35.584 athanasakos.java
23/02/2019  06:27 μμ            36.857 GoE2.java
23/02/2019  06:28 μμ            36.709 GoE4.java
23/02/2019  06:33 μμ            38.158 GoE4result.txt
23/02/2019  08:48 μμ            17.165 Fragkou_20190223.txt
23/02/2019  08:59 μμ            34.787 Fragkou_20190223.java
23/02/2019  09:05 μμ            25.083 Fragkou_20190223.class
23/02/2019  09:18 μμ    <DIR>          ..
23/02/2019  09:18 μμ    <DIR>          .
23/02/2019  09:18 μμ            40.515 stemmed.txt
              17 File(s)        479.678 bytes
               2 Dir(s)  120.909.328.384 bytes free

C:\stemSuite\alban1>
```

Two stemmed documents (by different stemmers) and the original (before stemming)



GoE4result - Notepad	stemmed - Notepad	input - Notepad
ABAND	ABANDON	ABANDONOHET
ABAS	ABAS	ABAS
ABB	ABB	ABBAS
ABDI	ABDI	ABDI
ABSID	ABSID	ABSIDË
ABSOLU	ABSOLU	ABSOLUTISHT
ABUZ	ABUZ	ABUZIMI
ABUZ	ABUZ	ABUZIMIN
ABUZ	ABUZ	ABUZIMIT
ABUZ	ABUZ	ABUZOJNË
ACA	ACAR	ACARIM
ACAR	ACARO	ACAROHESHIN
ADAP	ADAP	ADAPTUAR
ADE	ADEM	ADEMI
ADMINISTR	ADMINISTRA	ADMINISTRATA
ADMINIST	ADMINISTR	ADMINISTRONTE
ADOLESH	ADOLESHEN	ADOLESHENTË
ADRE	ADRE	ADRESË
ADR	ADRE	ADRESËN
ADR	ADRE	ADRESOI
ADR	ADRE	ADRESUAR
ADRIANOPO	ADRIANOPOJ	ADRIANOPOJËS
ADRI	ADRIA	ADRIATIK
ADRIAT	ADRIAT	ADRIATIKUT
AERONAU	AERONAU	AERONAUTIKA
AFA	AFA	AFATET
AFA	AFA	AFATEVE
AFER	AFER	AFER
AFER	AFERM	AFERMIT
AFË	AFËR	AFËRT
AFG	AFG	AFGAN
AFIR	AFIRM	AFIRMON
AFI	AFISH	AFISHUARA
AFR	AFRE	AFRESKE

A common mistake

We have said that the name of trial stemmer can have letters, digits and spaces and should start with letter. This is because this name will become the name of the class (java source code) that the code builder will produce. However, if we make the mistake and provide a name for the trial stemmer which is not valid for a java class, we can fix it. For example, we have seen earlier, that stemmer with id 63 is named "20190223 GoE 2" (see above the screenshot with table "sources"). The equivalent configuration file is

“20190223_GoE_2.txt” (because the “Save Config” button of Stemmer Builder replace spaces with underscores). In this case, the code builder is invoked with the next command:

```
C:\stemSuite\bin> java -jar CodeBuilder.jar 20190223_GoE_2.txt
```

And the result (the product of code builder) is the java source file 20190223_GoE_2.java

If you try to compile this file you will get an error message. This is because the file 20190223_GoE_2.java contains a class named 20190223_GoE_2 which is an invalid class name. The solution is to rename the class and the file to something acceptable. For example rename the file to “GoE2.java”, edit the file and change

```
public class 20190223_GoE_2 {
```

to

```
public class GoE2 {
```

Belgrade, January, 2020
Nikitas N. Karanikolas