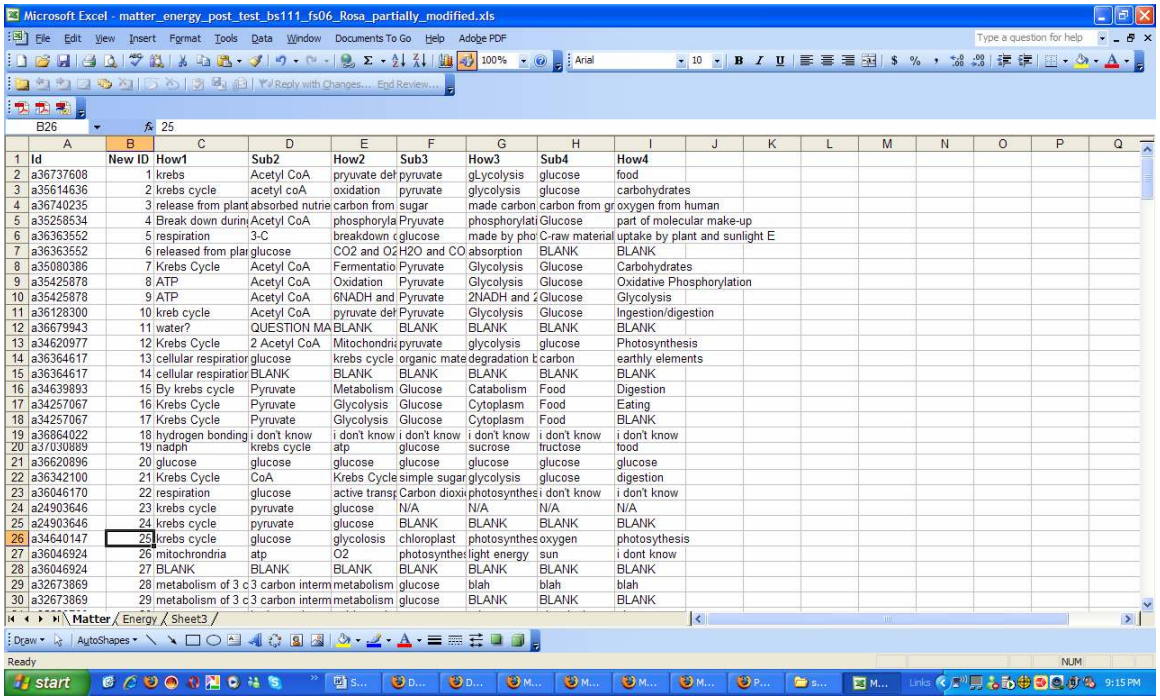
ANALYZING OPEN ANSWER QUESTIONS USING PASW TEXT ANALYSIS FOR SURVEYS, VERSION 3.0 PART I

PASW Text Analysis for Surveys (PTAfS), former SPSS Text Analysis for Surveys, is a program that allows classifying open-ended responses in a set of categories based on words and phrases that are recognized by the software. The process of recognizing words is referred to as '*extraction*' and the words and phrases are referred to as '*terms'*. The linguistic machinery of the software will put together terms that have similar meaning and will refer to them as '*concepts*'. This software has a great potential to be applied in the assessment of students learning in large-enrollment classes. The following steps will help a first-time user to start working with the software. Additional explanations for particular procedures can be found in either the User's Guide or by clicking the Help button on the software.

1. Preparing data file:
   Usually data will be collected in an excel datasheet. This datasheet must have a column with a unique ID (such as PID, or MSUNET ID) for each student and the answers for each question should be in the next columns:

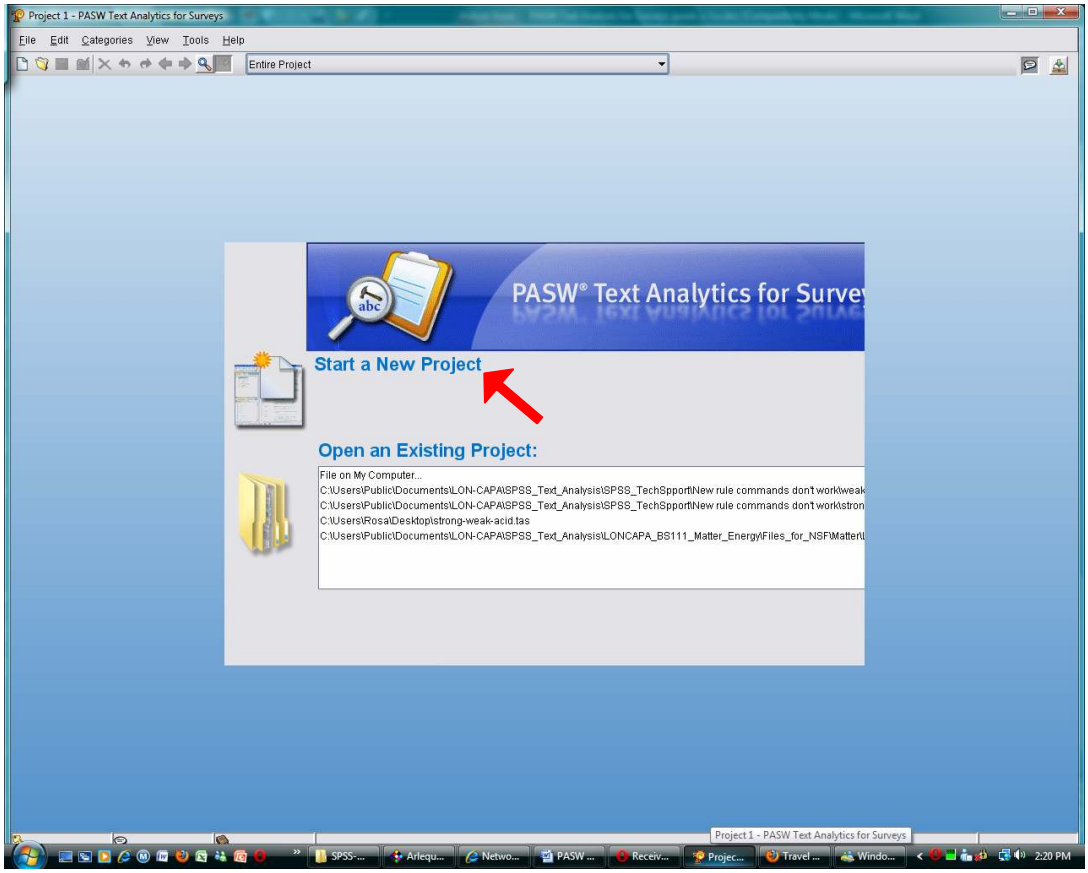


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Id	New ID	How1	Sub2	How2	Sub3	How3	Sub4	How4								
2	a36737608	1	krebs	Acetyl CoA	pryuvate def	pyruvate	gLycolysis	glucose	food								
3	a35614636	2	krebs cycle	acetyl coA	oxidation	pyruvate	glycolysis	glucose	carbohydrates								
4	a36740235	3	release from plant	absorbed nutrie	carbon from	sugar	made carbon	carbon from gr	oxygen from human								
5	a35258534	4	Break down durin	Acetyl CoA	phosphoryla	Pryuvate	phosphorylati	Glucose	part of molecular make-up								
6	a36363552	5	respiration	3-C	breakdown c	glucose	made by pho	C-raw material	uptake by plant and sunlight E								
7	a36363552	6	released from plar	glucose	CO2 and O2	H2O and CO	absorption	BLANK	BLANK								
8	a35080386	7	Krebs Cycle	Acetyl CoA	Fermentatio	Pyruvate	Glycolysis	Glucose	Carbohydrates								
9	a35425878	8	ATP	Acetyl CoA	Oxidation	Pyruvate	Glycolysis	Glucose	Oxidative Phosphorylation								
10	a35425878	9	ATP	Acetyl CoA	6NADH and	Pyruvate	2NADH and 2	Glucose	Glycolysis								
11	a36128300	10	kreb cycle	Acetyl CoA	pyruvate def	Pyruvate	Glycolysis	Glucose	Ingestion/digestion								
12	a36679943	11	water?	QUESTION MA	BLANK	BLANK	BLANK	BLANK	BLANK								
13	a34620977	12	Krebs Cycle	2 Acetyl CoA	Mitochondri	pyruvate	glycolysis	glucose	Photosynthesis								
14	a36364617	13	cellular respiratior	glucose	krebs cycle	organic mate	degradation t	carbon	earthly elements								
15	a36364617	14	cellular respiratior	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK								
16	a34639893	15	By krebs cycle	Pyruvate	Metabolism	Glucose	Catabolism	Food	Digestion								
17	a34257067	16	Krebs Cycle	Pyruvate	Glycolysis	Glucose	Cytoplasm	Food	Eating								
18	a34257067	17	Krebs Cycle	Pyruvate	Glycolysis	Glucose	Cytoplasm	Food	BLANK								
19	a36864022	18	hydrogen bonding	i don't know	i don't know	i don't know	i don't know	i don't know	i don't know								
20	a37030889	19	nadph	krebs cycle	atp	glucose	sucrose	fructose	food								
21	a36620896	20	glucose	glucose	glucose	glucose	glucose	glucose	glucose								
22	a36342100	21	Krebs Cycle	CoA	Krebs Cycle	simple sugar	glycolysis	glucose	digestion								
23	a36046170	22	respiration	glucose	active transp	Carbon dioxi	photosynthes	i don't know	i don't know								
24	a24903646	23	krebs cycle	pyruvate	glucose	N/A	N/A	N/A	N/A								
25	a24903646	24	krebs cycle	pyruvate	glucose	BLANK	BLANK	BLANK	BLANK								
26	a34640147	25	krebs cycle	glucose	glycolosis	chloroplast	photosynthes	oxygen	photosythesis								
27	a36046924	26	mitochrondria	atp	O2	photosynthes	light energy	sun	i dont know								
28	a36046924	27	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK	BLANK								
29	a32673869	28	metabolism of 3 c	3 carbon interm	metabolism	glucose	blah	blah	blah								
30	a32673869	29	metabolism of 3 c	3 carbon interm	metabolism	glucose	BLANK	BLANK	BLANK								

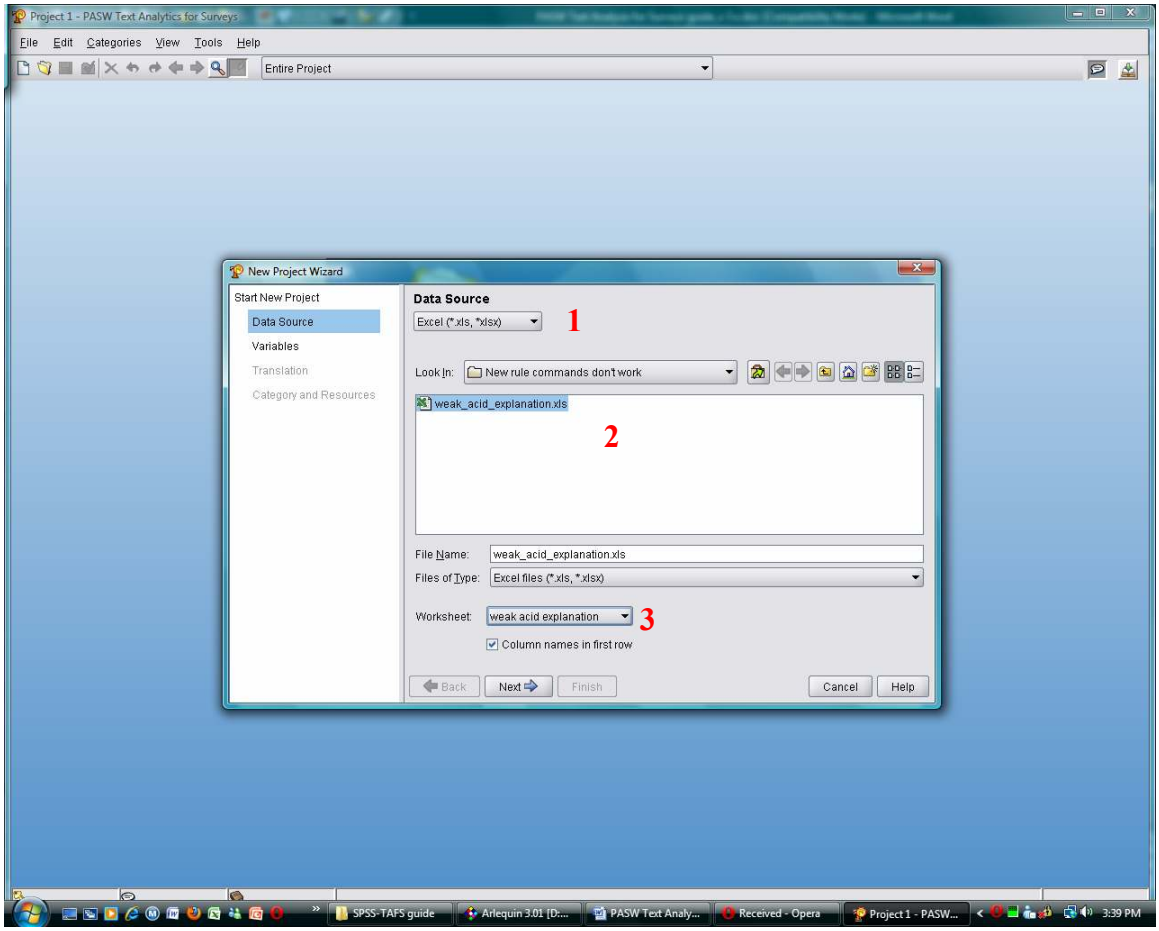
2. Manipulation of data before exporting them into PTAfS:
   Usually, students use different characters when they do not know an answer. It is convenient to substitute those characters by spelling-out the word, since PTAfS will not extract them automatically, i.e. '?' = QUESTION MARK, '!' = EXCLAMATION MARK. Students also may leave answers in blank; in such cases, those blank cells might be filled with the word BLANK. In other cases,

when students use '…' or random characters, it may be recommended to substitute by NA (no answer). A spell-check is also recommended. It is important not to alter the other responses of the students. After these changes, the file will be ready to be imported in PTAfS

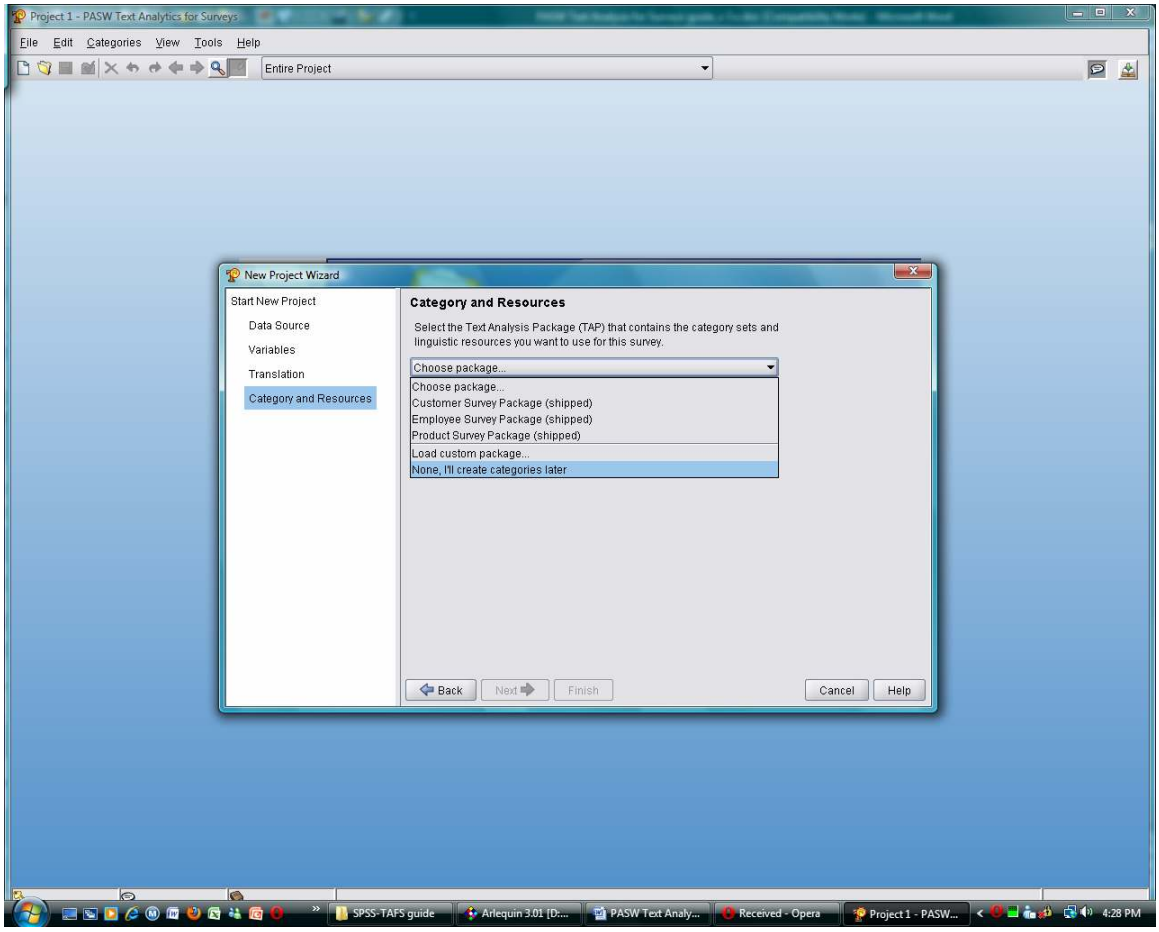
3. Creating a new project in PTAfS:
   To create a new project the data file (usually an excel file) has to be imported. A new project can be created when the program is opened.



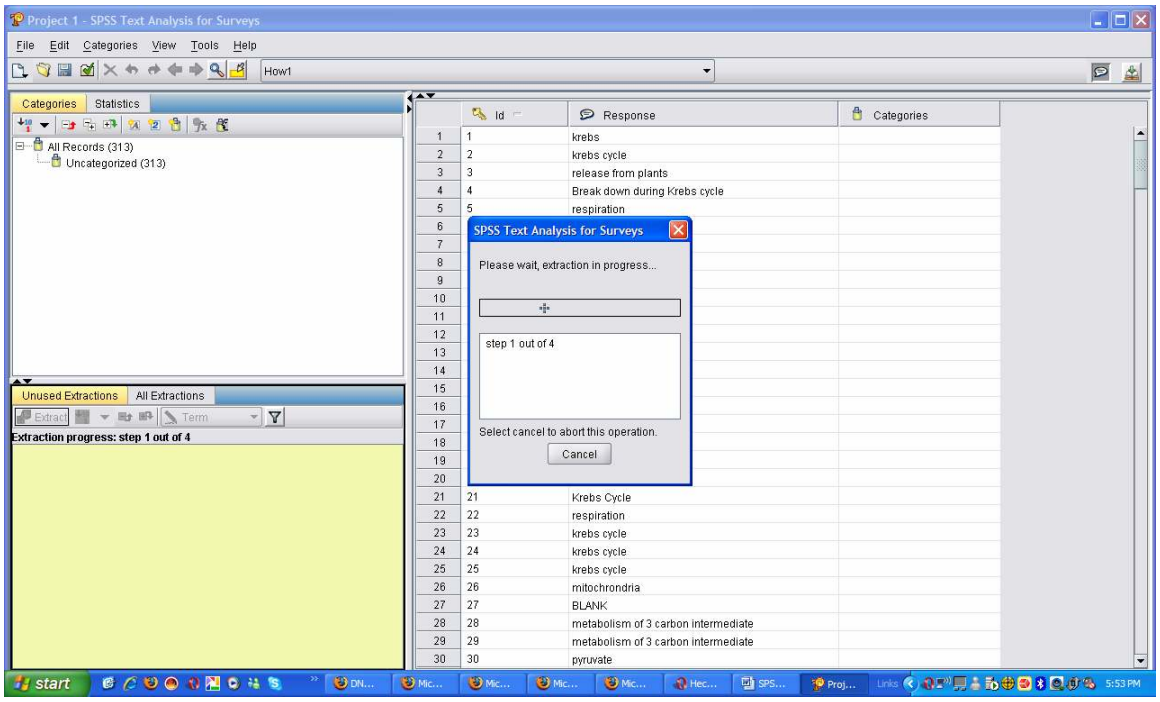
Click on *Starting a New Project*. The *New Project Wizard* window will pop up. The first thing it will ask is the data source. Here, 1- select the file type (excel in most of the cases), 2- its location, and 3- in which worksheet of the excel workbook the data for this project is.



In the next window, the *Variables* (*unique ID*, *the open-ended text,* and *references*) have to be defined. Also, we can define whether we want concepts extracted from all the questions, only the first one, or none. If we decide not to extract any concept at the moment, it can be done manually after creating the project (see 4. Extraction of Concepts). Next, we have the option of *Translation*, but as our projects do not need to be translated, we just pass this window. Finally, we get to select *Category and Resources*. The *Category and Resources* is a new feature of this version. Here, a Text Analysis Package (TAP) is chosen. A TAP contents categories and custom libraries. If we are starting a project from zero, select the option *"None, I'll create categories later*". If we are using one existing already, select *"Load custom package*". Then, we click *Finish* and we are ready start working with a new project (see screenshot below). For each step, we can always click the *Help* button if we have any question. Detailed information can be found on pages 41- 57 of SPSS Text Analysis for Surveys 3.0 User's Guide.

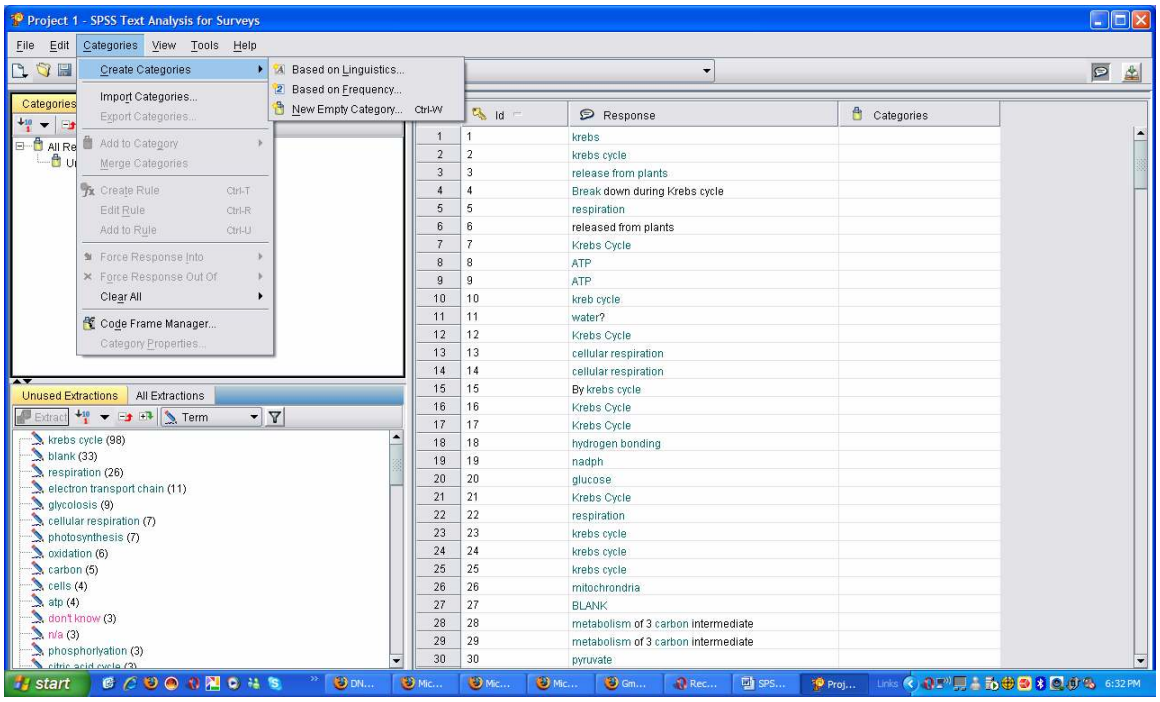


4. Extraction of Data:
   Once the project has been created, there will be three panels on the screen: two on the left and one on the right. The one on the lower left will be in yellow if no extraction has been performed. As mentioned above, *extraction* of concepts means to have the program recognize key words in your data. To extract the concepts, click on the button 'Extract'. A little window will pop up indicating the progress of the extraction (see screenshot below). The *extraction results* panel will include not only a list of concepts, but also types, concept pattern, and type pattern. Based on these results, categories will be created. More detailed information about extraction of data can be read starting on page 99 of the SPSS Text Analysis for Surveys 3.0 User's Guide.



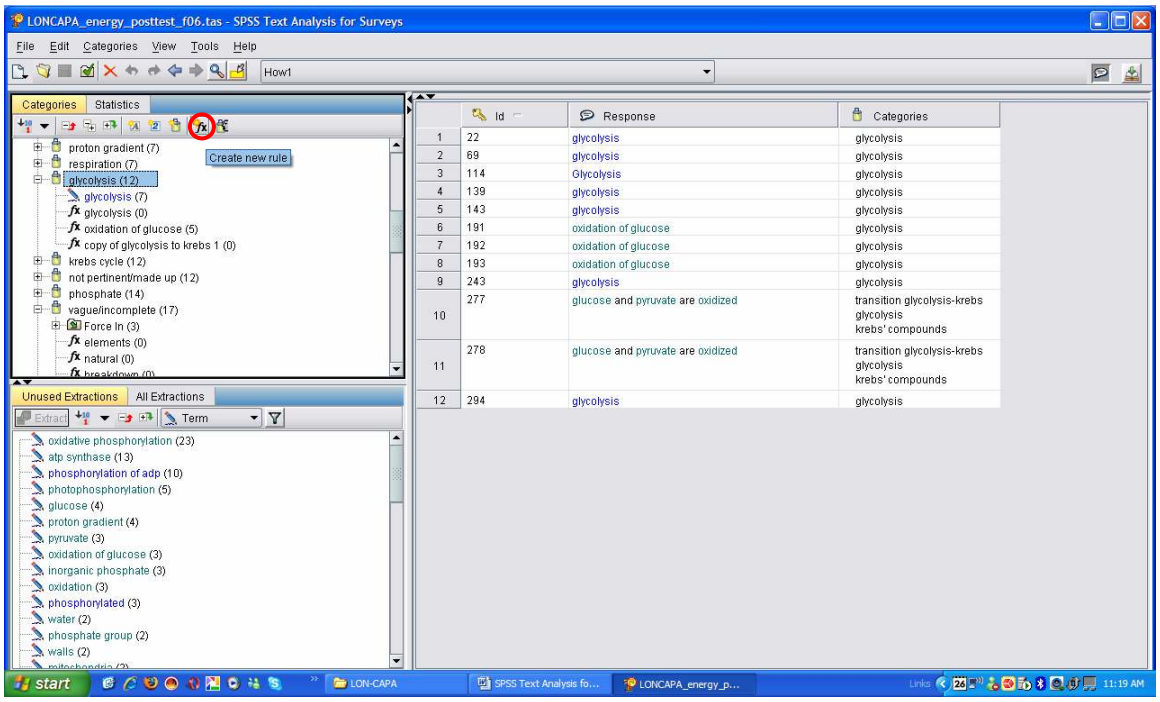
5. Creating categories:
   Once the terms (concepts) have been extracted, we want to group answers in a set of *categories*. It might be convenient to have some sort of "expert answer" to guide the creation of appropriate categories to group the data that are being studied. There are two ways to create categories automatically, by *linguistics* or by *frequency* that the terms appear. See the TAFS manual (pg 101-112) for more information on these procedures.

   Click 'Create Categories' under Categories (see screenshot below). One advantage of using linguistics is that the program will automatically create functions (see below), but in some cases these categories might not be meaningful. By using the option of frequency, the software will automatically group in a category terms with a certain occurrence (usually >25). A combination of both strategies may be convenient in some cases. Other categories that will group different terms but with a general same meaning (i.e. different ways student can say I DO NOT KNOW) can be created manually as an empty category and given a custom name. The Chapter 6 of the User's Guide has more detailed information about creating categories.

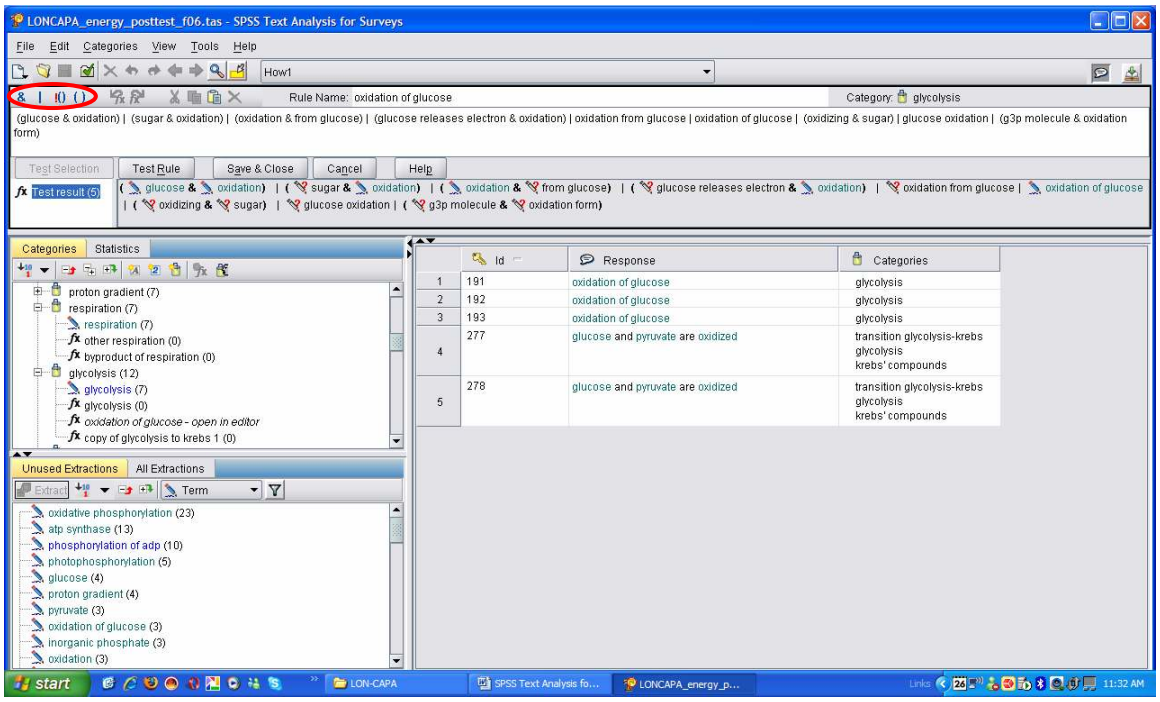


In v.3, this procedure is slightly different. Under Category, there are the options of *Build Categories* and *Build Settings*. Click on *Build Settings* and make sure that Categories will be built from *Types*. Here is where we select whether Categories will be built based on linguistic or frequency. After selections are made, click on the button on the left bottom corner that says *Build Categories*.

6. Creating functions or rules:
   To assure that responses are classified correctly, in some cases it is necessary to create functions that will assist in the classification process. These functions consist in a series of commands to include or exclude responses in the selected category having a given combination of terms (none, both, either, any). To create these functions, click on the category of interest and then click on the icon that has the function symbol (*fx*):



When we want several variations of a term included in a category, i.e. when different prepositions or articles are used in front or the target word, the command that has to be used is "OR" (= |). In those cases in which the combination of two words completes the meaning of a phrase in a category, we should use the command "AND" (= &); for example: if we have the phrase 'oxidations of glucose' and each word was extracted separately. 'Glucose' will be classified as a sugar and 'oxidation' as a chemical process, but 'oxidation of glucose' is a different process. If we want to make sure that process is correctly classified, we create a rule specifying that that specific combination of words (glucose & oxidation) should go in that category.



In the same example above, let's suppose that when a student answers 'oxidation of glucose' we do not want the word 'glucose' classified with 'sugars', then we create a rule using the command "NOT" (= !( )), i.e. glucose & !(glucose & oxidation), which reads 'include glucose but NOT when it is at the same time with the word oxidation'. It is important that when several commands are combined, we use parenthesis; otherwise all commands will be applied to the complete list of words and the rule will not work. As in the example above, there is a list of words all separated by the command OR, so ANY of those words will be accepted by the rule created. If a parenthesis is missing between those words that also have the command AND, then this command will also be applied to the whole list and no word will be included in this rule

Although all the command symbols are available on the computers' keyboard, they will not work if just they are typed in. The different commands will only work in a rule if they are included by clicking on the respective buttons of the program. For further information about creating rules, refer to pages 128-132 of the User's Guide.

In v.3, there are 3 additional commands that allow a better creation of rules:

Character	Description
+	The pattern connector used to form an order-specific pattern. When present, the square brackets must be used. (See examples)
[ ]	The pattern delimiter that is required if a pattern is being defined.

Character	Description
*	A wildcard representing anything from a single character to a whole word depending how it is used.

7. Libraries, Substitution and Exclude Dictionaries:
   All words that are included in PTAfS are grouped in libraries. Each one of these libraries has two dictionaries associated: the Substitution Dictionary contents synonyms, acronyms, or inflections of words in each library. The Exclude Dictionary is a list of words associated to each library that will not be extracted.

   Given the specialized nature of most of the biological terms, some words of interest will not be extracted by PTAfS. In this case, we must create a custom library for our project. For example, for the LON CAPA project, several words that were not recognized by the program were added, as well as synonyms or misspellings.

   On the right corner of PTAfS there are two icons: one is like a callout symbol and the other is like an open book with a green arrow, which are used to switch between the text analysis view and the dictionary editor view, respectively. In the dictionary editor view there are four panels: on the top left there are two: the list of libraries and the active library (the screenshot below it is showing the LON CAPA library); on the bottom left is the Substitution dictionary, a list of synonyms associated to the libraries above, and on the right is the Exclude dictionary, a list of terms removed associated with the libraries listed on the left. By default, there are 5 libraries: Core Library, Budget library, Opinion library, English library and Project library. The first four are built-in libraries and we are not going to edit them; the last one is empty and it is where we are going to start adding terms that are exclusive to our project to build our custom library. Thus, to have new words extracted, they have to be added to the Project library. When a word is added, it must be specified how the program will match the word in the data. By default it will match the exact term, but in those cases in which we want to extract a phrase, i.e. a process that has several words like "Electron Transport Chain", you should specify if you want the entire phrase or each word of the phrase extracted; there is also the option of having both tings done. Sometimes it may be a good idea to try using all three commands to see with which one we get the desired result. See pages 151-153 of the User's Guide for further information.

   Synonyms are added in the Substitution dictionary panel. In this panel we can add spelling variations or misspellings; also, when there are several technical names for a same process, acronyms or abbreviations. For example, the Krebs cycle is also known as acid citric cycle, or tricarboxylic acid cycle, Szent-Györgyi-Krebs cycle or TCA; students could use any of these options and we may prefer to have all of them as synonyms of simply Krebs cycle. Also, in some cases, we may want a group of words to be considered as a general category, i.e. different types of

sugars, like fructose, maltose, hexose, starch, carbohydrates… are just considered as sugars.