Chapter 6

SemanticExcel.com:
An Online Software for Statistical Analyses of Text Data based on Natural Language Processing

Aims and content:

The overall aim of this chapter is to present a guide in how to efficiently measure and statistically analyze text and numerical data using the online software SemanticExcel.com; we will focus on the following main functions:

1. Computing semantic similarity scores between the semantic representations of two sets of texts.
2. Testing whether the semantic representations of two sets of texts statistically differ using a paired semantic t-test.
3. Train the semantic representations to predict numerical values.
4. Predicting numerical values from the semantic representations of texts by applying a semantic trained (valence) model.
5. Visualize words based on statistically significant relationships along y and x-axes.

Exercises are provided for you to practice on in the end.

This chapter instructs on how to practically measure and statistically test text data based on natural language processing using the online software called Semantic Excel (www.semanticexcel.com/; see also Sikström, Kjell & Kjell, in progress). Semantic Excel has been developed to make statistical analyses of text easy. The aim is that researchers or professionals interested in quantification of text, should be able to use the software with minimal guidance; they should be able to start working on their analyses within minutes from logging in. Thus, Semantic Excel enables analyses of both words as well as numerical data within a user-friendly point-and-click environment. In Chapter 2 by Nielsen et al., we learned how to create semantic spaces, this chapter focuses on applying the semantic representations from semantic spaces. The semantic representations are in this chapter applied for different types of analyses. This includes computing *semantic similarities* between two words/texts and applying *semantic t-tests* to examine whether or not there is a statistically significant difference between two sets of texts (as in Chapter 4, Kjell, Kjell & Sikström, in press); as well as using *semantic training* to train the semantic representations to numerical values and predicting numerical values from text based on *semantic prediction* (as in Chapter 5, Kjell, Kjell & Sikström, in press). We will also learn how to publically save word norms and predictions so that other Semantic Excel users can use them in their research.

Language, the Semantic Space and their Semantic Representations

Semantic Excel comes with the possibility to choose from preprogrammed semantic spaces from more than 15 languages. That is, Semantic Excel enables you to map semantic representations from the semantic space to your word data. Table 1 presents information relevant to describe the semantic space that you select. For more information about the semantic spaces and semantic representations see the chapter by Hansen et al. (chapter 2) or Sikström, Kjell and Kjell (in progress).

Table 1.
Semantic spaces currently accessible in Semantic Excel

	Word Corpus				Output		
Language Space	**Size (nr. of words)**	**Contexts; Origin**	**Rows (words)**	**Columns (contexts)**	**Criteria for Dim. selection**	**Dim.**	**Performance/ Validation test**
English 1	8.70 e+11	5-gram; Google[1]	120,701	10,000	SSEI	512	SSEI = .0003
English 2	8.70 e+11	5-gram; Google[1]	120,599	50,000	SSEI	512	SSEI = .0001
Swedish 1	2.99 e+13	5-gram; Google[2]	120,448	10,000	SSEI	256	SSEI = .0164
Swedish 2	2.99 e+13	5-gram; Google[2]	110,006	50,000	SSEI	256	SSEI = .0112
Spanish	3.23 e+11	5-gram; Google[1]	120,406	10,000	Default	300	Not yet done
Dutch	4.58e+12	5-gram; Google[2]	120,386	10,000	Default	300	Not yet done
Romanian	1.21 e+13	5-gram; Google[2]	120,386	10,000	Default	300	Not yet done
Italian	5.62 e+10	5-gram; Google[1]	107,235	10,000	Default	300	Not yet done
German	1.90 e+11	5-gram; Google[1]	120,408	10,000	Default	300	Not yet done
French	4.97 e+11	5-gram; Google[1]	120,406	10,000	Default	300	Not yet done
Chinese	-	5-gram; Google[1]	38,714	10,000	Default	300	Not yet done
Czech	3.73 e+13	5-gram; Google[2]	120,384	10,000	Default	300	Not yet done
Finnish	7.22 e+11	5-gram; Google[2]	120,384	10,000	Default	300	Not yet done
Hebrew	-	5-gram; Google[2]	94,033	10,000	Default	300	Not yet done
Polish	3.80 e+13	5-gram; Google[2]	120,352	10,000	Default	300	Not yet done
Portuguese	2.49 e+13	5-gram; Google[2]	120,384	10,000	Default	300	Not yet done
Russian	4.06e+11	5-gram; Google[2]	120,334	10,000	Default	300	Not yet done
Persian	-	-	104,352	10,000	Default	300	Not yet done
Norwegian	1.94 e+11	-	120,384	-	Default	300	Not yet done
Danish	2.18e+10	-	120,384	-	Default	300	Not yet done

Notes. The Logarithmic frequency plus 1 was used to transform the matrix for all languages. Singular value decomposition to generate 300 to 512 dimensions were used for all languages. Dim = Dimensions; SSEI = Semantic space error index.

This Table is borrowed from Sikström, Kjell and Kjell (in progress).

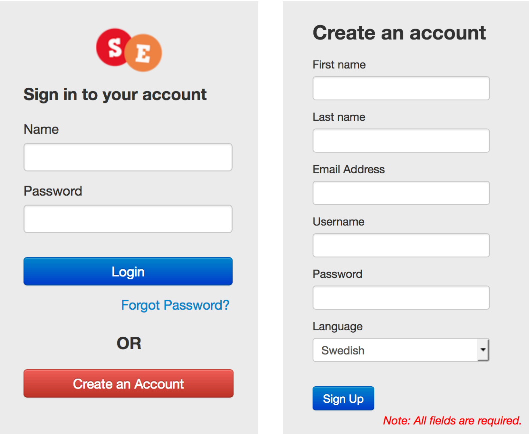
[1] The 5-grams come from Google N-gram Database, version July 1, 2012; see https://books.google.com/ngrams.

[2] Google N-gram CD with 10 European languages (https://catalog.ldc.upenn.edu/LDC2009T25)

Set up an account and getting started

To start analyzing text data go to www.semanticexcel.com and set up an account. Press the red "Create an Account" button and fill out the request information (Figure 1). The "Website tour" button in the upper right corner is a good way to familiarize yourself with the website.

Further, you find several peer-reviewed published articles using Semantic Excel (and the Matlab version *Semantics*) under the "Publication tab". To ask, answer and read questions about Semantic Excel and natural language processing you can get involved in the "User community".



Language: Swedish

Note: All fields are required.

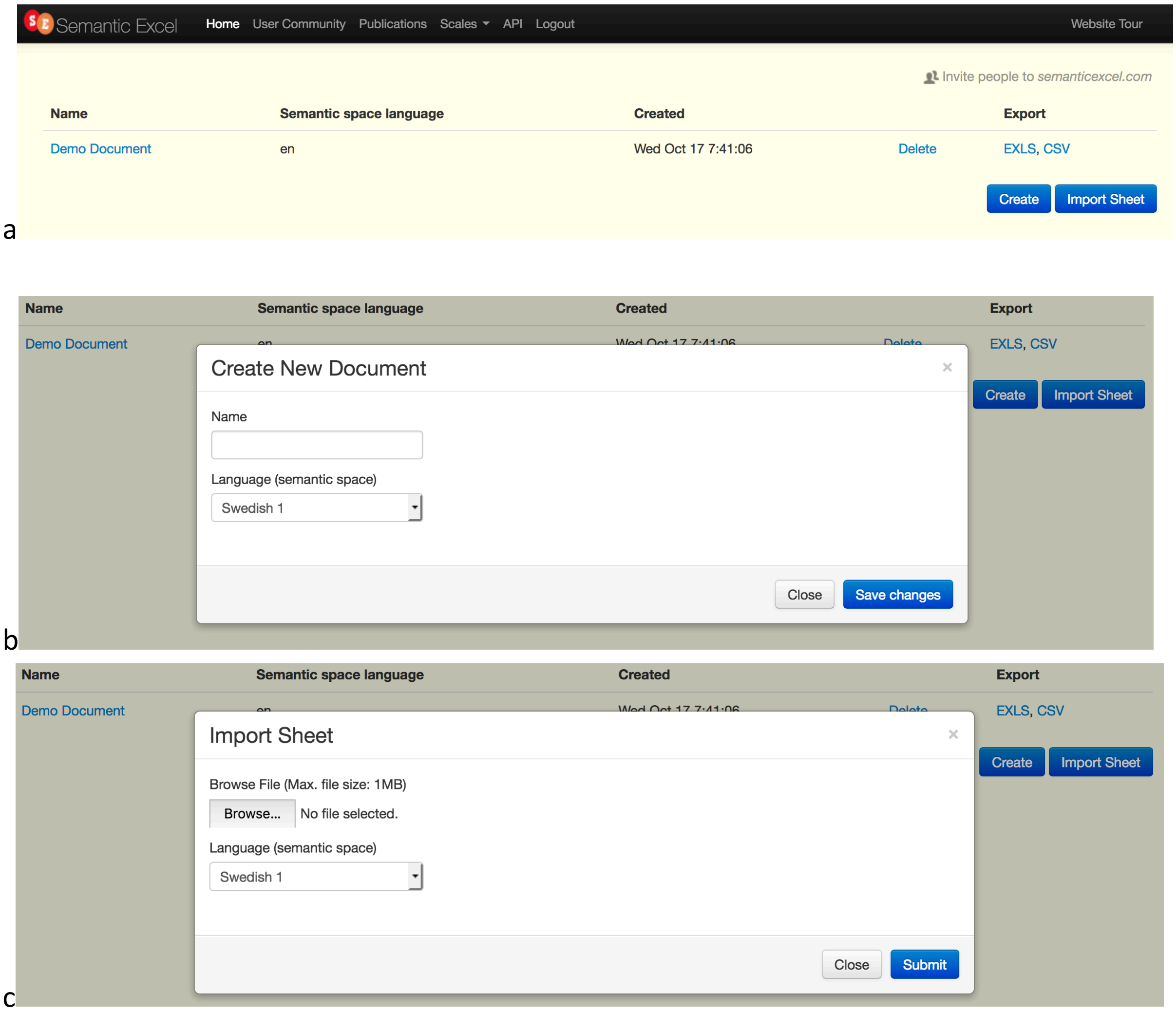
Figure 1. Create an account to start analyzing text data

Importing and exporting data

To import your research data into Semantic Excel, there are two different ways using **Create** or **Import** (Figure 2a). Selecting **Create** gives a pop-up window (Figure 2b) allowing you to name your document and then select a **Language** (i.e., a semantic space). Select the same language as your text data, and then press **Save changes**. This creates a document that can be found under **Home**. Selecting this document opens an empty spread sheet in which one can start writing or copy-pasting data to. Instead of choosing **Create**, you can select **Import Sheet**, which opens a window (Figure 2c) enabling you to **Browse** to a csv or excel file on your computer and then select **Language (semantic space)**. The created document is again found in the **Name** column in **Home** having the same name as the imported data file. Clicking the link will take you to your data; and to change the name of the document click on the current document name in the right upper corner.

The newly created document contains by default 200 rows, which are labelled 1-200; and 26 columns labelled A to Z. In case you require a larger sheet with unlimited document size you can in the right upper corner "**Apply for more cells**". The default size limit is set in order to avoid crashing the servers since running the natural language processing algorithm with loads of data often require a lot of computationally power.

It is easy to export the data again. This is useful when wanting to flexibly switch between different statistical programs such as Semantic Excel and SPSS or R. This is done by selecting **File** in the left upper corner, and pressing **Export** to csv or excel.



Name	Semantic space language	Created		Export
Demo Document	en	Wed Oct 17 7:41:06	Delete	EXLS, CSV

Create New Document

Name

Language (semantic space): Swedish 1

Import Sheet

Browse File (Max. file size: 1MB)

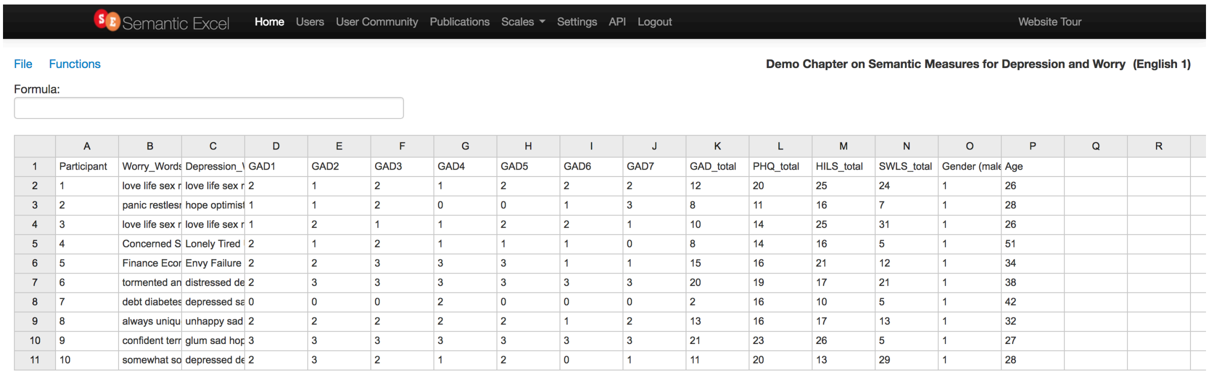
No file selected.

Language (semantic space): Swedish 1

Figure 2a-c. Demonstrates how to Create and Import Data into Semantic Excel.

Demo data on depression and worry

This tutorial uses a subset of data (Figure 3) from Chapter 4 and 5. There are 99 rows of data; the first row contains the column names and the following 99 rows each represent a participant. Column A includes participant number, column B and C contain the five descriptive word responses to the semantic questions of worry and depression, respectively. Column D through J contain the individual responses to the 7 items in GAD-7 (Spitzer, Kroenke, Williams, & Löwe, 2006); and column K through N contain the composite scores of GAD-7 (Spitzer et al., 2006), PHQ-9 (Kroenke, Spitzer, & Williams, 2001), HILS (Kjell, Daukantaitė, Hefferon, & Sikström, 2016), and SWLS (Diener, Emmons, Larsen, & Griffin, 1985). Column O and P hold information about gender (i.e., male=1; female=2; other=3) and age, respectively.



Demo Chapter on Semantic Measures for Depression and Worry (English 1)

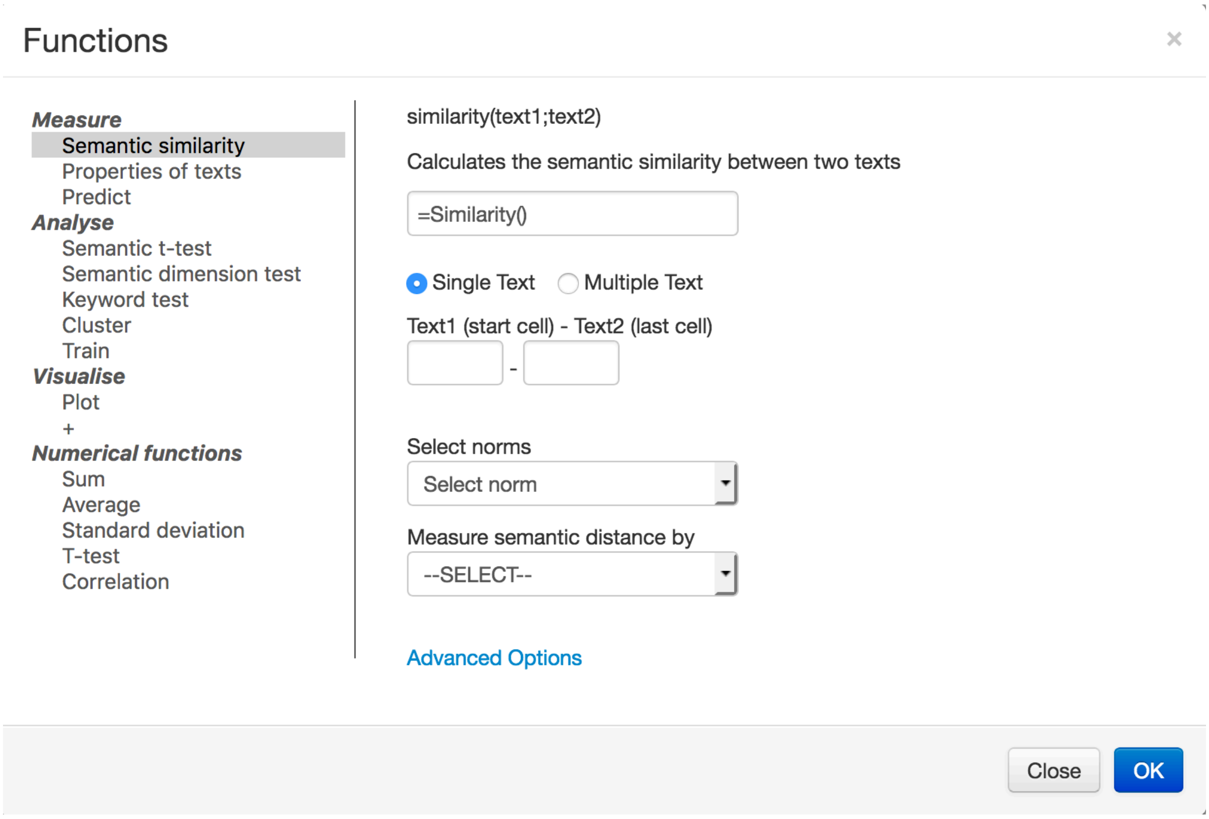
Formula:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Participant	Worry_Words	Depression_\	GAD1	GAD2	GAD3	GAD4	GAD5	GAD6	GAD7	GAD_total	PHQ_total	HILS_total	SWLS_total	Gender (male	Age		
2	1	love life sex r	love life sex r	2	1	2	1	2	2	2	12	20	25	24	1	26		
3	2	panic restless	hope optimist	1	1	2	0	0	1	3	8	11	16	7	1	28		
4	3	love life sex r	love life sex r	1	2	1	1	2	2	1	10	14	25	31	1	26		
5	4	Concerned S	Lonely Tired	2	1	2	1	1	1	0	8	14	16	5	1	51		
6	5	Finance Ecor	Envy Failure	2	2	3	3	3	1	1	15	16	21	12	1	34		
7	6	tormented an	distressed de	2	3	3	3	3	3	3	20	19	17	21	1	38		
8	7	debt diabetes	depressed sa	0	0	0	2	0	0	0	2	16	10	5	1	42		
9	8	always uniqu	unhappy sad	2	2	2	2	2	1	2	13	16	17	13	1	32		
10	9	confident terr	glum sad hop	3	3	3	3	3	3	3	21	23	26	5	1	27		
11	10	somewhat so	depressed de	2	3	2	1	2	0	1	11	20	13	29	1	28		

Figure 3. Demo data: a subset from Chapter 4.

Functions Overview

To carry out different analyses in Semantic Excel you must always be positioned in a cell. This indicates where you want the output/result to be printed. Then you can select **Functions** in the left upper corner in order to retrieve a pop-up window presenting the different analytic options (Figure 4). The functions are divided into four sections. The first section is labelled **Measure** and covers functions that return various measurements of text including semantic similarity between two words or predicted values based on previously trained models. The second section is called **Analyse** and comprises functions that performs statistically based tests. The third section is called **Plot**, which includes functions that visualize the data in different figures and word clouds. The last section is called **Numerical functions** and covers some of the most commonly used and basic analyses on numerical variables.



similarity(text1;text2)

Calculates the semantic similarity between two texts

=Similarity()

● Single Text ○ Multiple Text

Text1 (start cell) - Text2 (last cell)

Select norms

Select norm

Measure semantic distance by

--SELECT--

Advanced Options

Figure 4. Overview of Current Functions in Semantic Excel.

Semantic Similarity

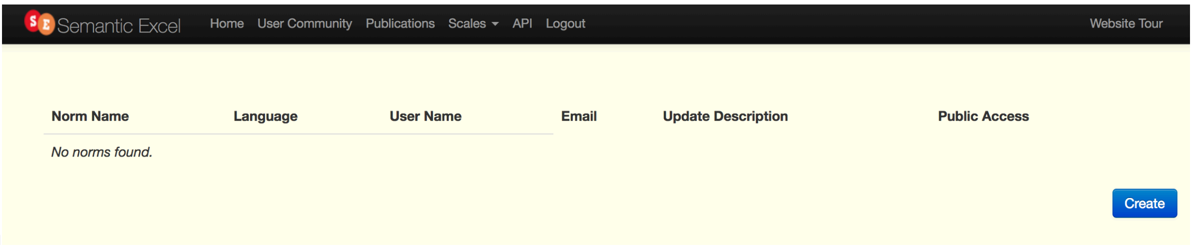
To compute the semantic similarity as measured by the cosine between vectors representing two sets of texts located in different cells, select **Semantic similarity** under **Measure** in **Functions**. To the right in the pop-up window you can now see that the default option is **Single Text** as marked with a blue circle. This option computes the semantic similarity between the texts contained within the cells specified in the boxes labelled **Text 1** and **Text 2**; and the result is returned in the cell where the cursor is positioned in the sheet. However, in line with the analyses in Chapter 4, we will compute the semantic similarity between the text responses to the worry question (i.e., column B) and the depression word norm.

Creating and using a word norm

To save a word norm, select **Scales** and then **My norms** in the main menu at the top of the page (Figure 5a), and then click the blue **Create** button. In the pop-up window (Figure 5b), select the **Language**, or more specifically the space you want to save the word norm to and give it a name in the **Name** box. Then insert the words you want the norm to comprise of in the **Text norm** box. In the **Explanatory comment**, elaborate on the technical and methodological details related to the norm. We encourage users to make their word norms public by ticking the **Make word norm public:** box. In these circumstances it is important to properly describe and reference your word norm so that those deciding to use it can properly reference it in their research, and potentially also give your email in case they want to contact you. We propose that you describe your norm by giving the following information: Reference to potential paper, word norm question used to gather the data; number of participants and their demographic

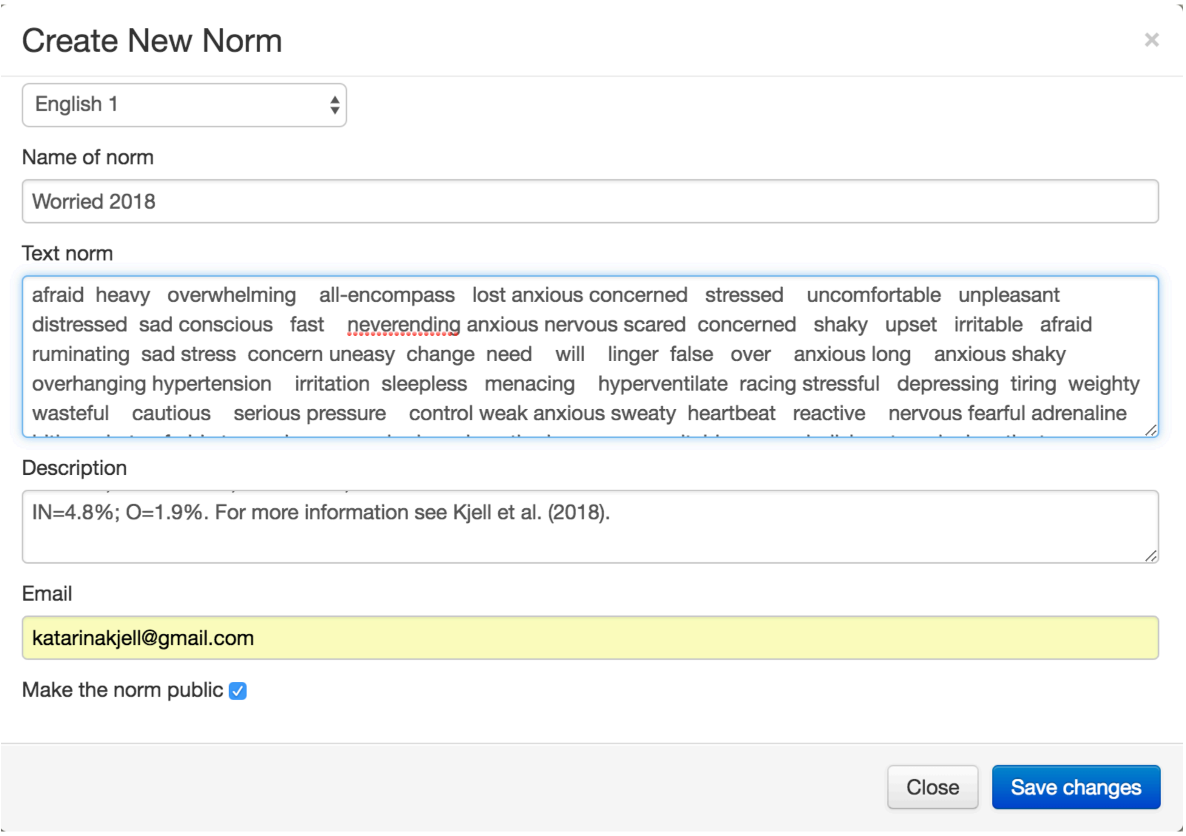
In our case for the worry norm this can look something like this:

> This word norm was collected in Kjell, Kjell, Garcia and Sikström (2018).
> Participants were asked to answer the following question: “Please write 10 words that best describe your view of being worried. Write descriptive words relating to those aspects that are most important and meaningful to you. Write only one descriptive word in each box.”
> Using Mechanical Turk, 104 participants' age ranged from 18-65 (M=28.73 SD=8.80) years; Female =52.9%; Male =46.2%; Other=1.0%; US=93.3%;
> Indians=4.8%; Other=1.9%. For more information see Kjell et al. (2018).



Norm Name	Language	User Name	Email	Update Description	Public Access
No norms found.					

a



Create New Norm

English 1

Name of norm

Worried 2018

Text norm

afraid heavy overwhelming all-encompass lost anxious concerned stressed uncomfortable unpleasant distressed sad conscious fast neverending anxious nervous scared concerned shaky upset irritable afraid ruminating sad stress concern uneasy change need will linger false over anxious long anxious shaky overhanging hypertension irritation sleepless menacing hyperventilate racing stressful depressing tiring weighty wasteful cautious serious pressure control weak anxious sweaty heartbeat reactive nervous fearful adrenaline

Description

IN=4.8%; O=1.9%. For more information see Kjell et al. (2018).

Email

katarinakjell@gmail.com

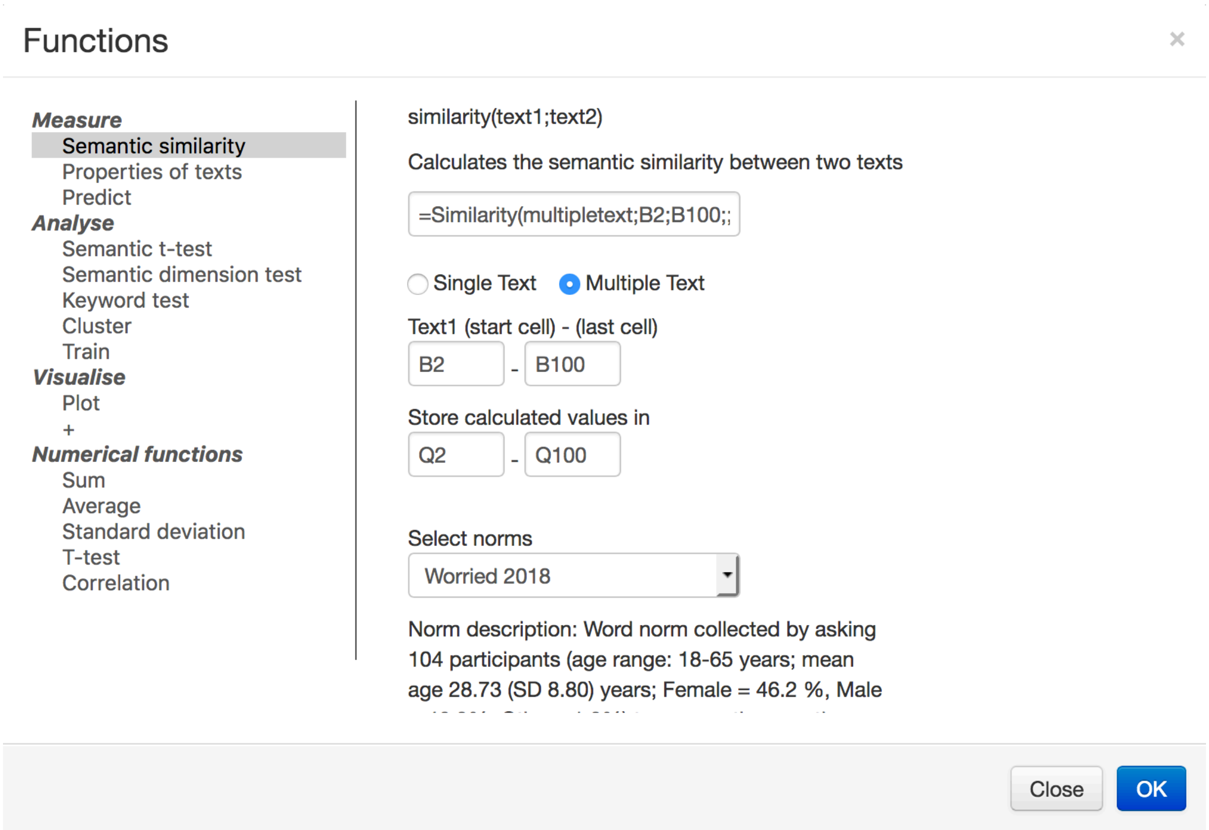
Make the norm public ☑

b.

Figure 5a-b. Demonstration of how to create a word norm

Now, since we made the norm public, you and anyone else using this same semantic space can use this word norm. Going back to the document we were in before (i.e., select **Home**, and click the link called *Demo Chapter on Semantic Measures for Depression and Worry* under **Name**), we can now apply the word norm to all the word responses. Select **Functions**, **Semantic Similarity** and then shift from the **Single text** to the **Multiple text** option (Figure 6). To measure the semantic similarity between each of the worry responses and the worry word norm, select all the worry word responses by specifying the cells in the **Text 1** boxes, in the first box write **B2** and in the second box write **B100** (; i.e., column B row 2 to row 100). Then select the **Select norms** option and select the Worried 2018 norm in the drop-down list. Lastly indicate, where you want the semantic similarity scores to be displayed by filling in the **Store calculated values in** boxes; to store them in column Q write **Q2** and **Q100**. Click **OK** and see how column Q gets populated with the semantic similarity scores. Label the column by writing, for example, writing "Worry semantic similarity scale" in the Q1 cell.

To examine the correlation between the computed semantic similarity scores and, for example, the GAD-7 total score (i.e., variable GAD_total in column K), select **Function**, and **Correlation** under **Numerical functions**; insert the column and cell numbers for the two variables (i.e., K2-K100; Q2-Q100) and press **OK**. To see the results, look at the cell where the cursor was positioned; when it does not say "*Calculating…*" anymore **right click** (or click with two fingers on Mac) on the cell and select **Show value**. Interestingly the two variables yield a Pearson correlation of .29 (p = .0039).



similarity(text1;text2)

Calculates the semantic similarity between two texts

=Similarity(multipletext;B2;B100;;

○ Single Text ● Multiple Text

Text1 (start cell) - (last cell)

B2 - B100

Store calculated values in

Q2 - Q100

Select norms

Worried 2018

Norm description: Word norm collected by asking 104 participants (age range: 18-65 years; mean age 28.73 (SD 8.80) years; Female = 46.2 %, Male

Figure 6. Settings to measure semantic similarity between word responses and a word norm.

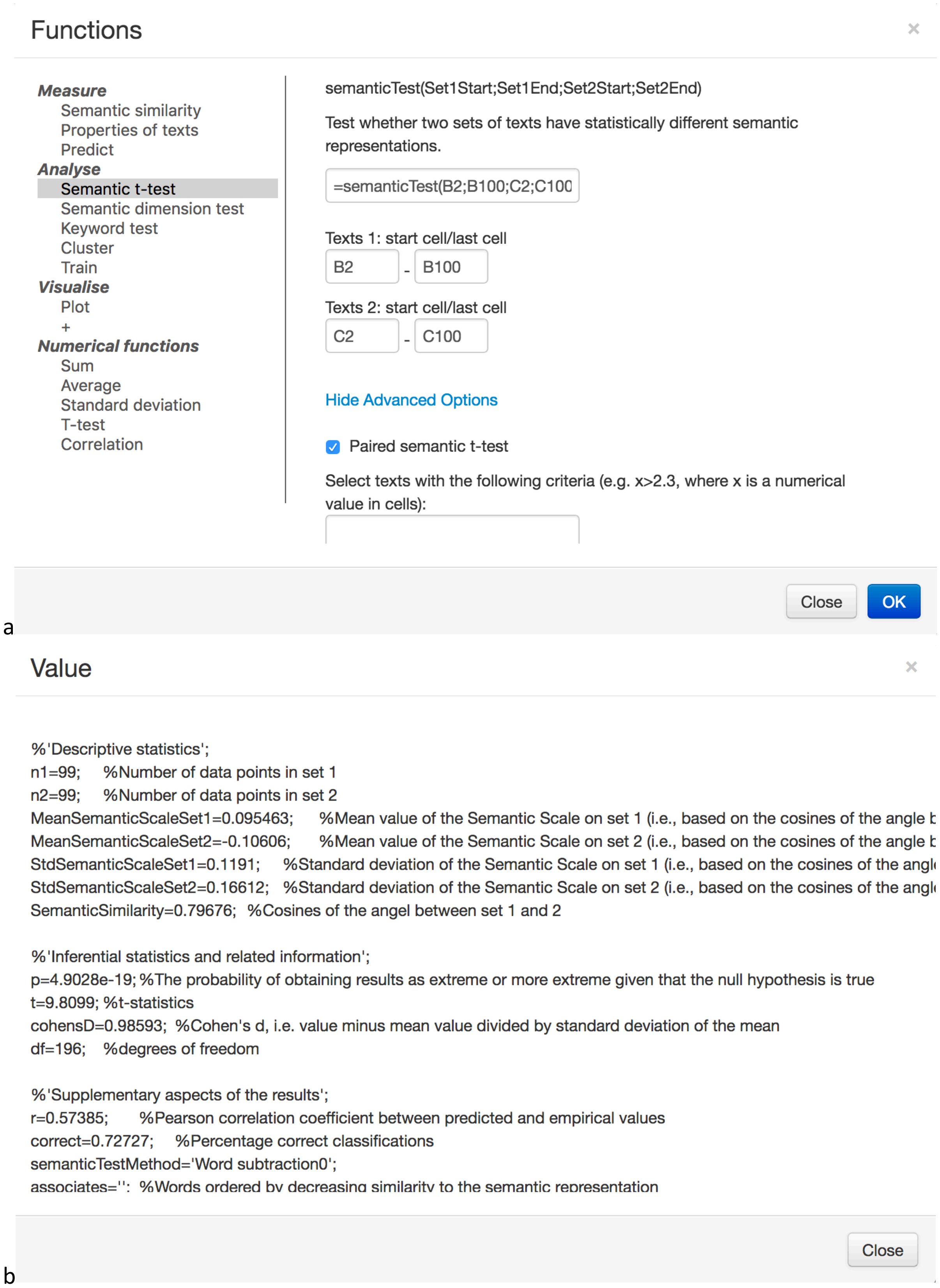
Using advanced settings to correct for frequency artifacts

It has been found that controlling for artifacts linked to the occurrence frequency of words may increase the validity when adding semantic representations together (for more details see Kjell et al., 2018). To turn this correction function on in Semantic Excel, click **Advanced Settings** and in the **Set Parameter here** dropdown list under **CONTEXT GENERATONS** select **Correct for frequency artifacts during creation of semantic representations**. In the appearing box change the number 0 (i.e., the parameter is turned off) to 1 (i.e., the parameter is turned on). This change would influence subsequent analyses, however, for this chapter we will leave it turned off (i.e., set to 0). The advanced settings also include a large number of additional parameters that controls Semantic Excel. A short description of their function can be seen on the same row as they are displayed. However, this chapter do not cover these parameters.

Semantic t-test

Next, we use the semantic representations to examine whether there is a significant difference between worry and depression word responses. In Semantic Excel, this is done by selecting **Semantic t-test** under **Analyse** in **Functions** (Figure 7a). To compare the worry and depression word responses, insert **B2** to **B100** in the **Text 1** boxes and **C2** to **C100** in the **Text 2** boxes. By default, Semantic Excel performs a Student's t-test, however, since we here want to compare the word responses within participants, tick **Advanced Options** and then select **Paired semantic t-test**. Press **OK** to execute the analysis.

Examine the results by right-clicking as described before. The pop-up window (Figure 7b), first shows *Descriptive statistics* such as *N*, *M* and *SD*; secondly, *Inferential statistics and related information* such as t-value, p-value and effect size; and thirdly the *Supplementary aspects of the results*. In short, our paired semantic t-test shows that worry and depression word responses (N = 99) significantly differ [$t_{(196)} = 9.81$, $p < .001$] with a strong effect size (Cohen's d = 0.99). (As a side note, the percent signs that you see throughout the presentation of the results are there because they are in Matlab used to comment text out. Hence, this enables you to easily copy-and-past the results from the pop-up window into Matlab for further analyses).



semanticTest(Set1Start;Set1End;Set2Start;Set2End)

Test whether two sets of texts have statistically different semantic representations.

=semanticTest(B2;B100;C2;C100

Texts 1: start cell/last cell: B2 - B100

Texts 2: start cell/last cell: C2 - C100

☑ Paired semantic t-test

Select texts with the following criteria (e.g. x>2.3, where x is a numerical value in cells):

```
%'Descriptive statistics';
n1=99;    %Number of data points in set 1
n2=99;    %Number of data points in set 2
MeanSemanticScaleSet1=0.095463;    %Mean value of the Semantic Scale on set 1 (i.e., based on the cosines of the angle b
MeanSemanticScaleSet2=-0.10606;    %Mean value of the Semantic Scale on set 2 (i.e., based on the cosines of the angle b
StdSemanticScaleSet1=0.1191;    %Standard deviation of the Semantic Scale on set 1 (i.e., based on the cosines of the angl
StdSemanticScaleSet2=0.16612;    %Standard deviation of the Semantic Scale on set 2 (i.e., based on the cosines of the angl
SemanticSimilarity=0.79676;  %Cosines of the angel between set 1 and 2

%'Inferential statistics and related information';
p=4.9028e-19; %The probability of obtaining results as extreme or more extreme given that the null hypothesis is true
t=9.8099; %t-statistics
cohensD=0.98593;  %Cohen's d, i.e. value minus mean value divided by standard deviation of the mean
df=196;    %degrees of freedom

%'Supplementary aspects of the results';
r=0.57385;      %Pearson correlation coefficient between predicted and empirical values
correct=0.72727;    %Percentage correct classifications
semanticTestMethod='Word subtraction0';
associates='';  %Words ordered by decreasing similarity to the semantic representation
```

Figure 7ab. Shows how to compute and understand the results from a paired semantic t-test

Semantic training

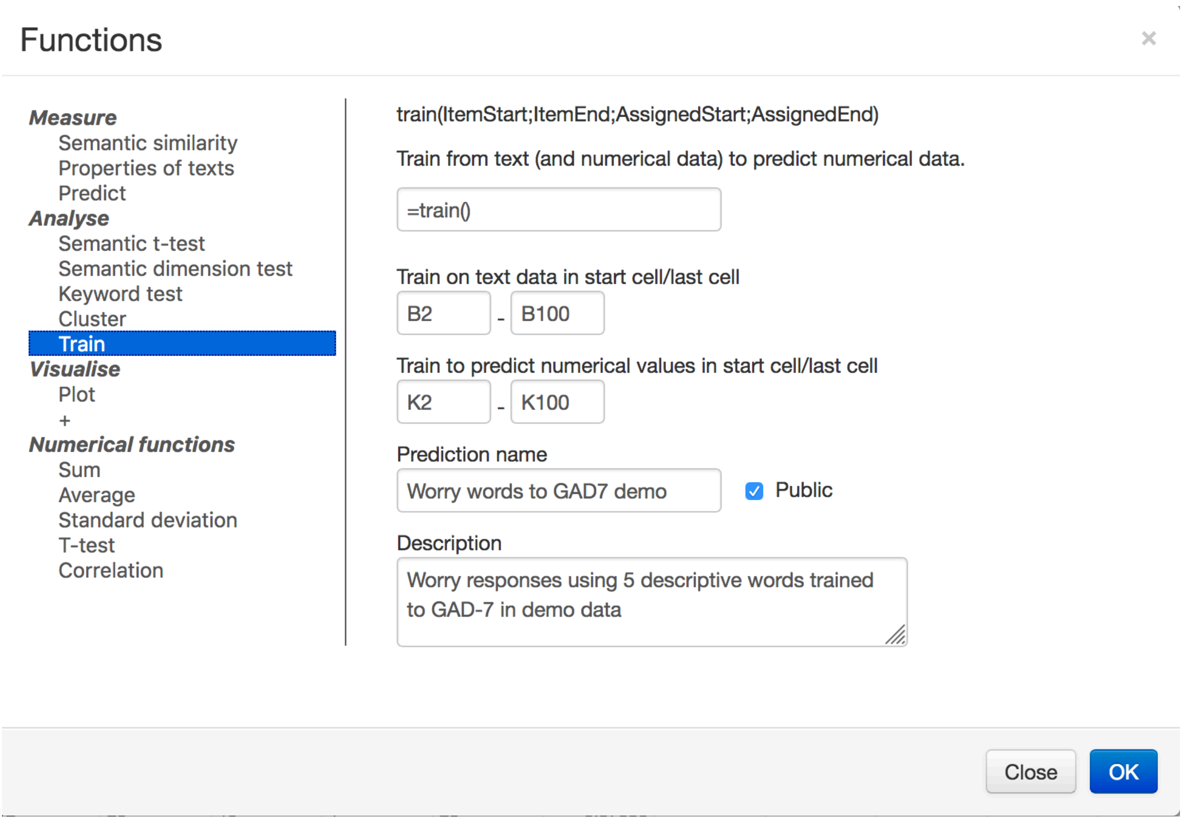
Training the semantic representations of words to a numerical variable may be used to examine whether there is a statically significant relationship between words and numbers. Chapter 5 included, for example, the examination of whether the semantic representations of worry word responses could statistically predict the rating scale score for worry (i.e., GAD-7). This is in Semantic Excel done by going to **Train**, under **Analyse** in **Functions**. In the **Train on text data in start cell/last cell** boxes insert B2 and B100, which hold the worry words responses; and in the **Train to predict numerical values in start cell/last cell** boxes insert K2 and K100, which hold corresponding composite rating scale scores of GAD-7 (Figure 8a). By giving the prediction a name it will be saved so that you can use it to predict the GAD-7 score from other texts. As with word norms it is also possible to make a trained model public by ticking the **Public** box and write a **Description**. To increase the chances of having others benefit from your trained models it is important to sufficiently describe it. And we propose that you include some information from the results, so for now just write "Worry responses using 5 descriptive words trained to GAD-7 in demo data" in the Description box and press OK.

Right clicking on the cell where the cursor was positioned and selecting **Show value** presents the results in a pop-up window (Figure 8b). The results reveal that the semantic representations of the worry word responses (*N*=99) significantly predict the GAD-7 composite scores with a Pearson correlation of .34 (*p* < .001) between the actual and predicted scores (using leave-10%-out cross-validation; i.e., see *NGroups* in *Supplementary aspects of the results*). Even though the correlation is significant it is considerably smaller than that presented in Chapter 5, which most likely is due to the lower number of participants used in this analysis. This information is now suitable to include in the description of the prediction that we created during the analysis. Select **My Predictions** in the main menu, and then click on the **paper-and-pen icon** next to the prediction that we named "Worry words to GAD7 demo chapter". In the pop-up window, it is possible to update the information (Figure 8c). We propose adding and updating information regarding how to reference the prediction, what was trained, number of participants and their demographics, correlation and associated p-value. In our case, we add:

> This prediction is based on the responses to the semantic question for worry (Kjell, Kjell, & Sikström, in progress 2018) using 5 descriptive words trained to the Generalized Anxiety Disorder scale 7 (Spitzer, Kroenke, Williams, & Löwe, 2006). It is based on the demo data for the chapter by Sikström, Kjell and Kjell.
> r=0.34056; *p*=0.00028115.
> N = 99; male = 36; female = 63; age = 34.85 (SD = 9.49) years.
> Cite as: Sikström, Kjell and Kjell Analyzing in Semantic Excel: The online software that facilitates text analyses based on Natural Language Processing.

Using advanced settings to change training algorithm

In advanced settings, it is possible to change the type of algorithm used to train the data. In the third row of the result output you can read that the default algorithm that we used in the previous training was regression. However, if we want to predict a categorical variable such as gender or different conditions (rather than the interval rating scale scores) it is possible to change the algorithm to logistic regression. This function is set by clicking **Advanced Options** in **Train**, and then in the **Set Parameters here** dropdown list select **Type of predictor used during training**, which is positioned under the heading **TRAIN**. In the new dropdown list that appears it is now possible to change from linear **regression** to **logistic** regression.



train(ItemStart;ItemEnd;AssignedStart;AssignedEnd)

Train from text (and numerical data) to predict numerical data.

=train()

Train on text data in start cell/last cell

B2 - B100

Train to predict numerical values in start cell/last cell

K2 - K100

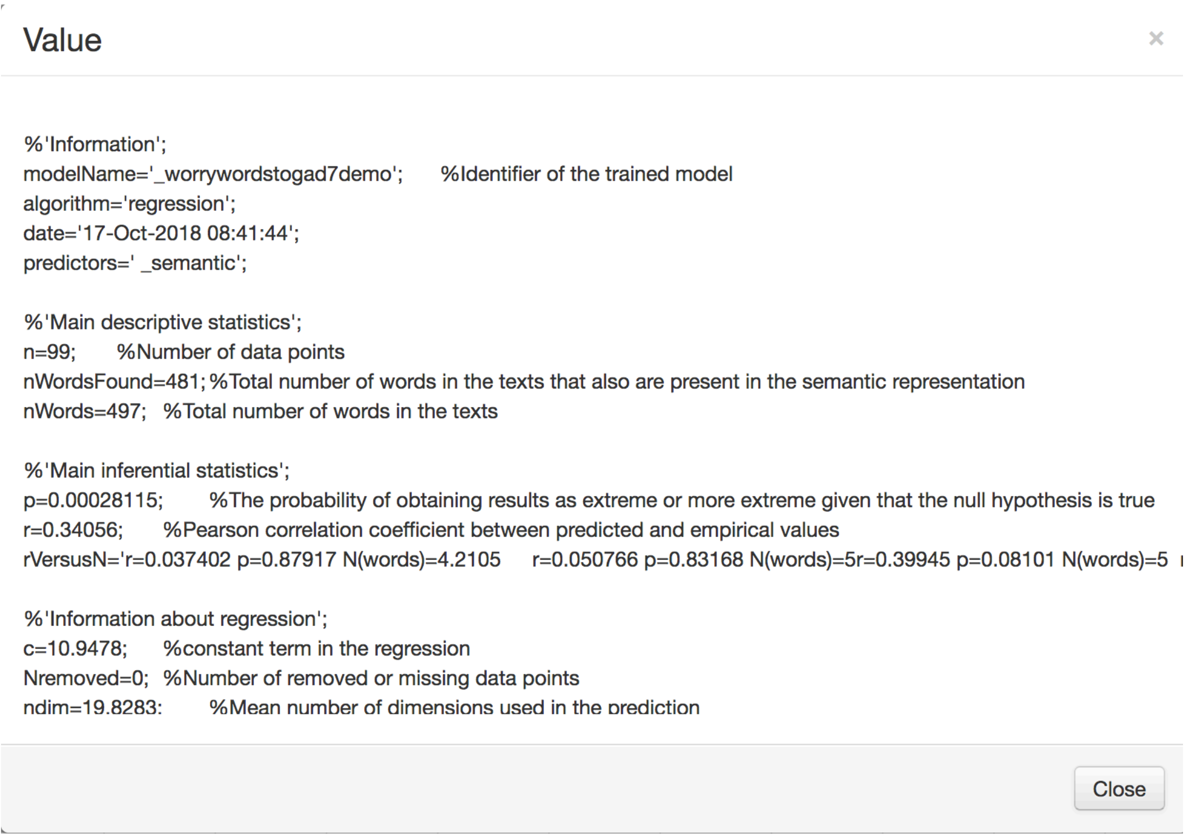
Prediction name

Worry words to GAD7 demo ☑ Public

Description

Worry responses using 5 descriptive words trained to GAD-7 in demo data

a



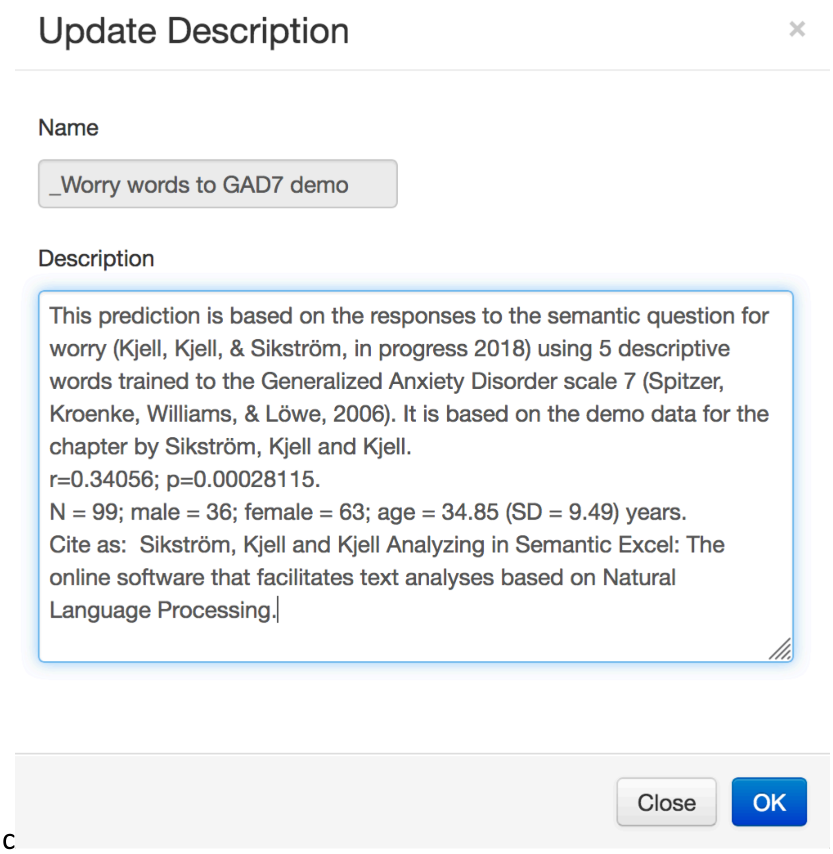
```
%'Information';
modelName='_worrywordstogad7demo';      %Identifier of the trained model
algorithm='regression';
date='17-Oct-2018 08:41:44';
predictors=' _semantic';

%'Main descriptive statistics';
n=99;      %Number of data points
nWordsFound=481; %Total number of words in the texts that also are present in the semantic representation
nWords=497;   %Total number of words in the texts

%'Main inferential statistics';
p=0.00028115;      %The probability of obtaining results as extreme or more extreme given that the null hypothesis is true
r=0.34056;     %Pearson correlation coefficient between predicted and empirical values
rVersusN='r=0.037402 p=0.87917 N(words)=4.2105      r=0.050766 p=0.83168 N(words)=5r=0.39945 p=0.08101 N(words)=5

%'Information about regression';
c=10.9478;     %constant term in the regression
Nremoved=0;  %Number of removed or missing data points
ndim=19.8283;      %Mean number of dimensions used in the prediction
```

b



Name

_Worry words to GAD7 demo

Description

This prediction is based on the responses to the semantic question for worry (Kjell, Kjell, & Sikström, in progress 2018) using 5 descriptive words trained to the Generalized Anxiety Disorder scale 7 (Spitzer, Kroenke, Williams, & Löwe, 2006). It is based on the demo data for the chapter by Sikström, Kjell and Kjell.
r=0.34056; p=0.00028115.
N = 99; male = 36; female = 63; age = 34.85 (SD = 9.49) years.
Cite as: Sikström, Kjell and Kjell Analyzing in Semantic Excel: The online software that facilitates text analyses based on Natural Language Processing.

c

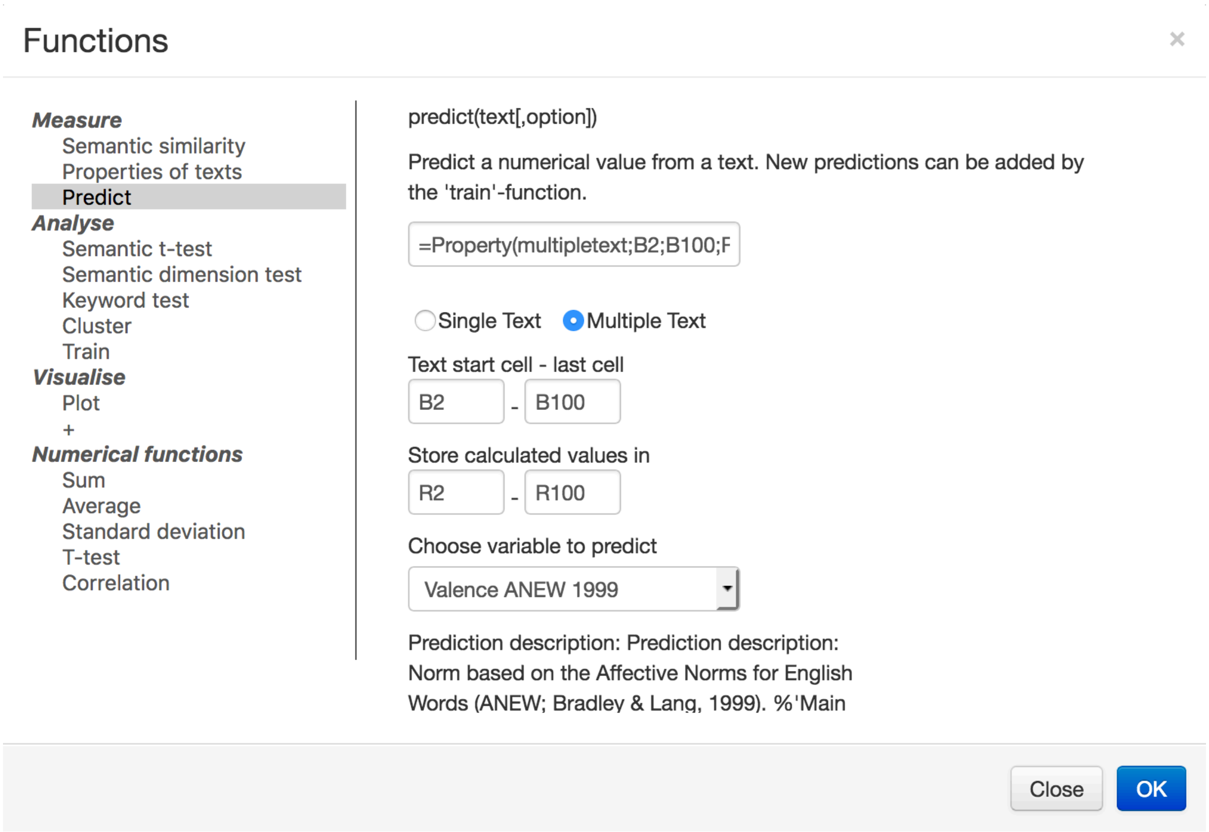
Figure 8a-c. Training semantic responses to rating scale scores and saving the model as a prediction.

Semantic prediction

Semantic predictions may be used to predict numerical values from texts, based on a previously trained model as described above. In the current demo sheet, we could for example apply the previously trained model named "Worry words to GAD7 demo" onto the depression word responses. However, in line with Chapter 5, we will next predict the valence of the worry word responses and then examine how these predictions correlate with the GAD-7. This will be achieved by using a public prediction called **Valence ANEW 1999**.

Go to **Functions** (Figure 9), and under **Measure** select **Predict**. Since we are going to predict the valence from several responses choose the **Multiple Text** option; in the **Text start cell – Last cell** boxes insert **B2** and **B100** and in the **Store calculated values in** boxes insert **R2** and **R100**. Lastly, in the **Choose variable to predict** dropdown list, select the **Valence ANEW 1999** prediction. ANEW stands for *Affective norms for English words* (Bradley & Lang, 1999) and comprises a word list of more than 1000 words that participants have rated using a rating scale based on figures symbolizing negative to positive affect. The description of the Valence ANEW 1999 prediction states that the training of the word list to the rating scale valence scores achieved a correlation of r = .74, (*p* < .001) between the actual and predicted values. Clicking **OK** will return the valence predicted values in column R. To find out the correlation between the predicted valence scores and the GAD-7 scores, go to **Correlation** under

Numerical functions in **Functions** and insert the cells holding the GAD-7 scores (i.e., **K2** through **K100**) and the predicted valence scores (i.e., **R2** through **R100**). This yields a correlation of r = -.32 (*p* = .0012).



predict(text[,option])

Predict a numerical value from a text. New predictions can be added by the 'train'-function.

=Property(multipletext;B2;B100;F

○ Single Text ● Multiple Text

Text start cell - last cell: B2 - B100

Store calculated values in: R2 - R100

Choose variable to predict: Valence ANEW 1999

Prediction description: Prediction description: Norm based on the Affective Norms for English Words (ANEW; Bradley & Lang, 1999). %'Main

Figure 9. Settings to Predict Valence from Worry Word Responses and Storing the Values in Column R

Summary

In this chapter, we have showed how to carry out analyses discussed in Chapter 4 and 5. This involved semantic similarity scales, semantic t-test, semantic training, semantic prediction and plotting words. We have also demonstrated how to save word norms and semantic predictions, and how to make them public for others to use. We have also briefly introduced some advanced settings to show the flexibility and many alternatives that Semantic Excel holds. We hope that Semantic Excel can be a useful tool for researchers and professionals that are interested in measuring, plotting and statistically analyzing text and numerical data. To further understand Semantic Excel, we recommend that you do the exercises listed below.

Exercises

The following exercises are influenced by Chapter 4 (questions 1-3) and Chapter 5 (questions 4-6). Note that for some of the questions you have helpful hints in the end.

1. ***Semantic similarity scales: Is there a significant correlation between the semantic similarity values between depression word responses and the depression word norm and the PHQ-9?***
   a) Compute the semantic similarity values between depression words and the depression word norm.
   b) Correlate the depression semantic similarity values with the PHQ-9.

2. ***Plot using the semantic similarity scale: Visualize depression and anxiety word responses according to the depression semantic similarity scale.***
   a) Plot words comparing depression words and worry words on the x-axes and the semantic similarity scale between depression words and the depression word norm on the y-axes.
   b) Try changing color of the words in your plot.
   c) Try remove or add dotted lines between the words and their associated cross.

3. ***Semantic t-test: Is there a significant difference between the depression word responses by females versus males?***
   a) Compute a semantic t-test comparing the depression word responses by females and males$^{\text{Hint 1}}$.

4. ***Semantic training: To which of the four included rating scales do depression word responses have the strongest relationship.***
   a) Train the depression word responses to the PHQ-9, GAD-7, HILS and SWLS and compare the four result outputs.

5. ***Semantic training: Are there statistical relationships between depression word responses and a) age and b) gender.***
   b) Is there a relationship between word responses and participants' age? Train the worry and depression word responses to the age variable.
   c) Is there a relationship between text responses and gender? Train the depression word responses to gender$^{\text{Hint 2}}$. Compare the result with exercise 3.

6. **Semantic prediction: Are the semantic predicted a) valence and/or b) arousal from depression word responses correlated with PHQ-9.**
   a) Predict the level of valence of the depression word responses. Correlate the semantic trained valence scale with the PHQ-9.
   b) Predict the level of arousal of the depression word responses. Correlate the semantic trained arousal scale with the PHQ-9.

Hint 1: The data has been arranged according to gender, so that males are positioned in B2-B50 and females are in B51-B100.
Hint 2: Remember to change from linear regression to logistic regression in advanced settings.

References

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Retrieved from

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*(1), 71-75.

Kjell, O. N. E., Daukantaitė, D., Hefferon, K., & Sikström, S. (2016). The Harmony in Life Scale Complements the Satisfaction with Life Scale: Expanding the Conceptualization of the Cognitive Component of Subjective Well-Being. *Social Indicators Research, 126*(2), 893-919. doi:10.1007/s11205-015-0903-z

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2018). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, No Pagination Specified-No Pagination Specified. doi:10.1037/met0000191

Kroenke, K., Spitzer, R. L., & Williams, J. W. B. (2001). The PHQ-9: validity of a brief depression severity measure [Electronic version]. *J Gen Intern Med, 16*(9), 606-613.

Sikström, S., Kjell, O. N. E., & Kjell, K. (2018, October 25). Semantic Excel: An Introduction to a User-Friendly Online Software Application for Statistical Analyses of Text Data. https://doi.org/10.31234/osf.io/z9chp

Spitzer, R. L., Kroenke, K., Williams, J. W. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The gad-7. *Archives of Internal Medicine, 166*(10), 1092-1097. doi:10.1001/archinte.166.10.1092