CPIDR® 5.1 USER MANUAL

Michael A. Covington

caspr

COMPUTER ANALYSIS OF SPEECH
FOR PSYCHOLOGICAL RESEARCH

Institute for Artificial Intelligence
The University of Georgia
Athens, Georgia 30602-7415 U.S.A.
www.ai.uga.edu/caspr

2012

CPIDR® 5.1 User Manual

Michael A. Covington
Institute for Artificial Intelligence
The University of Georgia

2012 August 17

Introduction

CPIDR® 5.1 (Computerized Propositional Idea Density Rater, pronounced "spider") is a computer program that determines the propositional idea density of an English text automatically.

It is well known that propositional idea density, in the sense of Kintsch (1974) and Turner and Greene (1977), can be approximated by the number of verbs, adjectives, adverbs, prepositions, and conjunctions divided by the total number of words (Snowdon et al. 1996). In earlier papers (Brown et al., 2007, 2008), we refined this technique and used a part-of-speech tagger, plus adjustment rules, to obtain accurate idea density measures. CPIDR 3 is the latest product of this research program.

Authorship and version history

The name CPIDR is a registered trademark of the University of Georgia Research Foundation, Inc. It has been applied to several programs.

- The first CPIDR (with no version number) was a prototype idea density rater implemented in Prolog by Cati Brown;

- CPIDR 1 was a Java program implemented by Tony Snodgrass, using a somewhat more sophisticated rule set (Brown et al. 2007);

- CPIDR 2 was the same program, ported to C# by the same author and using the same rule set;

- CPIDR 3 was an early version of the current CPIDR program, coded in C# by Michael A. Covington and using a considerably refined rule set described further by Brown et al. (2008), and was the first CPIDR released to the public.

- CPIDR 4 added numerous co-authors but was not released.

- CPIDR 5 is the current product, and specifically version 5.1 is the version released in 2012 and documented here.

Coding done by CPIDR 5.1 is in most circumstances slightly more accurate than that done by CPIDR 3. However, the main reason CPIDR was revised was to improve performance and remove dependence on outside software. For scientific integrity, **when using CPIDR in research, you should always give the exact version and date,** which are displayed when you select Help, About CPIDR in the main menu. The version is also written at the beginning of each saved output file.

Differences between CPIDR 3 and CPIDR 5.1

CPIDR 3, which remains available at www.ai.uga.edu/caspr, is open-source freeware subject to GPL. It incorporates MontyTagger, which is itself GPL freeware. CPIDR 5.1 is proprietary software belonging to the University of Georgia Research Foundation, Inc. (UGARF). It is not freeware and is not open-source.

CPIDR 5.1 does not rely on an external tagger. It is self-contained and should run faster and more reliably in virtually all settings. In particular, the three-minute startup delay of CPIDR 3 is gone; CPIDR 5.1 starts much more quickly.

CPIDR 3 could (often) run under MacOS or Linux using Mono, but CPIDR 5.1 is for Windows only and requires .NET Framework 3.5 SP1 (included in all current, properly updated versions of Windows).

CPIDR 5.1 includes a 95% confidence interval for every idea density measurement. This will be described in what follows.

CPIDR 5.1 can be called by other software (using the file CPIDRmethods.dll). For information about how to do this, contact the authors.

Installation requirements

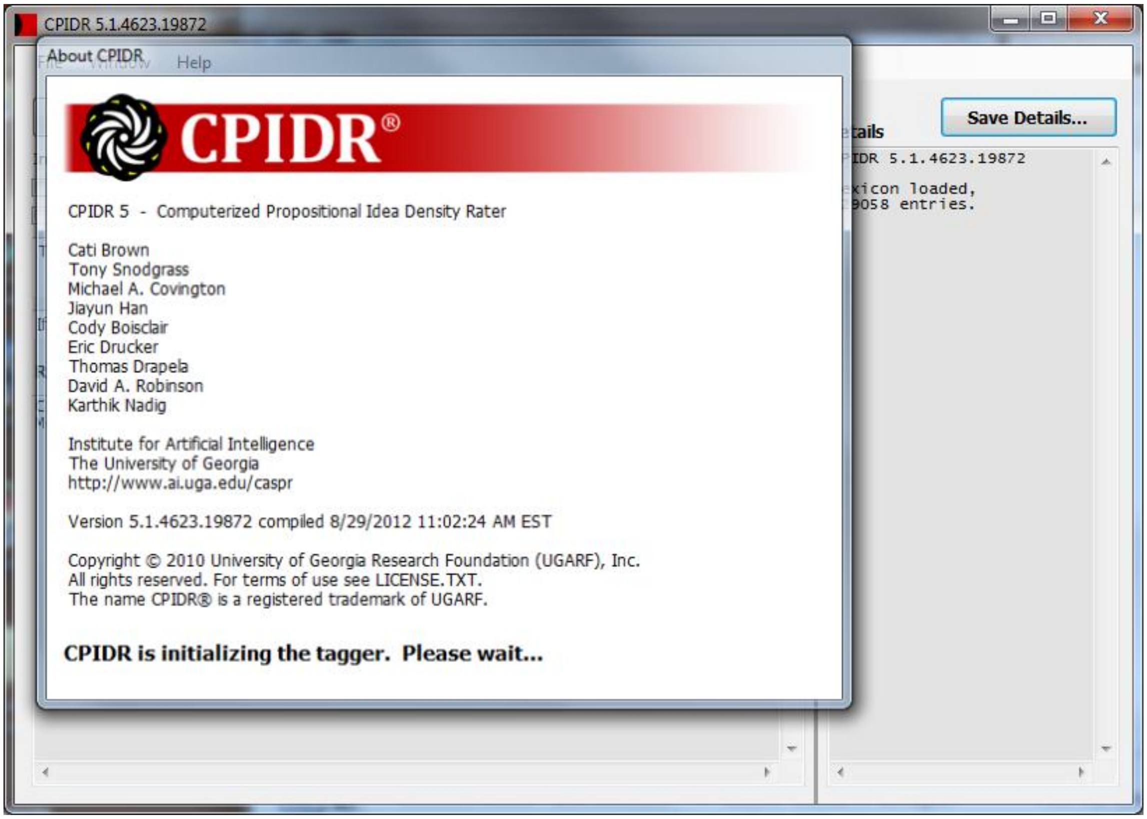
CPIDR 5.1 runs on Windows XP or later using .NET Framework 3.5 SP1, with both 32-bit and 64-bit CPUs. To install CPIDR, simply launch the supplied MSI file. During installation, you will be prompted to download .NET Framework from Microsoft if you do not already have it.

As input, CPIDR 5 accepts ASCII or Unicode text files or input typed on the keyboard or pasted from the Windows clipboard. "Smart quotes" (the characters " " ' ') as well as ASCII quotes (' ") are acceptable.

Basic operation

When CPIDR 5.1 is installed, a shortcut to it is placed in your Programs menu.

When you launch CPIDR 5.1, there will be a brief pause while the tagger is loaded and configured. During this time, a splash screen giving basic information about CPIDR is displayed:



```
IDR 5.1.4623.19872

xicon loaded,
9058 entries.
```

CPIDR 5 - Computerized Propositional Idea Density Rater

Cati Brown
Tony Snodgrass
Michael A. Covington
Jiayun Han
Cody Boisclair
Eric Drucker
Thomas Drapela
David A. Robinson
Karthik Nadig

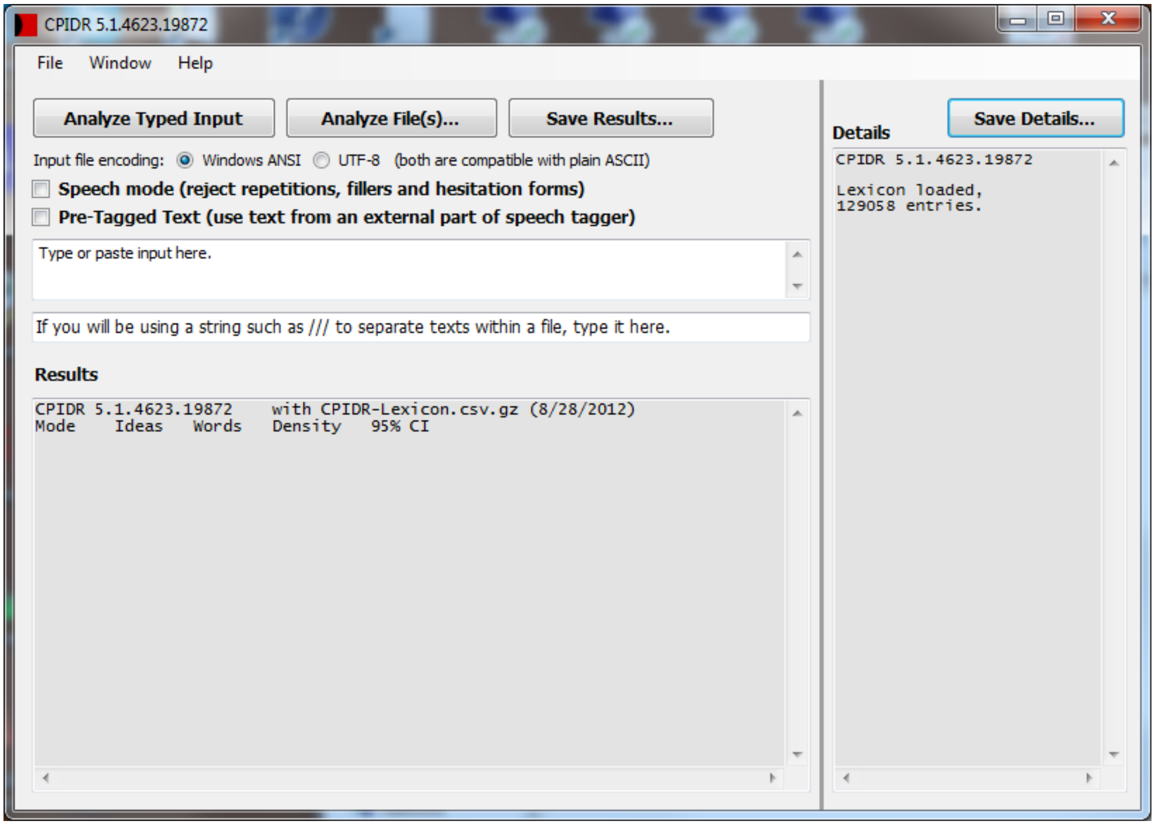
Institute for Artificial Intelligence
The University of Georgia
http://www.ai.uga.edu/caspr

Version 5.1.4623.19872 compiled 8/29/2012 11:02:24 AM EST

Copyright © 2010 University of Georgia Research Foundation (UGARF), Inc.
All rights reserved. For terms of use see LICENSE.TXT.
The name CPIDR® is a registered trademark of UGARF.

CPIDR is initializing the tagger. Please wait...

The main CPIDR screen looks like this:



Input file encoding: ● Windows ANSI ○ UTF-8 (both are compatible with plain ASCII)

□ Speech mode (reject repetitions, fillers and hesitation forms)

□ Pre-Tagged Text (use text from an external part of speech tagger)

Type or paste input here.

If you will be using a string such as /// to separate texts within a file, type it here.

```
CPIDR 5.1.4623.19872    with CPIDR-Lexicon.csv.gz (8/28/2012)
Mode     Ideas    Words    Density    95% CI
```

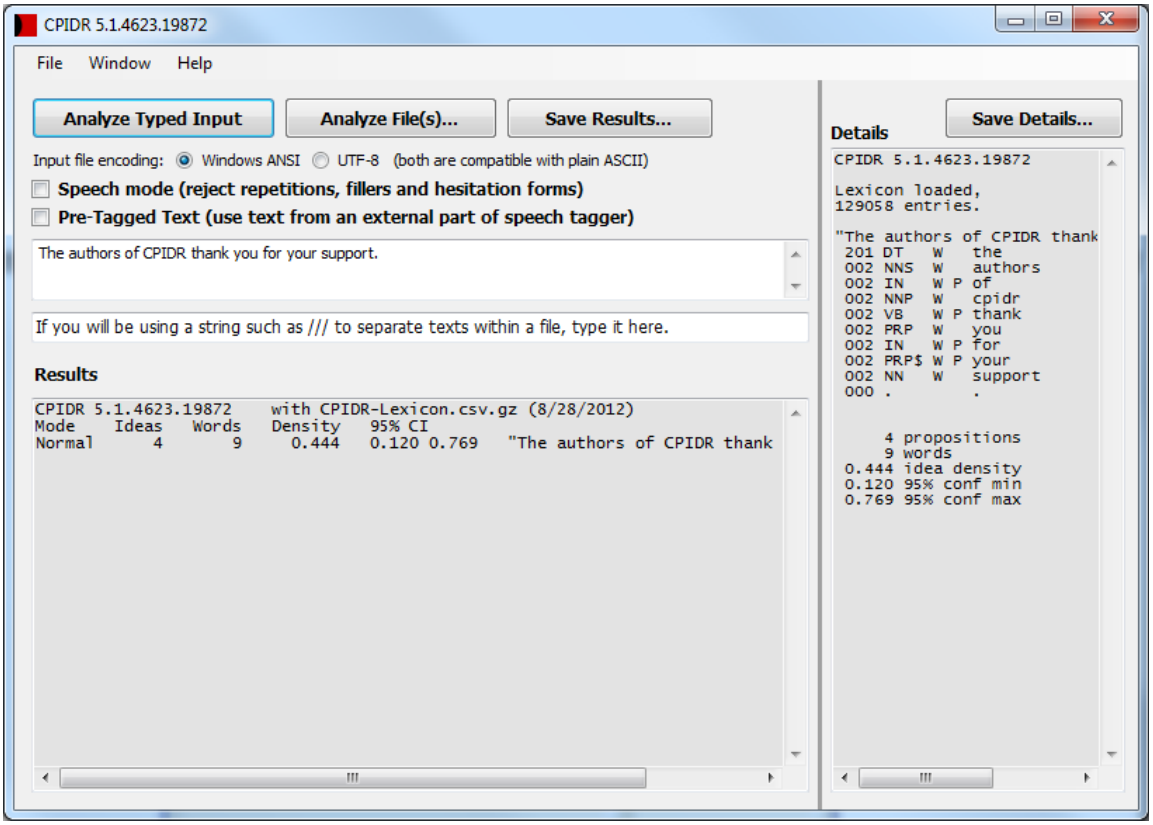
```
CPIDR 5.1.4623.19872

Lexicon loaded,
129058 entries.
```

Operation is fairly self-explanatory. Type one or more sentences into the white box, or paste them from the clipboard, and click “Analyze Typed Input,” or else place your input in text files and choose “Analyze File(s).”

In the latter case, you are allowed to select multiple files, and they will be processed in alphabetical order by full path and filename. If you wish, you can treat a single text file as containing multiple texts. In that case, type in, at the indicated place, the symbol, such as ---, that you will be using on a line by itself to separate texts.

Here is an example of the analysis of a sentence:



Input file encoding: ● Windows ANSI ○ UTF-8 (both are compatible with plain ASCII)

□ Speech mode (reject repetitions, fillers and hesitation forms)

□ Pre-Tagged Text (use text from an external part of speech tagger)

The authors of CPIDR thank you for your support.

If you will be using a string such as /// to separate texts within a file, type it here.

```
CPIDR 5.1.4623.19872    with CPIDR-Lexicon.csv.gz (8/28/2012)
Mode     Ideas   Words   Density   95% CI
Normal       4       9     0.444   0.120 0.769   "The authors of CPIDR thank
```

```
CPIDR 5.1.4623.19872

Lexicon loaded,
129058 entries.

"The authors of CPIDR thank
 201 DT   W   the
 002 NNS  W   authors
 002 IN   W P of
 002 NNP  W   cpidr
 002 VB   W P thank
 002 PRP  W   you
 002 IN   W P for
 002 PRP$ W P your
 002 NN   W   support
 000 .        .


     4 propositions
     9 words
 0.444 idea density
 0.120 95% conf min
 0.769 95% conf max
```

The Results window shows the idea count (proposition count), word count, idea density, the 95% confidence interval, and an identifying string (the first 37 characters of the text, or if the text had some from a file, the filename). As you analyze more sentences and files, more lines are added to this window.

The Details window shows you how the sentence was analyzed:

```
"The authors of CPIDR thank you for your support."
 201 DT   W   the
 002 NNS  W   authors
 200 IN   W P of
 002 NNP  W   cpidr
 200 VB   W P thank
 002 PRP  W   you
 200 IN   W P for
 200 PRP$ W P your
 002 NN   W   support
 000 .        .
```

Here 000, 002, 200, and 201 are the rules (in CPIDR's rule set) that acted upon each word; DT, NNS, IN, etc., are part-of-speech tags; W and P indicate which items were counted as words and as propositions. We recommend that you look briefly at the Details window to make sure words and propositions are being counted correctly.

95% confidence intervals

Every idea density measurement is accompanied by a 95% confidence interval. This is something you can ignore if all you want to know is the idea density for a specific text.

If, however, the text that you are analyzing is a sample of something much larger – if, for example, you are trying to characterize an author's style from one sample of his or her writing – then the 95% confidence interval tells you how much the idea density of the much larger, unobserved set of texts could differ from what you have seen so far. Specifically, there is a 95% probability that the idea density of the much larger set is between the values shown. If you are analyzing a small sample, the interval is large; it becomes smaller if the sample is larger.

This is based on a familiar statistical calculation, the standard error of a proportion.

Speech mode

If you check "Speech mode" in the main window, CPIDR will reject most repetitions (i.e., will not count them as new propositions, though they remain in the word count) and will reject hesitation forms and interjections more aggressively, as is appropriate for unedited transcribed speech.

How to save results to a file

The "Save Results" button lets you save the contents of the Results window as a tab-delimited text file suitable for importing into Excel. The "Save Details" button saves the detailed analysis onto a file.

You can also use the mouse and right mouse button to copy material from the Results or Details window to the clipboard, then paste it into another program.

On the main menu, “Window, Clear Output Windows” clears all the displayed results so that you can start afresh.

CPIDR in a DLL

If you want to write software that uses CPIDR’s idea-density measurement methods, contact UGARF for technical information and examples. The measurement methods are located in file CPIDRmethods.dll and can be called by other software.

How CPIDR works

The premise of CPIDR is that although it is *roughly* correct to equate every verb, adjective, adverb, conjunction, and preposition with an idea (proposition), numerous readjustment rules are needed to get an accurate count. CPIDR 3 does not understand every sentence in full and therefore does not produce perfect proposition counts, but it has been shown to be more reliable than most if not all human raters.

The part-of-speech tags are those of the Penn Treebank (Santorini 1995; not later versions). The most important ones are:

.	sentence-ending punctuation
CC	coordinating conjunction
CD	cardinal number
DT	determiner
IN	preposition, except *to*
JJ, JJR, JJS	adjective (positive, comparative, superlative)
MD	modal verb
NN, NNS	noun (singular, plural)
RB, RBR, RBS	adverb (positive, comparative, superlative)
TO	*to* (preposition or infinitive)
VB, VBZ, VBD, VBN, VBG	verb (various forms)

The full set of readjustment rules of CPIDR 3 (the open-source version) is documented in the sourthe file *IdeaDensityRaterRules.cs* which is installed with CPIDR 3 (in the *src* folder). This file is copiously commented so that non-programmers can read it. It is not delivered with CPIDR 5, which uses a somewhat refined set of rules.

Many of the rules condense complicated verb phrases into single propositions. For example, *may have been singing* is just one proposition (following Turner and Greene, 1977, who do not treat tense or modality indicators as propositions). *May not have been singing* is two propositions, not five.

Subject-aux inversion is undone in order to handle questions correctly. For example, *Has he resigned?* is changed to *he has resigned* so that subsequent rules handling *has resigned* will apply. In the Details window, this is displayed as:

```
“Has he resigned?"
 002         has/moved
 002 PRP  W   he
 402 VBZ  W   has
 200 VBD  W P resigned
 000 .        ?
```

indicating the original and moved positions of *has*.

In some cases, an auxiliary verb moves too far; for example, *Is he president?* is changed to *he president is*, but the proposition count is still correct.

The accuracy of CPIDR

For detailed tests of CPIDR 3 see Brown et al. (2008). CPIDR 5 is very similar but slightly more accurate on the example sentences of Turner and Greene (1977). We do not go along with Turner and Greene on every sentence.

CPIDR 3 always counts Verb + Preposition + Noun Phrase as two propositions (treating *come to Boston* exactly like *sing in Boston*). Turner and Greene usually do the same, but they do not count *to* as a proposition in their sentences 2 (*Fred went to Boulder*) and 53 (*...refusing to come to the party*).

In Turner and Greene’s sentence 46 (*Jimmy ate an orange and a banana*), the tagger mistakenly tags *orange* as an adjective, leading CPIDR 3 to count an extra proposition. All taggers are approximate, and by preferring to take *orange* as an adjective, the tagger gets correct results more often in other English text.

CPIDR (all versions) tends not to count propositions consisting of attributive nouns. This may be correct behavior. Turner and Greene’s sentence 52.f is:

The fine-quality wool of the Merino sheep causes their popularity.

CPIDR 3 counts 3 propositions here; CPIDR 5 counts 4 propositions; and Turner and Greene count 5, taking “Merino” to be like an adjective. We think it is quite likely that “Merino sheep” functions as a single compound word for most readers and speakers of English who know that it means.

References

Brown, Cati; Snodgrass, Tony; Covington, Michael A.; Herman, Ruth; Kemper, Susan J. (2007) Measuring propositional idea density through part-of-speech tagging. Poster presented at Linguistic Society of America, Anaheim, California. Available at: http://www.ai.uga.edu/caspr.

Brown, Cati; Snodgrass, Tony; Kemper, Susan J.; Herman, Ruth; and Covington, Michael A. (2008) Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods* 40 (2) 540-545.

Frijters, Jeroen (2004) IKVM, an implementation of Java for Mono and the .NET Framework. http://www.ikvm.net (and SourceForge).

Kintsch, W. A. (1974) The representation of meaning in memory. Hillsdale, NJ: Erlbaum.

Liu, Hugo (2004) MontyLingua: An end-to-end natural language processor with common sense. http://web.media.mit.edu/~hugo/montylingua.

Santorini, Beatrice (1995) Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). University of Pennsylvania.

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996) Linguistic ability in early life and cognitive function and Alzheimer’s disease in late life: Findings from the Nun Study. JAMA 275:528–532.

Turner, A., and Greene, E. (1977) The construction and use of a propositional text base. Technical report 63, Institute for the Study of Intellectual Behavior, University of Colorado, Boulder.