CASPR Research Report 2007-03

CPIDR 3 USER MANUAL

Michael A. Covington

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

2007

CPIDR 3 User Manual

Michael A. Covington
Artificial Intelligence Center
The University of Georgia

2007 August 17

Introduction

CPIDR 3 (Computerized Propositional Idea Density Rater, third major version) 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 an earlier paper (Brown et al., 2007), 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 has been applied to several programs:

- A prototype idea density rater implemented in Prolog by Cati Brown;
- A Java program implemented by Tony Snodgrass, using a somewhat more sophisticated rule set (Brown et al. 2007);
- The same program, ported to C# by the same author and using the same rule set (CPIDR 2);
- The current C# program, coded by Michael A. Covington and using a considerably refined rule set (CPIDR 3), described further by Brown et al. (in preparation).

CPIDR 3 is a product of the CASPR project (Computer Analysis of Speech for Psychological Research) at The University of Georgia. It is distributed as open-source freeware under the General Public License; see the file LICENSE.TXT, distributed with CPIDR, for particulars. CPIDR includes two additional open-source components, MontyLingua (Liu 2004) and IKVM (Frijters 2004), and inherits GPL from MontyLingua.

A future version of CPIDR will be self-containined, not relying on MontyLingua or IKVM.

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.

Installation requirements

CPIDR 3 runs on any Windows 2000, XP, or Vista system with .NET Framework 2.0 installed. To install CPIDR, simply launch the supplied MSI file. During installation, you will be prompted to download .NET Framework 2.0 from Microsoft if you do not already have it.

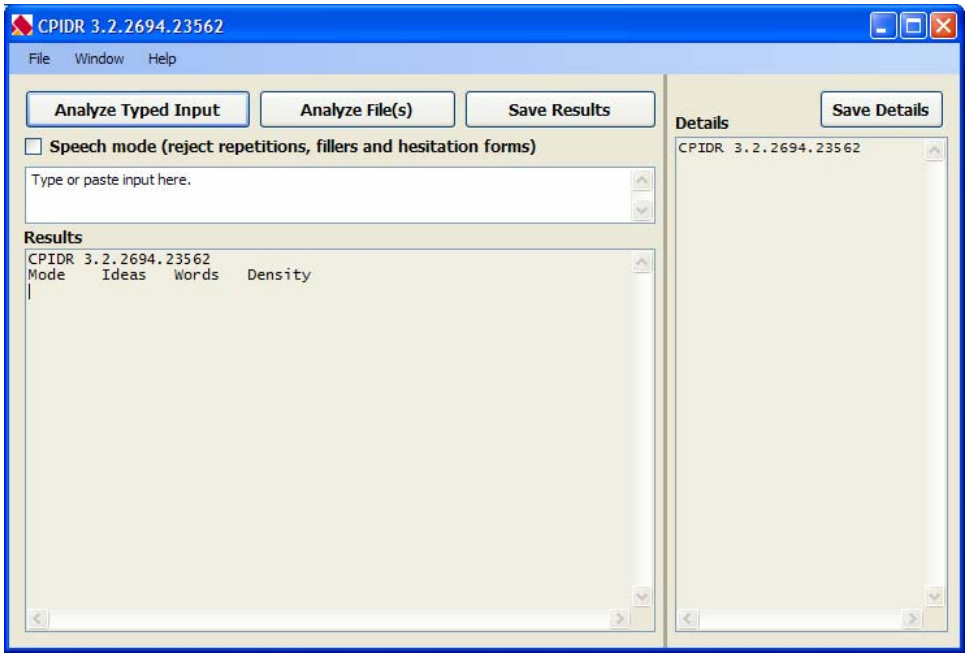
As input, CPIDR 3 accepts ASCII or Unicode text files or input typed on the keyboard or pasted from the Windows clipboard. “Smart quotes” (the characters “ ” ‘ ’) as well as ASCII quotes (' ") are acceptable.

Basic operation

When CPIDR 3 is installed, a shortcut to it is placed in your Programs menu.

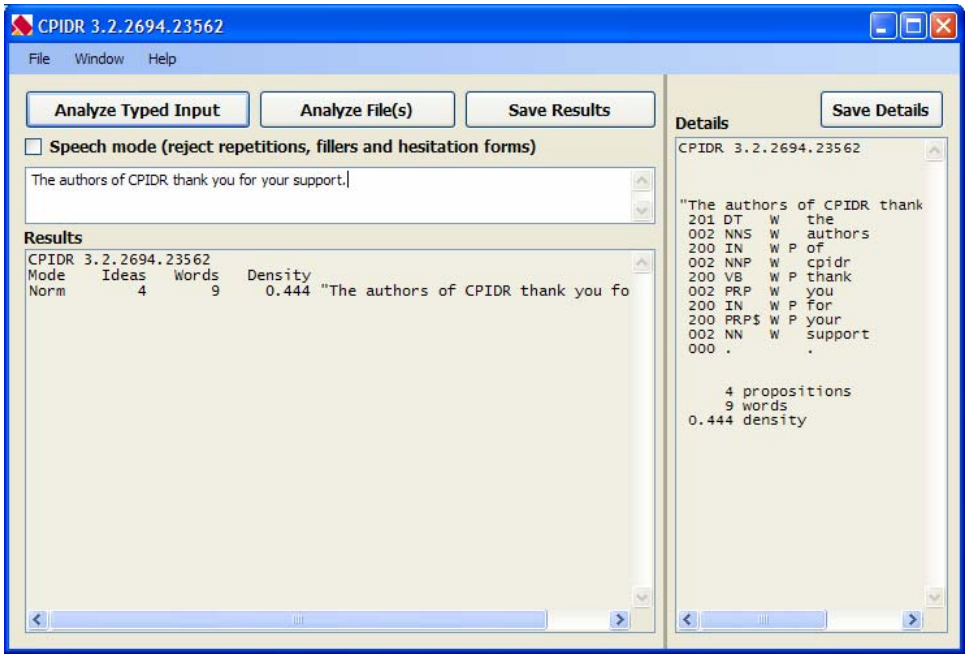
When you launch CPIDR 3, there will be a one- to three-minute pause while the MontyLingua tagger is loaded and configured. During this time, a splash screen giving basic information about CPIDR 3 is displayed.

The main CPIDR screen looks like this:



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.

Here is an example of the analysis of a sentence:



The Results window shows the idea count (proposition count), word count, idea density, 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.

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.

How CPIDR 3 works

The premise of CPIDR 3 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 is documented in the 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.

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 3

For detailed tests of CPIDR 3 see Brown et al. (in preparation).

CPIDR 3 agrees entirely with the proposition counts given by Turner and Greene (1977) for all but three of their 69 example sentences.

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 MontyLingua tagger mistakenly tags *orange* as an adjective, leading CPIDR 3 to count an extra proposition.

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. (in preparation) Automatic measurement of propositional idea density from part-of-speech tagging.

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.