



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

# Building specialised corpora from the web

Silvia Bernardini

EBRALC 2011



<http://bootcat.sslmit.unibo.it/>

- Basic pipeline
  - Select initial seeds (terms, keywords)
  - Query SE for random seed combinations
  - Retrieve pages and format as text (corpus)
  - Extract new seeds via corpus comparison
  - Iterate
  - Designed for translation students
  - Also used for reference corpus building
    - Leeds Internet Corpora

# NB!

Sketch Engine

Home Settings Change password

Search

user: adriano ferraresi used tokens: 0 / 1,000,000 days left: 747 Search in enTenTen

### French\_cuisine: WebBootCaT: Downloading data...

39%

Successfully processed files	20	Errors	19
Files remaining	59	- unable to retrieve	8
Data downloaded	1994 kB	- invalid content-type	0
Tokens retrieved	21,721	- file size out of range	0
Tokens per file (avg)	1,086	- cleaned file size out of range	11
Time elapsed	1:01	- keywords filter applied	0
Estimated time remaining	1:33	- unable to convert to text	0
Average file processing time	1.6 s	- duplicate	0

[Cancel processing](#)

For Help, press F1

# Our task

- *Stella's project*
- Our contribution
  - We will build a WaC corpus of recipes
    - from different countries/cultures
    - Written in English
    - For an English-language (international) audience
  - To serve as a reference corpus against which to study the way in which Brazilian culture is represented in English language cookbooks
  - Interesting on its own
    - Teaching, translation, Web studies, cooking!

# Target population

- **Language**
  - English
- **Recipe or not recipe?**
  - Cooking blogs?
  - Newspaper columns?

## Primary (sel.) criteria

- Geographic distribution
  - Continents? Countries?  
Regions? Languages?
- Or coverage of best known cuisines
  - French, Mexican...?
- Or UK/US/IR/CA/AU/NZ vs. rest of the world?
- How about kosher, /vegan/raw/fruitarian, ...?

## Secondary (desc.) criteria

- Favour variety
  - different Websites (from different locales)
  - different types of dishes (appetizers/mains/desserts; breads/soups/salads)
  - different lengths of texts
  - different subsets within primary subdivisions?

# Before we start

- Make sure you have
  - AntConc
  - BootCaT
  - A Bing API key
  - a text editor
- And you know where the BootCaT corpora folder is in your system

# Step 1. Build the “seeding” corpus

- We build a small corpus manually
  - 10 texts
  - representative of the target variety
  - As varied as possible
- What queries? (make note)
- Download texts (plain)
  - Single file (copy and paste)
  - Multiple files (save-as-text)



**Search the recipe you are looking for in the search box or  
browse the site**

**your  
Course!**

APPETIZERS



*Pesto Recipe - Pesto Genovese ( fresh basil Pasta Sauce)  
(for 5 persons)*

Ingredients: 10 handful of basil; 9 cloves of garlic; 50 gr walnuts; 50 gr pine-nuts; 50 gr ewe's milk cheese; 100 gr. parmesan cheese; salt; 1 glass of extra virgin olive oil.

To prepare with the old manner: pound the all in a mortar in order to melt it and get a sauce. Start with the basil and



09-20-2007

#1

## Chuck Love

Moderator



Join Date: Jan 2007

Location: PA

Posts: 1,699

### 😊 Another Ligurian Recipe - Salsa di Noci

There seemed to be some interest in the William-Sonoma Ligurian recipe for pesto with trofiette, so I thought I'd share another one. A sauce from that part of Italy. It's not their famous pesto but one just as common in Liguria, Salsa di Noci - Walnut Sauce. Walnuts are widely used in Italian cookery in a variety of preparations, some of the best known are in recipes from Liguria

#### Ingredients:

1/2 lb. walnuts, shelled  
2 oz. fresh, crustless bread  
3 tbs. extra virgin olive oil  
1 tbs. fresh marjoram  
salt  
1 clove garlic  
1 cup cream

#### Method:

Scald the walnuts, and peel off the skins.  
Dip the bread in water and squeeze out most of it.  
In a mortar (or a food processor "on pulse") pound the walnuts together with the bread, garlic, marjoram and salt to achieve a smooth paste.  
Place the mixture into a mixing bowl, drip in the olive oil, stirring constantly.  
Add the cream, stir well.  
This sauce is ready to use. In the original Ligurian recipe, soured milk, called *prescinseua* in dialect, is added instead of the cream. *Prescinseua* can be replaced with plain yogurt for a closer to authentic taste.



# My Italian cuisine “seeding” corpus

- [Venetian cuisine: culinary traditions and typical recipes of Venice ...](http://www.veneziasi.it/en/cuisine-venice-tradition/venetian-cuisine.html)

[www.veneziasi.it/en/cuisine-venice-tradition/venetian-cuisine.html](http://www.veneziasi.it/en/cuisine-venice-tradition/venetian-cuisine.html)

Venetian cuisine: venetian food and traditional recipes of Venice, how to make **venetian recipes**.

- [Venice Recipes and Curiosities from the Venetian Cuisine](http://venicexplorer.net/tradizione/cucina-veneziana/index.php)

[venicexplorer.net/tradizione/cucina-veneziana/index.php](http://venicexplorer.net/tradizione/cucina-veneziana/index.php)

Venice **Recipes** and Curiosities from the **Venetian** Cuisine. Up one level At first page  
Next page page 0 of 20, Testo italiano. Thanks to the forum "Gastronomia e ...

- [Art Of Venetian Cooking | Food & Wine](http://www.foodandwine.com/articles/art-of-venetian-cooking)

[www.foodandwine.com/articles/art-of-venetian-cooking](http://www.foodandwine.com/articles/art-of-venetian-cooking)

Daniela, Giampaolo's wife, uses generations-old **Venetian recipes**; Marika, who is married to Gianluca and runs a New York City catering company called ...

- [Cooks.com - Recipes - Venetian](http://www.cooks.com/Recipes-Venetian)

[www.cooks.com > Recipes](http://www.cooks.com/Recipes-Venetian)

Place shrimp and scallops in a large bowl. In a small bowl combine bread crumbs and olive oil, spoon over sea food. Mix gently to coat seafood ... Ingredients: 5 ...

# Step 2. Get seeds from your corpus

- Get keywords from AntConc
  - Open program
  - File=> open file=> *choose*
  - Word List=> start
  - Tool preferences=> keyword list=> choose files=> *ref\_corp\_euoparl\_en.txt*=> apply
  - Keyword List=> start
- Select top n (~50) by copying and pasting in a text editor
- Get rid of inappropriate ones
  - verbs better than nouns
  - general words better than specific ones
  - keep in mind your purpose!
- Save text file

# Keyword candidates

dough

salt

cup

minutes

baking

~~Biscotti~~

flour

oven

~~scungilli~~

inch

g

until

Add

jam

olive

brown

bowl

chopped

~~crostata~~

cups

garlic

pasta

pepper

~~sardines~~

sauce

teaspoon

sugar

sheet

eggs

bake

Ingredients

juice

lemon

parsley

~~Place~~

tbsp

grams

recipes

heat

mix

delicious

ingredients

F

gr

oz

salad

slices

recipe

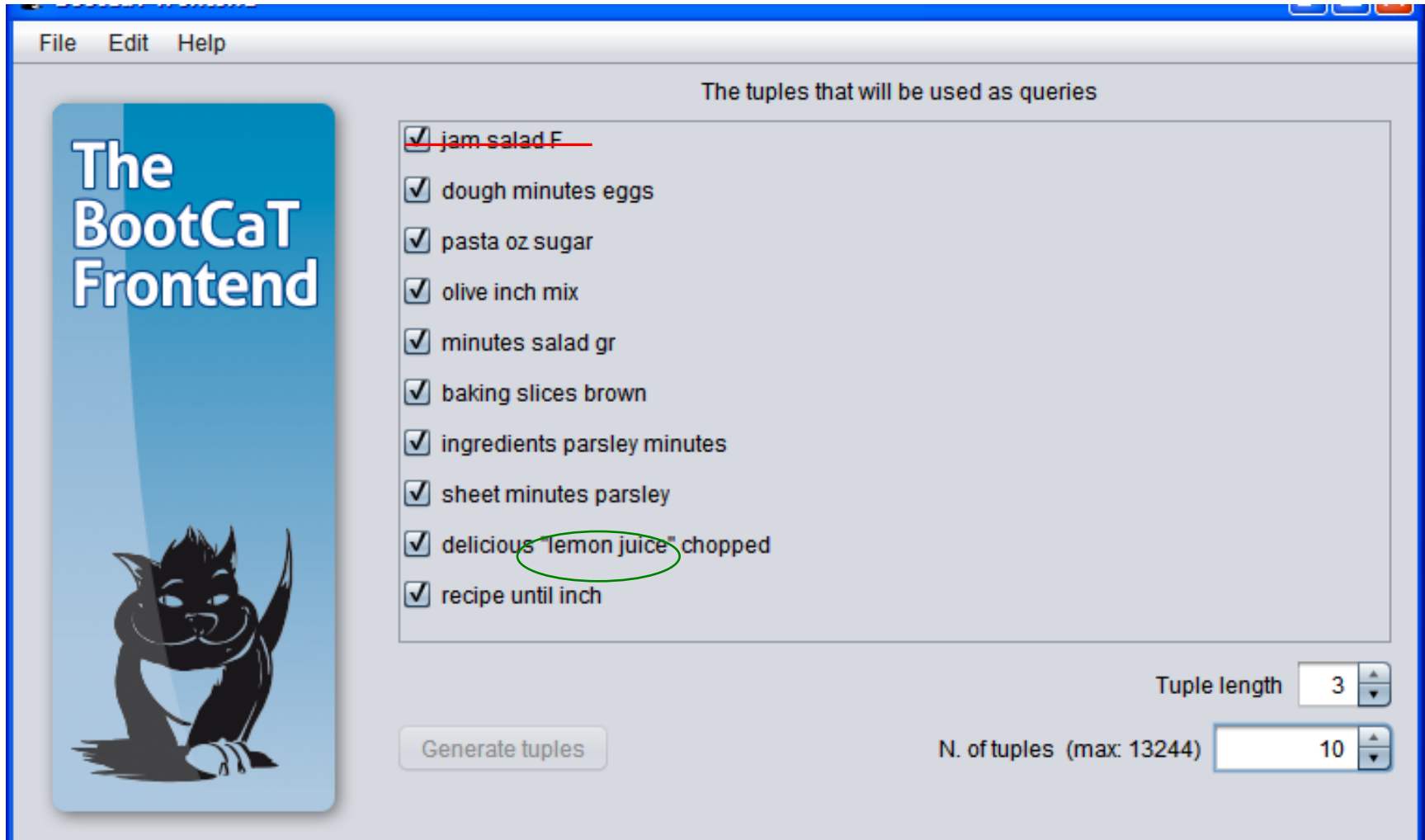
wine

water

# Step 3. Your first WaC corpus

- Open BootCaT
  - read welcome msg & click next
- Choose a name for your corpus
  - Mine is: it\_cuisine
- Select language
  - English United States (?)
- Ignore “more options”
  - Black lists and white lists
- Click next

# Step 3. Your first WaC corpus

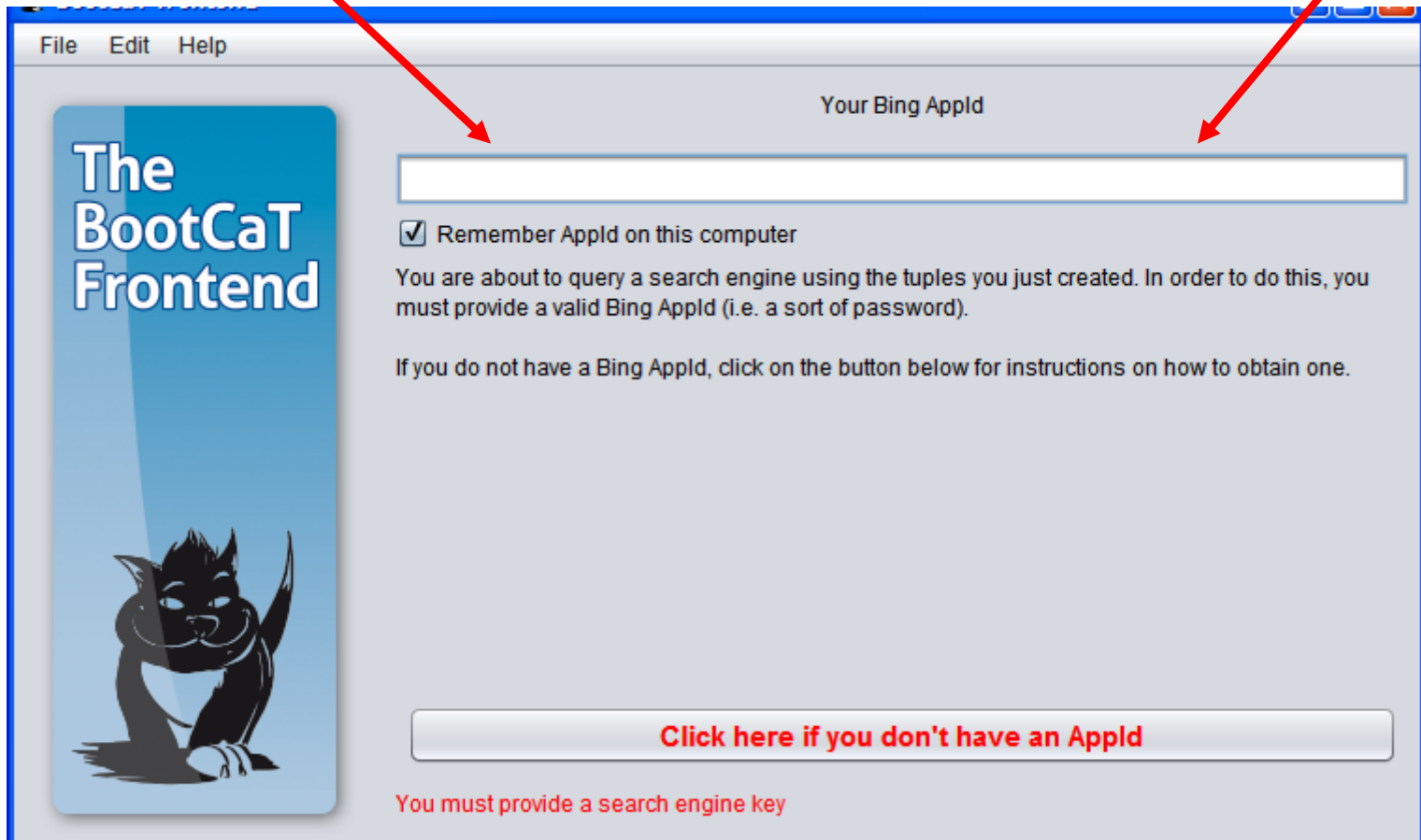


The screenshot shows the 'The BootCaT Frontend' application window. On the left is a logo featuring a black cat silhouette and the text 'The BootCaT Frontend'. The main area is titled 'The tuples that will be used as queries' and contains a list of ten items, each with a checked checkbox:

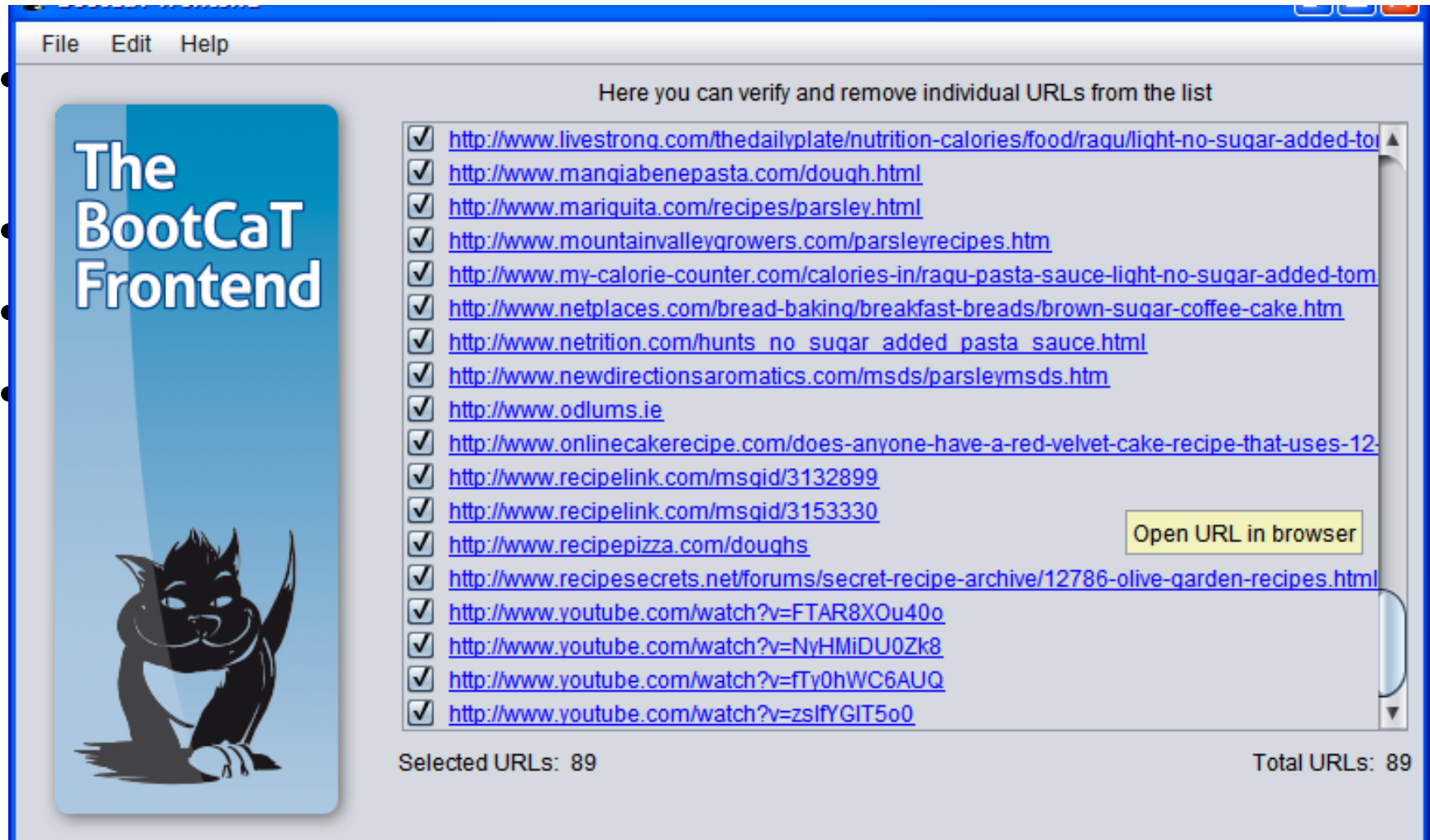
- ~~jam salad F~~
- dough minutes eggs
- pasta oz sugar
- olive inch mix
- minutes salad gr
- baking slices brown
- ingredients parsley minutes
- sheet minutes parsley
- delicious "lemon juice" chopped
- recipe until inch

At the bottom right, there are two controls: 'Tuple length' set to 3 and 'N. of tuples (max: 13244)' set to 10. A 'Generate tuples' button is located at the bottom left of the list area.

# Make sure you have a Bing AppId



# Step 3. Your first WaC corpus cont'd



The screenshot shows the BootCaT Frontend interface. On the left, there is a blue vertical banner with the text "The BootCaT Frontend" and a black silhouette of a cat. The main area is titled "Here you can verify and remove individual URLs from the list". It contains a list of 18 URLs, each with a checked checkbox to its left. A yellow button labeled "Open URL in browser" is positioned to the right of the list. At the bottom, it shows "Selected URLs: 89" and "Total URLs: 89".

File Edit Help

Here you can verify and remove individual URLs from the list

**The BootCaT Frontend**

- <http://www.livestrong.com/thedailyplate/nutrition-calories/food/raqu/light-no-sugar-added-to>
- <http://www.manqiabenepasta.com/dough.html>
- <http://www.mariquita.com/recipes/parsley.html>
- <http://www.mountainvalleygrowers.com/parsleyrecipes.htm>
- <http://www.my-calorie-counter.com/calories-in/raqu-pasta-sauce-light-no-sugar-added-tom>
- <http://www.netplaces.com/bread-baking/breakfast-breads/brown-sugar-coffee-cake.htm>
- [http://www.netrition.com/hunts\\_no\\_sugar\\_added\\_pasta\\_sauce.html](http://www.netrition.com/hunts_no_sugar_added_pasta_sauce.html)
- <http://www.newdirectionsaromatics.com/msds/parsleymsds.htm>
- <http://www.odlums.ie>
- <http://www.onlinecakerecipe.com/does-anyone-have-a-red-velvet-cake-recipe-that-uses-12>
- <http://www.recipelink.com/msgid/3132899>
- <http://www.recipelink.com/msgid/3153330>
- <http://www.recipepizza.com/doughs>
- <http://www.recipesecrets.net/forums/secret-recipe-archive/12786-olive-garden-recipes.html>
- <http://www.youtube.com/watch?v=FTAR8XOu40o>
- <http://www.youtube.com/watch?v=NyHMiDU0Zk8>
- <http://www.youtube.com/watch?v=fTy0hWC6AUQ>
- <http://www.youtube.com/watch?v=zsIfYGIT5o0>

Open URL in browser

Selected URLs: 89 Total URLs: 89

# Step 3. Your first WaC corpus cont'd

- At this point you can either
  - Go back and add the wrong URLs to the list of domains to be excluded
  - Manually untick the obviously wrong URLs
    - E.g. get rid of youtube and Amazon
  - Click “Next” and then “Build corpus”



wait...wait...wait...wait...wait...wait...wait...

wait...wait...wait...wait...wait...wait...wait...

wait...wait...wait...wait...wait...wait...wait...

wait...wait...wait...wait...wait...wait...wait...

wait...wait...wait...wait...wait...wait...wait...

**In the meantime...**

wait...wait...wait...wait...wait...wait...wait...

wait...wait...wait...wait...wait...wait...wait...

wait...wait...wait...wait...wait...wait...wait...

wait...wait...wait...wait...wait...wait...wait...

wait...wait...wait...wait...wait...wait...wait...

# Step 3. Your first WaC corpus cont'd

- Browse the corpus
  - click on “open corpus folder”
  - Open text file named “corpus” in text editor
    - Textpad, Notepad++, SciTE...
  - notice that all files (seeds, tuples) have been saved to a folder with the name you chose, within “BootCaT Corpora” folder
- Are you (reasonably) happy with the corpus you have built?
  - How many texts not belonging to target population?
    - non-recipes?
    - non English?
    - not about desired cuisine?
    - ...

# Step 4. Let us try again

- Not happy with the texts you downloaded?
  - Can try something different
- Go back to AntConc/your manual corpus
- Select “clusters”
- Tick “N-Grams”
  - Min. size: 3
  - Max.size 3
  - Start

# N-gram candidates

and bring to  
minutes or until  
a large bowl  
a slotted spoon

and cook for

bring to a

~~Chocolate Almond Biscotti~~

~~in water bay~~

juice in water

lemon squeeze the

Shut off heat

slotted spoon remove

squeeze the juice

~~the juice in~~

to a boil

water bay leaf

~~a boil Lower~~

and add a

~~Anthony Iannacone s~~

cookie sheet and

for minutes or

from the oven

~~Iannacone s Pizza~~

If you have

into a log

it will be

large bowl and

mix in the

of water and

on a lightly

or until the

over and bake

pinch of salt

pot place the

pound g of

Remove from the

~~s Pizza Rustica~~

salt and pepper

sheet and bake

teaspoon of salt

# Step 5. Repeat...

- Use the top 50 clusters as BootCaT seeds
  - Copy and paste, one per line
  - Remove those you don't like
  - No need to add quotes
- Remember to add blacklist websites
  - youtube.com, wikipedia.org, ...
- Remember to exclude “unlikely” tuples
- Remember to evaluate URLs before retrieving the corpus
  - Yahoo answers?
  - Blogspot?
- Build corpus



# Step 6. Fine-tuning

- Second corpus more about cooking (?)
  - Clusters identify genres better than keywords
  - Keywords identify topics better than clusters
- But still not about Vietnamese/Kosher/Vegan cuisine ☹️
- Solution: We can manually add seeds 😊
  - Think up an appropriate (word or) phrase
    - “Italian cuisine” “Italy” ...

# Step 6. Fine-tuning

- Go to the BootCaT corpora folder
- Open the “seeds” file for the corpus you have just built (it\_cuisine\_2)
- Copy the seeds
- Open BootCaT
- Create a new corpus (it\_cuisine\_3)
- Go through the usual steps
  - BUT for tuple length choose “2”
- Until you get to the “Collect URLs” screen
  - **DON'T CLICK COLLECT URLS!**

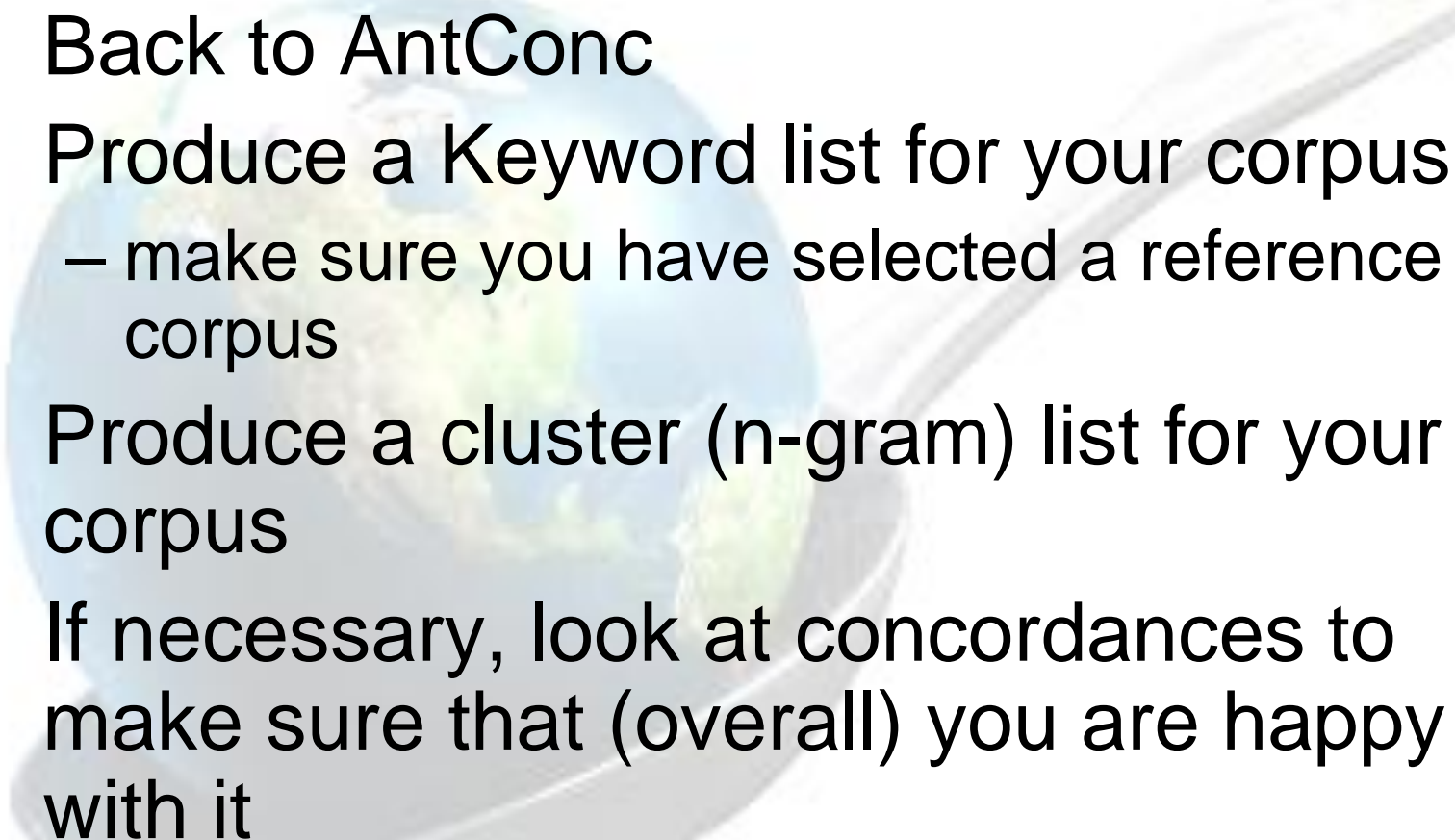


# Step 6. Fine-tuning

- Using a text editor, open the “tuples” file in the “~it\_cuisine\_3” BootCaT corpus folder
- Add manually to each line (query) the word or phrase specific to your cuisine (**in inverted commas!!**)
- Save, close and go back to BootCaT
- NOW click “Collect URLs”
- Edit the output if necessary
- Build corpus



# Time for some corpus analysis

- Back to AntConc
  - Produce a Keyword list for your corpus
    - make sure you have selected a reference corpus
  - Produce a cluster (n-gram) list for your corpus
  - If necessary, look at concordances to make sure that (overall) you are happy with it
- 

# Step 7. Going bilingual

- Back to AntConc
- Copy the top 20 keywords (or clusters) from the last corpus you built
- Translate them into Portuguese (using Google translate if you like)
- Add any other word or phrase that you think is relevant
  - Ricetta, ricette, ingredienti, preparazione, per 4 persone
- Repeat the whole corpus building procedure (used for the last corpus) for Brazilian Portuguese



[BSc \(Hons\) Surf Science and Technology](#)

[Course detail and](#)

[Career, further opportunities](#)

[More information](#)

[Entry requirements](#)

## The BootCaT Frontend



Limit search to the following Internet domain (e.g. .edu):

Exclude the following Internet domains (e.g. .com, wikipedia.org):

Maximum number of URLs to return for each query (i.e. tuple)

10

URLs returned by search engine

Collect URLs

## Presentation of the Master Course

A brief presentation of the Master of Science in Civil Engineering, entirely taught in English

### Presentation

The International Master Course in Civil Engineering is an international graduate program (Laurea



[Home](#)

[Prospect](#)

[How to apply](#)  
[International Students](#)

[Learning activities](#)

[Why this programme](#)

[Current students](#)

[Graduating students](#)

[Send to a friend](#)

# Evaluating/Reporting

- Carefully explain construction method
  - Reasons for choosing seeds
  - Chosen seeds
  - Number of tuples, iterations etc.
- keyword/cluster lists
  - Report / comment /compare
- Manually evaluate a sample of texts
  - Sharoff 2006
    - <http://corpus.leeds.ac.uk/serge/publications/wacky-paper.pdf>