Diatopix Exercise: Level I

Introduction

Diatopix, developed by Patrick Drouin at the *Observatoire de linguistique Sens-Texte* (OLST) at the University of Montreal, is a Web search tool that allows a user to view and compare search results for one or two search terms, and summarizes results in graph form, broken down by the countries of origin of the pages found. It can be very useful, for example, for users who want to study regionalisms or regional variants (either lexical or structural), or who want to compare the frequency with which two potential synonyms or quasi-synonyms are used. You can learn more about Diatopix on Patrick Drouin's Web page, http://www.mapageweb.umontreal.ca/drouinp/, and on the Web page of the OLST, http://olst.ling.umontreal.ca/.

Diatopix works by automating series of Yahoo! Web searches, and by restricting these searches to pages identified by their country of origin (the U.S., Canada, the UK, Ireland, Australia and New Zealand for English; France, Canada, Belgium, Switzerland and Luxembourg for French; Spain, Mexico, Chile, Argentina, Venezuela, Colombia, and Cuba for Spanish). Diatopix also allows for searches to be restricted (more or less) by domain using a very simple method based on co-occurrence.

Doing these exercises will help you learn to:

- Use Diatopix to search for information about words' or terms' usage,
- Understand how Diatopix queries are formulated, and
- Evaluate different challenges that can be encountered in searching using tools like this one.

Getting ready

1. Open the Web browser of your choice, for example, Internet Explorer (accessible using the shortcut that appears on the Windows Task bar) or Firefox (accessible from the Start menu**)**.
2. Open the page http://olst.ling.umontreal.ca/~drouinp/diatopix/index_en.html. The English Diatopix query page appears.

Searching the Web with Diatopix

1. Read the introductory information and the description of how the tool works.
2. Try the suggested query:
    a. Type *truck* in the first *String* field and *lorry* in the second field;
    b. Click on the **GEOGRAPHICAL DISTRIBUTION** button.

3. The results appear, presented in bar-graph form, with a bar representing the numbers of hits for each term per million pages, and a separate column for each region.
    a. What information can you draw from comparing the different frequencies for one of the terms? By comparing the results for the two terms?
4. If you want to save the graph for future reference, copy it to a Word document, PowerPoint presentation, etc.:
    a. Right-click on the graph;
    b. From the contextual menu that appears, choose the option **COPY**;
    c. Open your document;
    d. Paste the graph into your document (**CTRL** + **V**).
    e. Don't forget to also note any pertinent information about the graph that isn't shown (e.g., the date, any options you chose, your goal in doing the search, etc.).
5. Underneath the graph, a table presents the numbers of hits per million pages in each region. By clicking on the number of hits in each cell, you can link to the Yahoo! query that produced the results. Click on the value in a few of the cells to observe how the queries are formulated and defined and to view some of the pages found by the queries.
    a. Can you think of some challenges in using this kind of approach for searching?
    b. Are you likely to get some results that are not exactly what you wanted (*noise*)?
    c. Are you likely to miss some pertinent results (*silences*)?
6. Click on the browser's **BACK** button to return to the results page, and then again to return to the Diatopix query page.
7. Repeat the query, but this time with the plural forms of the two terms (*trucks*, *lorries*).
    a. Are the results similar to those for the singular forms? Is the distribution among the different countries and between the terms similar?
    b. Can variations in the numbers of hits tell you anything about the use of these terms in the singular and plural forms? Can you think of another way to test this kind of hypothesis using Diatopix?
8. Do another query, this time for the spelling variants *standardize* and *standardise*.
    a. Are the results more or less what you expected to see? What do they suggest to you about spelling variations between countries?
    b. What other forms should you query in order to get a complete portrait of the use of these verbs?
    c. What does this tell you about the challenges of searching for verbs using this kind of a tool?
    d. How are these challenges in English likely to compare to those in French and Spanish?
9. Try a new query, this time using the terms *lift* and *elevator*. Evaluate the results.
    a. Are they what you expected to see? Why or why not?

b. Can you think of any reasons why the results might not be exactly as you expected?

10. Click on some of the values in the table to view the pages that generated the results.
    a. Are your suspicions about the sources of difficulties confirmed?
    b. Did you find any unexpected sources of challenges?
    c. What does this mean about the way in which these kinds of search results should be considered and used?
11. Try a new query, this time using the terms *chip* and *crisp*. Evaluate the results.
    a. Are they what you expected to see? Why or why not?
    b. Do the pages found indicate the source of the difficulties here?
12. Repeat the query, but this time using *chips* and *crisps*.
    a. Are the results closer to what you had expected to find?
    b. Can you think of why the results might be somewhat different from those of the last query?
13. Repeat the query again, but this time with the expressions *bag of chips* and *packet of crisps* in the *String* fields.

    Note: There is no need to enter quotation marks around these expressions in the *String* fields. Diatopix will do this automatically when it generates the Yahoo! queries.

    a. Are the results closer to what you had expected to find? Why or why not?
    b. How does this change in the form of the query affect the numbers of occurrences found? Do you think the results are as reliable for this search as for the previous ones? Are they reliable enough to draw conclusions from?
14. Look at the pages found for *bag of chips* in the U.S. and the U.K.
    a. Does this expression mean the same thing in the British pages as in the American ones? How can you tell?
15. Repeat the query for *chip* and *crisp*, this time limiting the search to the domain of *Food* by selecting it from the pull-down *Domain* menu.
    a. How do the results compare to the previous sets?
16. Click on the values in one or more of the cells of the results table to view the queries that were used to generate these results.
    a. How is the domain indicated in the queries?
    b. How can restricting searches by domain help to improve results of these searches?
17. Repeat the *truck* and *lorry* query, but this time restricting it to the automotive domain by selecting *Automobile* from the pull-down *Domain* menu.
    a. How are the results different from the first ones? How many hits are found, in comparison to the original query?
    b. What are some of the challenges of using this approach to restrict the domain?
    c. Did the distribution between the two terms in the numbers of hits change? What could this be related to? Can you think of a way of testing your hypothesis using Diatopix?

18. Return to the Diatopix query page and click on one of the links to the other languages in which Diatopix is available.
19. Try out the suggested query in this language, and evaluate the results. Try out a few more queries, to look at some of the issues that interest you.
    a. Can you think of any reasons why the Diatopix approach could be more or less effective in this language? For a specific type of search or unit? In general?
    b. What interlinguistic variations could lead to different challenges in this language?

Questions for reflection

1. How can Diatopix help you with challenges you might encounter in translation or writing? In what kinds of situations?
2. Can you rely completely on the results of a Web-searching tool such as Diatopix? Why or why not?
3. What are some of the major challenges that you may encounter in using this kind of a tool (e.g., sources of noise or silences)?
4. Are these challenges unique to Diatopix? Or are they pertinent to other Web searching applications?
5. How can you try to avoid some of these challenges? What effects could these techniques have on the results and their reliability?
6. What is the main lesson that we can learn from these experiments, and that we can apply to the use of most — if not all — translation and terminology tools?