A Terminology Crosswalk for LABSTAT:

Mapping General Language Words and Phrases

To BLS Terms

Final Report to the United States Bureau of Labor Statistics

Stephanie W. Haas, Ph.D.

stephani@ils.unc.edu
School of Information and Library Science
University of North Carolina at Chapel Hill

September 29, 2000

Executive Summary

The overall goal of this project was to develop a reference resource to map common general language words and phrases for United States Bureau of Labor Statistics (BLS) data to the agency terms. The major deliverable is the LABSTAT Crosswalk (LSC); an information structure that organizes general language words and phrases into semantic clusters and subgroups, and then maps them to BLS terms and the information products in which they can be found. The LSC currently consists of thirty-five tables, or clusters, and lists approximately 4,400 words and phrases.

Part of the BLS mission, in common with other government agencies, is to disseminate information to citizens who need it. One method of dissemination is via the web. Finding information on a web site can be difficult and frustrating to people for a number of reasons. Many of these problems hinge on language problems - differences between what experts (i.e., BLS experts) and non-experts (i.e., the end users of the web site) call BLS-related concepts and data.

- The BLS may use the word in a more technical sense, which usually means that the definition is more restricted than when the word is used in general language.
- There can be more than one common phrase for a concept, and the BLS has settled on using a single phrase.
- The BLS is limited in the information it actually gathers, so that words that represent ideas within the scope of BLS information are at too fine a level of detail to be associated with available data.
- The BLS may gather and organize information in ways that do not directly correspond to words or ideas frequently used outside the agency.

All of these problems are the result of various flavors of language mismatch; the language used by end users does not correspond with the technical language used by the agency. The remedy that this work investigated was that of creating a crosswalk between general language and the BLS terminology; establishing mappings between the words and phrases that ordinary people use, and those used on the LABSTAT web page.

Three strategies were used to collect general language words and phrases for the end user side of the crosswalk.

- Interview BLS advisors and experts about their experiences with user terminology mismatches.
- Examine the actual communications from users when looking for information.
- Scan a wide range of published and broadcast information.

The LABSTAT Crosswalk (LSC) is structured as a collection of 4-column tables.

Column 1 contains general language words and phrases gathered from the sources listed above.

Column 2 will list the corresponding BLS terms.

Column 3 describes the semantic concept represented by a group of Column 1 words and phrases, from the general language perspective.

Column 4 will list specific surveys, tables, and other data resources in which the words, phrases, or concepts are applicable.

The LSC as it exists now is ready for additions and refinement by BLS experts.

- Some of the larger, less well-formed clusters (e.g., *Work Arrangements*) could be subdivided.
- BLS term mappings need to be entered into Column 2.
- BLS publications, tables, series, and other data products need to be entered into Column 4.
- For words and phrases that could lead to information outside the scope of BLS information, links or pointers could be provided to help users find the appropriate source. This is not feasible for all words and phrases, but certainly could be done for those that BLS consultants see most often.

Some consideration should be given to policies and procedures for maintenance, both to reflect changes in BLS information, and to enlarge the collection of general words and phrases.

Designs for the tools that will be based on the LSC can be developed. This includes

- Determining the functions of the tools, and their priority.
- Looking at the best way for end users to "enter' the crosswalk, for example, via semantic categories that are designed for browsing. This could entail investigating existing organizational structures that are familiar in everyday life. Any browsing structure would then need to be tested by a sample of non-expert users.

The idea of templates, or a semantic grammar for BLS information based on the LSC should be explored.

1. The Language Mapping Problem

Experts in any discipline develop a sublanguage for talking about their field - a set of words, expressions, and sentence structure that express the concepts, entities, and processes of interest. If non-experts need access to information in the discipline, one hurdle they face is that of terminology - finding the "correct" words. When members of the general public attempt to use the United States Bureau of Labor Statistics (BLS) LABSTAT, via the Selective Access Application (http://www.bls.gov/hlpselec.htm), they face precisely this problem. Users must formulate their queries by specifying values for a series of variables (e.g., industry, age, etc.) for which brief definitions are provided. But the names and terms that users have for data items may not be the same as the agency's terms for the same items. This type of communication problem is exacerbated when non-experts have a general understanding of the concept, but the technical, "expert" definition is somewhat different. For example, the BLS uses the term *income* to refer to all sources of money an individual accrues, including pay, rent, dividends, etc. In common conversation, people frequently use *income* to refer to money earned in exchange for work, usually in association with a job. In looking for information on *income*, therefore, the user may retrieve some tables or series, but they may or may not actually be what he or she is looking for.

This sort of terminology mismatch can cause many problems when retrieving, manipulating or even understanding the data.

- When an end user inputs a term that he or she uses for a concept, it may not match anything in the BLS information.
- When an end user inputs a term that he or she uses for a concept, it may retrieve information, but it may not be what was actually sought.
- An end user may actually find the information that will answer his or her question, but may not recognize it because it is called by an unfamiliar name.
- An end user may not fully understand the definitions, scope, or applicability of information that is found.
- It may be difficult for the end user to use help facilities, i.e., on the web page. This is the "dictionary problem" - you can't find the entry unless you know what it is!

The overall goal of this project was to develop a reference resource to map common general language words and phrases for BLS data to the agency terms. The major deliverable is the LABSTAT Crosswalk (LSC); an information structure that organizes general language words and phrases into semantic clusters and subgroups, and then maps them to BLS terms and the information products in which they can be found. The LSC currently consists of thirty-five tables, or clusters, and lists approximately 4,400 words and phrases.

The outline of this report is as follows. It starts by discussing issues of sublanguage and terminology in general and as they cause difficulties specifically with BLS terminology. Then it describes the processes for collecting and organizing general language words and phrases, and the rationale behind them. The structure of the LSC is presented next, along with some suggestions for how it could be used. Finally, I discuss specific term tables

and clusters, characteristics of the words and phrases that were collected, and implications for continued development and deployment of the LSC. The Appendix lists titles, a brief description, and any specific notes for each table. The full tables are not included in this report for reasons of space, but are available from Fred Conrad (Conrad_F@bls.gov) or me (stephani@ils.unc.edu).

2. Background

Part of the BLS mission, in common with other government agencies, is to disseminate information to citizens who need it. One method of dissemination is via the web. Finding information on a web site can be difficult and frustrating to people for a number of reasons.

- They may not be familiar with the structure of the web site, and where different kinds of information can be found. One reason this occurs is that web sites are often structured according to the internal structure of the organization, rather than in a way that would make sense to external users.
- Using search functions to find information may not be successful. A search term or phrase may return too many pages for people to scan. People may give up at this point, especially if the first few pages do not seem to be related to their need. On the other hand, users may not know what search terms might be useful, especially if they are searching for information for which the organization or agency uses its own technical terms.
- Even if the user finds the correct page or even table, he or she may not be able to recognize that this is what is needed, or be able to interpret it correctly.

Many of these problems hinge on language problems - differences between what experts (i.e., BLS experts) and non-experts (i.e., the end users of the web site) call BLS-related concepts and data.

The notion of *terminology*, or *jargon* is well-established. Experts in a field or discipline use special words or terms, which have specialized definitions, to communicate especially with other experts in their field. Experts need to have precise names for objects in their realm of discourse, and need to be able to make distinctions that non-experts may not perceive. Well-established terminologies include medical vocabulary, names for chemicals and related processes, and legal language. Frequently, words or phrases that are part of such terminology are recognizable to the non-expert as a specialized term, even if the definition is not known. The average person may not know what *blepheritis* means to a medical practitioner, but can tell that it sounds like "medical jargon" (Cabré, 1999; Pearson, 1998).

Terminology associated with BLS information is a slightly different, and in some ways, more complex problem. Many of the concepts of concern are familiar to the average person, and are things that we experience and talk about in everyday life, such as *wage* or *benefits* or *bonus*. Therefore, when a user needs information about one of these concepts,

he or she will search for it using these common words. Unfortunately, that may not always work, and the failure can take several, sometimes subtle forms.
The BLS may use the word in a more technical sense, which usually means that the definition is more restricted than when the word is used in general language. For example, the BLS distinguished between *income* and *wage* or *salary*. *Income* refers to the total incoming money a person accrues, regardless of the source. It may include rent, inheritance, dividends, etc. *Wage* or *salary*, on the other hand, are limited to regular payments a worker receives in exchange for his or her work. Most people recognize the association of *wage* or *salary* with a job (although the subtle differences about what is and is not included may not be understood). But we also use the word *income* to mean the same thing. So a user may request information on *police officer income*, intending to find information about how much a police officer earns, not how much money he or she gains from all sources. In general conversation, this request is usually interpreted correctly, or the speaker may be asked for clarification. In searching on a web page, either phrase makes sense, but different information will be returned. Even worse, the end user may not recognize that the answer is the answer to the wrong question; it will look plausible in either case. Riggs (1993) mentions the difficult relationship between common concepts and the use of general language words as names for them.

Another problem may arise when there is more than one common phrase for a concept, and the BLS has settled on using a single phrase. For example, in general language *health* and *medical* are treated as synonyms in a variety of phrases; *health care – medical care*, *health insurance – medical insurance*, and so on. As a component of the Consumer Price Index (CPI), however, the BLS uses the phrase *medical care*. This may not be a serious problem; given a list of phrases to browse, most people will recognize that the two mean the same thing. If a search does not happen to give the user the opportunity to see what terms are used, however, then he or she will be frustrated at not being able to find information that "should be there". The user may or may not think to use the common synonym on a second search attempt. Haas (1999a, 2000a) discussed these problems.

A third side to this problem is that the BLS is, of course, limited in the information it actually gathers. Any question an end user asks may need some negotiation or compromise to match data that is actually available. For example, it does not collect information on all possible occupations. It is important that the agency have clear definitions for those titles it does use, so that occupations may be categorized consistently over time and space. It takes time to develop these definitions, so all the latest occupations may not be covered. For example, the user may have to infer that *webmaster* is in the general category of mathematical and computer scientists, even though it is not specifically mentioned there. Further, there is a limit to the level of specificity that the occupation classifications can reach, both because of the effort involved in collecting data, and the need for a reasonable sample size. So a user may want information specifically about CIO's, but have to settle for information about executive, administrative, and managerial workers. Finally, of necessity the agency has settled on one scheme for classifying work, but there are many ways of thinking about occupations. Is a *ski instructor* a teacher or someone providing personal service?

A related problem concerns how information is gathered and presented. The most common example is the infamous *inflation*, which is in common use in discussing the economy in general, but which does not have a direct analogue in BLS data. Another common example of this has to do with definitions of race and ethnicity. Categories such as *Japanese-American* or *Arabic* are not identified by the BLS, although *Black* and *Hispanic* are.

Yet another problem is that of the end user who is looking for information that seems within the scope of BLS information, but is not. For example, the BLS does gather data on producer costs and prices, e.g., the cost to an enterprise of providing salary and benefits to its employees. But that does not mean that all producer data is available from the BLS. It does not gather data on corporate profits or dividends, for example.

All of these problems are the result of various flavors of language mismatch; the language used by end users does not correspond with the technical language used by the agency. The remedy that this work investigated was that of creating a crosswalk between general language and the BLS terminology; establishing mappings between the words and phrases that ordinary people use, and those used on the LABSTAT web page. A crosswalk is a data structure that captures the mapping of synonyms and near-synonyms between two sets of terms, in this case, the BLS terms on one side, and terms used by the general public on the other. I use the term "crosswalk", rather than "thesaurus", because a thesaurus may represent many relationships among terms other than synonymy, such as antonymy (opposite) or the part-of relationship. Based on my earlier study of BLS vocabulary (Haas, 1999b, 2000b), I hypothesize that there exist many-to-many mappings between some terms. An end-user may use a general term that could mean any of several BLS terms, similarly, a single BLS term may correspond to several different general language words.

3. Methods

3.1 Criteria

The original proposal for this work identified three criteria that the crosswalk should satisfy.

1. It should be built from the end user's perspective. The entry points should be terms that end users have used or might use, and they should lead to agency terms. This has the potential for building some bias into the crosswalk. For example, there may be some agency terms and concepts that users rarely want; these agency terms may not be included in the crosswalk. This would imply that there are areas of the agency information repository that are used only by the agency experts; they are either of no interest to the public, or the public is not aware of their existence (a different problem). This approach differs from that taken by Haas & Hert (2000), in that we started from the agency terms and found

corresponding general language words. Given that the goal of this work is to provide better help to end users, however, this potential for bias should not be a problem.

2. The crosswalk should provide broad, but shallow coverage of users' terms and concepts. This criterion is based on the 80/20 principle; if we can quickly provide shallow coverage of a large number of terms, then that will alleviate 80% of users' difficulties. The remaining 20% will require more in-depth coverage, or may always require consultation with an expert. (The 80/20 numbers are convention, not based on any rigorous measure of the terms and queries.) The broad but shallow approach should allow a quick start-up, providing an initial mapping on the term level (e.g., synonym or near-synonym pairs). Subsequent iterations could then provide coverage of the remaining (presumably rare) terms, along with definitions or examples of the concepts This is in contrast to Haas (1999b) and Haas & Hert (2000), where we examined a single complex concept and its associated terms in depth.

3. The crosswalk will be developed as either a flat file in Microsoft Word, or an SQL database.

The first criterion was met by considering where users find or learn about the words and phrases with which they attempt to identify information. In other words, what words and concepts do they have when they arrive at the LABSTAT page? As originally stated, the criterion mentioned the possibility that there were areas of BLS information that were sought only infrequently. It is clear that the complement to this problem can also occur; users come to the BLS looking for information that is not actually within the agency's purview. These issues are discussed later in this section.

As was indicated originally, there is really no sure way of measuring coverage, especially when one considers the range of language possible, as well as the sorts of confusions caused by users looking for information in the wrong place. However, this criterion can be considered to have been met successfully in a couple of ways.

- Words in some clusters can be listed thoroughly, if not exhaustively. Examples include geographical area terms and words referring to gender.
- The words and phrases that were found cover a range of register and precision. *Register* roughly refers to the formality of the language; *desk jockey* is a more informal (or even slangy) phrase than *office worker*.
- I was able to identify different classes of words and phrases, according to novelty, register, and technical expertise. This classification will aid in the future collection of words. The classes are discussed in Section 5.
- The collected phrases suggest that it may be possible to define a type of grammar, or set of templates, that describe possible combinations of word clusters. This is explored in Section 5, and may form the basis for my next work project.

Although these factors do not lead directly to a measurement of coverage, they do suggest some useful ways of thinking about coverage, and perhaps working around the coverage problem.

The third criterion refers to the structure of the final product; the resulting Microsoft Word table structure is described in Section 4.

3.2 Collection Strategies

I used three strategies to collect general language words and phrases for the end user side of the crosswalk.

3.2.1 Strategy 1: Interviews

The first strategy was to interview BLS advisors and experts about their experiences with user terminology mismatches. These interviews were fruitful in a couple of ways, providing general insight into some common problem areas. Some of these are fundamental to understanding the basic organization of BLS information; misunderstandings at this level can lead to multiple naming and interpretation problems. These include:

- Understanding the difference in perspective between establishment and household surveys. For example, in counting the number of jobs, or new jobs, the answer depends in part on whether you ask an existing company about new jobs, or ask individuals, who may be newly self-employed, but would not show up on a company payroll. Information about benefits is another area where the source of the information is important.
- "Terms that don't exist". This was one person's name for words and phrases that are commonly used for economic concepts outside of the agency, but which don't have a precise mapping to data within the agency. Everyone's favorite example of this is *inflation*.
- Concept access. People frequently want to find "everything you have about X", where X is some large concept such as "employment in the Midwest", or "union activity". Full answers to these questions require combining data from different surveys and even different agencies. The difficulty is that different sources may use different assumptions and definitions in data collection and analysis, e.g., for geographical area, that make meaningful combination questionable.
- Occupation titles and industry classifications. The level of specificity at which data is available was frequently cited as confusing or frustrating to end users. In addition, the technical definitions of occupation categories and industry classifications frequently conflict with common usage. A further related problem is the speed with which new occupations and businesses arise; it is impossible for the BLS to keep up to the minute.
- Information the BLS does not collect. The scope of BLS information is frequently confusing; it collects labor statistics, but not information about labor law. It has information about work-related injuries, but is different from OSHA. It has some information about industries, but does not look at commerce *per se*. These distinctions in coverage can be somewhat opaque to end users.

Other problems represent specific confusions about individual words or concepts. Examples included:

- Geographical regions and their names and definitions.
- Differences among types of data, e.g., *index, rate, ratio*. This often indicates a general unfamiliarity with statistical concepts themselves, as well as their names.
- *Seasonal adjustment*. This is a common phrase, and non-experts may have a vague idea of what it means or what it is for, but not have a clear idea of how it is done or why it changes what data is available.
- The difference between ethnicity and race.
- Definitions of important terms such as *unemployment* (which does not mean that one is collecting unemployment payments) or *earnings* (which may include some "extra" payments but not others)

More examples can be found in Section 5 and the Appendix.

3.2.2 Strategy 2: Communications from Users

The second strategy was to examine the actual communications from users when looking for information. There were two sources here: search logs from the FedStats web page, and part of an archive of email messages sent to the BLS from users. Difficulties in using search logs as a source of data have been discussed in other places (see for example, Hert 1999). For the purposes of this investigation, perhaps the most problematic issue is that although the words used can be seen, the intention behind the words, or what the individual actually wanted, cannot be known. On the other hand, this served as a good source of words that people actually use. As mentioned earlier, it is likely that a single word or phrase could map to more than one BLS term; this must therefore be treated as a common occurrence rather than an unusual event. If the user can see the possible interpretations that are available and choose among them, the lack of contextual information about his or her needs may not be a significant problem. For example, if someone enters *New York* as a geographical area name, it could refer to the city, the metropolitan area, or the state. The user may or may not realize the ambiguity. If he or she is then offered a choice of interpretations, along with a description of the data that is or is not available for each choice, he or she can identify which interpretation is most useful given the available data, regardless of whether that was the original intention.

The email archive was especially helpful. Individuals send email to the BLS staff when they are looking for information they cannot find (but assume exists somewhere), when they do not understand the information they have found, or when they are having trouble navigating the web site itself (among other reasons). Not only are the messages a good source of words or phrases, the phrases are embedded in sentences or even entire paragraphs. An added benefit is the length of the messages. It is well known that the phrases entered into web search pages are usually quite short - 2 or 3 words. This limitation was generally not seen in the email. The responses given to correspondents, which were also included in the archive, were also helpful in highlighting the ways in which a question could be mapped to BLS data. For example, a response could state in essence "I'm not sure what you mean by X, but if you mean Y, then here is where you look, if you mean Z, try this." (Of course, it is also possible that the questioner really

meant Q!) The main focus of the email analysis was to gather words and phrases, however, not to analyze the discourse.

3.2.3. Strategy 3: Published and Broadcast Information

The third source of words and phrases, and the most fruitful one, was a wide range of published and broadcast information. The rationale behind this approach was that people learn about labor-related information from a variety of sources, including the media, conversations with others, their job contract and paycheck, textbooks, economic specialists and their reports, etc. When they start to search for information on the web page, they come in with an understanding of agency terms and concepts that they have gained from their own life experience combined with these other sources. Therefore, I undertook to scan a variety of these sources to collect words and phrases in much the same way. There was no attempt to seek out only authoritative sources; on the contrary, the purpose was to find any words that could be considered relevant by a non-expert. One of the results of this strategy is that many of the words in the crosswalk refer to concepts that do not have direct correspondence with BLS data, but this accurately reflects the end user's experience. If he or she comes in looking for information on the earnings of *Chinese-American webmasters*, it is true that the question cannot be answered, but it is also true that since there is information on ethnicity and occupations, this is not a misplaced request. A function of the LSC will be to help support the negotiation and compromise necessary to find the information that comes closest to answering the question, and also to inform the user about what information is (and is not) available.

Another result of this strategy is that many of the collected words and phrases are more casual, or slangy than the "official" BLS terminology, or the discourse normally used by more formal sources. Words such as *cyberjobs*, *desk jockey* or *cube farm* may not be formal, but they are meaningful to some portion of citizens who want information about occupations and working conditions and should be treated as carefully as those phrased in more traditional ways. Indeed, someone who is more familiar with these types of names or descriptions may be quickly stymied by the BLS terminology; help may be especially crucial.

I approached the environmental scanning task in two ways. The first method deliberately targeted specific types of sources. In looking at these sources, I scanned all sections, columns or departments, not just those focused on business or the economy, since discourse on work-related topics can show up almost anywhere. Sources included:

- Mainstream newspapers, including the *Raleigh News & Observer*, the *New York Times*, the *Wall Street Journal*, and the UNC-CH student newspaper. This has been ongoing, (and I have been able to enlist the occasional help of friends and family members).
- General and specialty magazines. These included various weekly and monthly magazines aimed at a general audience, news magazines, business and economics magazines, women's magazines, environmental magazines, those aimed at sports fans, college and high school students, hobbyists, and so on. In general, I examined one year's worth of these.

- Tabloid-style newspapers and magazines. These are not necessarily "mainstream" or "authoritative", but do cover economic and labor-related topics. In addition, they are aimed as a somewhat different audience than the preceding sources. They are often written in a different linguistic register, and therefore use different vocabulary.
- Web sites. I looked at business and employment-related sites, as well as those associated with various news media. In general, I visited each site only once, although I looked at "back issues" dated pages when they were available. As before, I looked at both "mainstream" and "tabloid" sources, again noting differences in language and design.
- Radio and television. The deliberate sampling was aimed mainly at local and national news shows.

The second method of scanning was more opportunistic. I tried to pay attention to as much discourse surrounding me as possible, looking for additional words or phrases that could be interpreted as pertaining to BLS topics. Obviously, this coverage was spotty. An important advantage, however, was that it included conversations, both those in which I was a participant and those I merely overheard. People who were familiar with my work often acted as secondary collectors, noting interesting phrases they heard or used. In addition, this strategy widened the range of broadcast media I was able to pull from. Even movies can be a source of terms!

4. LABSTAT Crosswalk Structure and Organization

4.1 Table Structure

The LABSTAT Crosswalk (LSC) is structured as a collection of 4-column tables.

Column 1 contains general language words and phrases gathered from the sources discussed earlier.

Column 2 will list the corresponding BLS terms. There may be more than one corresponding term, perhaps used in different program areas or publications. The BLS term may or may not be included in Column 1. In some cases, there will be no corresponding term. This may occur because the word or phrase is outside the scope of BLS information. In some cases, the corresponding BLS term may be more general than the Column 1 word or phrase. This is likely to occur, for example, in occupation titles, where a specific title is subsumed under a more general category. Column 2 will be filled in by BLS experts; an important part of this work will be determining when a word or phrase is outside BLS scope, and therefore what action to recommend.

Column 3 describes the semantic concept represented by a group of Column 1 words and phrases, from the general language perspective. For example, in the *Diversity* table, there are relatively neutral words such as *diversity*, and there words with more negative connotations, such as *discrimination*. The Column 3 entries are *general diversity* and *general discrimination*. ("General" as opposed to diversity

or discrimination along any one dimension, such as race or ethnicity.) In most cases, a semantic concept label applies to subsequent rows until a new label is listed.

Column 4 will list specific surveys, tables, and other data resources in which the words, phrases, or concepts are applicable. For example, the *Geographic Area* table lists words for national, state, county, and zip code (among other types of regions). BLS information is not always available at all these levels. This column will be filled in by BLS experts. This column could also be used to direct users to other agencies or sources of information, if known. For example, detailed questions about population statistics are often more appropriately answered using information from the Bureau of Census.

The LCS currently consists of thirty-five 4-column tables, containing an approximate total of 4,400 general words and phrases. (This total includes words and phrases that may occur in more than one table.) Each table is in its own Microsoft Word file. The Appendix lists information for each table:

Table name
Approximate number of words or phrases
Brief descriptions of contents
Notes about the contents, taken from the actual table files.

I have not included the actual tables in this report, for reasons of space. Those who wish to obtain them should contact Fred Conrad (Conrad_F@BLS.gov) or me (stephanie@ils.unc.edu).

The nature of the relationships among words and phrases within each cell varies in part according to the degree of synonymy in the words, and the number of entries. In some tables, entries in a single cell are synonyms. In other tables, most notably the occupation and industry tables, the cells contain much larger groupings, for example, all occupations in the medical specialties. There are several reasons for this. (I've used occupation titles for the example here, but the same reasoning applies to industries.)

- It is clearly infeasible to hope to list all of the possible occupation titles (or even 80% of them) that people use and could look for.
- Occupation titles change frequently, according to duties, changes in organizations' structures, fashion, etc.
- BLS information is generally not given at the level of specific job titles; broader occupation groupings are used.
- Other institutions with which users may be more familiar organize occupation titles in other ways. Examples include telephone directories, newspaper want ads, Yahoo-type directories, or college departments and job fairs. It might be helpful to look at these as a way of structuring the general language side of the crosswalk.

When a user calls or emails for help finding information about a specific job title, the consultant usually gives the closest equivalent, if that seems unambiguous, or steers the user to the occupational classification to allow him or her to choose the best category. This model might be more effective than trying to create an exhaustive list.

The rows within the tables are not ordered in any particular way, although I've usually tried to put general words first, and more specific ones later. For example, in the *Benefits* table, I've put words such as *benefits* and *fringes* first, then types of benefits (*health insurance*, *retirement plan*) later.

4.2 Organization of Collected Words and Phrases

4.2.1 Ambiguity

There are at least three types of ambiguous words or phrases in the LSC:

1. Some words or phrases may be ambiguous in general language, but when they are associated with the information that the BLS provides, the other meanings drop out. *Labor*, for example, may refer to a stage in the birth of a baby, or to work and employment. Since the BLS only deals with the second meaning, the first is clearly out of scope of the BLS.
2. Some words may or may not be ambiguous in general language, and are still ambiguous in reference to BLS information. Some form of additional clarification may be necessary in order to provide appropriate information. The infamous *pay* terms (Haas, 1999b; Haas & Hert, 2000) are an example of this.
3. Some words or phrases may or may not be ambiguous in general language, but are ambiguous in the sense that they could refer to BLS data or data that other agencies provide, especially from the perspective of the non-expert user. Again, additional clarification may be necessary in order to steer the person in the right direction.

4.2.2 Synonymy

A simple definition of a crosswalk is that it identifies words in two different languages (general language and BLS terminology, in this case) that mean the same thing. If an end user enters a word or phrase that we can identify as a synonym of a BLS term, it is easy to find the BLS information associated with that term. For example, if *medical care* is found as part of the scanning process, and *health care* is a BLS term, it is fairly straightforward to link the two as meaning the same thing. At this point, one would have a small group of two terms that are names for the same concept. Both are found in general language, but one is "privileged" as being the BLS term for a component of the Consumer Price Index (CPI). We could add to the group by finding other phrases that mean the same thing. The crosswalk would then consist of several words and phrases from general language that are mapped to a single BLS term.

Of course, the situation is not that simple. Although there are indeed many instances of fairly exact synonyms, there are also many examples of "near-synonyms" or "quasi-synonyms". A near-synonym is a word or phrase that means almost the same as another. The difference may be one of tone or register (*chick* and *girl*), one may be slightly broader or narrower in meaning (*nurse* and *RN* or *elderly* and *older American*) or the degree of synonymy may be dependent on context or usage (*retired* and *old*, or *unemployed* and *not in the labor force*). In addition, there are differences in dialect, local usage, age, social standing, native or non-native language, and so on that can complicate judgments of similarity. In these cases, identifying useful groupings may be less

straightforward than in the first case. Obviously, there is some level of agreement on when words mean the same thing – this is information that most dictionaries and general thesauri provide. Even there, however, the words listed as synonyms need to be interpreted for "flavor" or nuances of meaning. For example, *Roget's Thesaurus* (4th Edition, 1977, Harper & Row) lists these words in the same cluster as *unemployed*:

> idle, fallow, otiose, unemployed, unoccupied, disengaged, jobless, out of work, out of employ, out of a job, out of harness, free, available, at liberty, at leisure, at loose ends, unemployable, lumpen, leisure, leisured, off duty, off work, off. (Concept 708.17)

Obviously, some of these are closer to the technical meaning of *unemployed* than others.

4.2.3 Conceptual Grouping

Synonymy is only one organizational scheme that can be applied to the collected words within the framework of a crosswalk. Another scheme clusters possible values for the same variable or parameter together. For example, synonymy would put all the words for *female* in one group (*woman*, *girl*, *women*, etc.) and all the words for *male* in another. But we could then associate these two groups together as representing two possible values for the variable or concept of *gender*. This organization is especially clear-cut in demographic categories (e.g., *race*, *gender*, *education*), but can also apply to somewhat less well-formed (or well-agreed upon) categories such as *industry* or *occupation*. The key here is that ultimately, the clusters must be tied to BLS data organization, which provides a handy grounding point for the crosswalk. There is, however, a crucial disadvantage to depending solely on BLS information organization; it is one of the things that people find confusing! Clearly there must be some kind of anchoring in BLS data organization in the crosswalk at some point; I would argue, however, that it should not necessarily be used as the top level or sole organization scheme of the collected words and phrases. In other words, the crosswalk can mediate not only between general language words and BLS terminology, but between general language organization or clusters and BLS organization. We are still left with the problem of exactly how to determine "the right clusters".

For this project, I started by identifying the "easy" ones. This include many of the demographic classes, and ones that intuitively have broad agreement in general language that corresponds to BLS concepts. Clusters in this group include

- Age
- Disabled
- Education
- Ethnicity
- Race
- Gender
- Marital status
- Geographical area
- Job categories
- Industry

For some of these, such as *gender*, the internal structure is clear – there are two basic meanings to which to map words. Others, notably *job categories* and *type of industry* are

generally identifiable as having a "crisp" meaning, especially in the BLS context. However, membership in the cluster, as well as its internal structure (i.e., which words are synonyms or near-synonyms) is harder. *Race* and *Ethnicity* are an interesting pair. The BLS has clear definitions for each, with very few possible values. The general public, and indeed, other government agencies, recognize more. In addition, there is frequent user confusion over where some words fit. For example, where does *Indian* belong? From the BLS perspective, it does not matter - no data is collected regardless of its meaning. An end user may find this more confusing.

For other clusters, I tried to err on the side of inclusion, rather than trying to identify smaller groupings. In fact, calling these *clusters* may even be misleading - *general topics* or *descriptive groups* may be more accurate at this point. A good example of this is the *work arrangements* table. It includes information on working hours, shifts, locations (e.g., traveling, home, office, factory floor), job requirements and skills, and so on. This table is evidence of the many things people can say about how they work, in addition to what their occupation is! Many of the words or phrases in this cluster could be interpreted (that is, mapped to different BLS terms) in more than one way, often depending on context, or other words in the request. Many of these words or phrases may not even have equivalent BLS terms, or map to information the BLS collects. As BLS experts provide the mappings to terms and data, they may wish to further divide these clusters, or filter out those that have no mappings so they can be handled differently.

If less ad hoc, or finer-grained classification is desired, there are a couple of possible approaches that could be applied iteratively until term groupings of the desired size and semantic scope were reached.

1. BLS experts could sort the words into groupings that made sense to them according to their expert knowledge of the organization and meaning of BLS information. They could give judgments of "goodness of membership" to identify core and outlying members. They could also attach names or labels to the groups; these names could represent the mapping to BLS terms where it exists, or other technical or common terms where it does not. The difficulty with this approach is that the structure would then reflect the existing BLS structure, which was addressed above.

2. An alternative would be to give members of the general public the same task. We could assume that the resulting clusters would reflect a general non-expert view of the similarities and differences among the words.

3. A third approach would be to use a general thesaurus (e.g., *Roget's*) to identify clusters. One drawback to this approach is the problem of coverage, especially of slang or new words. Another problem is that it would still require interpretation of which clusters were best, especially for ambiguous words.

4. For some clusters, finally, there may be other existing organizational structures that would both reflect common ideas about similarities of the words within the cluster, and serve to produce smaller, more precise categories. For example, industry types and occupation categories could be organized according to telephone directory or Yahoo-like directory structures. These are familiar to many users (especially those on the web), and also offer more than one layer of specificity.

There are two other characteristics of words that complicate the development of obvious clusters. One is that some words are truly ambiguous, and can therefore fall into more than one category. I have tried to place these words into all the appropriate categories, although I expect that I have overlooked some possible meanings.

The more interesting characteristic is that some words combine meanings from more than one cluster. *Actress*, for example, incorporates gender and occupation title. Other words hint at this possibility, often because of social custom and usage. For example, *student* indicates that an individual is attending school. Because people are usually students when they are younger, however, there is a frequent assumption that a student is young. Similarly, *retired* implies that someone was working and now is not, (generally permanently rather than just being between jobs), but there is also a common assumption that a retired person is older - over 65. Unless we have the opportunity to ask the user using these words, we cannot tell if this is deliberate, or merely imprecise usage.

5. Discussion and Recommendations

This section discusses specific term tables and clusters, characteristics of the words and phrases that were found, and implications for continued development and deployment of LSC. In addition, it suggests the next stage of analysis; developing a set of templates to represent common phrase structures.

5.1 Uses of the LSC

Fundamentally, the purpose of the LSC is to guide non-expert users from their starting point, a question phrased in general language based on whatever level of understanding of BLS concepts they have, to the BLS information source that provides the best answer to their question. The LSC structure supports this purpose by providing a framework in which general language words and phrases are mapped to BLS terms, and secondarily, to the information sources in which that term is applicable. But the LSC itself should be seen more as the "backend" for a variety of user aids, not as an aid itself. Following are some thoughts and recommendations concerning some kinds of tools that LSC could support.

Alphabetical list of words. A comprehensive list of the collected words and phrases in alphabetical order. The BLS "translation" or mapping would be paired with those words having a corresponding term. If there was none, the column would be blank, thus directly

or indirectly informing users that either the information is not available at the BLS, or that another word must be used. The major advantage of this tool is that it would be relatively easy to construct. One disadvantage is that it could quickly lead to frustration, as users tried to find words that either had no entries, or that had no mapped term. Another is that it has no inherent semantic or conceptual structure. This lack discourages browsing as a means of finding pertinent words. It also does not help the users learn more about the BLS information organization, which would help improve success on subsequent searches.

User-oriented categories. One version of this structure would have a small set of fairly high-level categories that reflected how non-experts think about BLS-related information. Users could then drill down to find more specific groupings within the category, and then browse a short list of related words or terms. Candidate top-level categories might be "Employment and Unemployment", "Demographics", or "Wages and Salaries". An important advantage of this structure is that it starts by reflecting a more general perspective on BLS information, starting with information that non-experts are likely to have and understand. As users start to drill down, the smaller groupings could then reflect ties to the BLS terminology structure and data products. The difficulty with this approach is in determining just what "user oriented categories" are. Earlier in this report I suggested that telephone directories or other common directories might serve as models for occupation and industries, but this notion would need investigation to determine whether such directories would be comprehensive enough. Another approach would be to have non-expert users evaluate the categories to see if they could understand what the categories meant, and predict what words or phrases might be found in each. This experiment could be expanded by including experts in the participant pool. This would show where experts' and non-expert's perspectives of the conceptual space overlapped, that is, what categories they recognized in common, and where they differed.

Links to existing classifications. Some categories, especially those that cannot be exhaustively enumerated such as occupation titles or industry types, are associated with existing classification systems. Providing a direct link to those would allow the user to select the term that best matches their information need. A disadvantage is that some additional definitions or annotations might be needed to help the user decide; the class names themselves might not be sufficient.

Links to examples. The general language words and phrases in the LSC can be used in giving examples of BLS terms and categories. These could be linked from the LABSTAT pages themselves, as the user is asked to make each decision in constructing a series. Many of the current definitions and choices would be enriched by examples that illustrate general language concepts as well as more technical ideas and instances.

Elaborating BLS data products. This suggestion is an extension of the previous one, and would involve using general words and phrases drawn from the LSC to give examples of the information found in a particular series or table as a whole, rather than one variable at a time.

5.2 Observations on Words and Phrases

5.2.1 Word Variants

Individual words occur in many regular forms, as well as a few not-so-regular ones. Variations observed in the LSC tables include:

- Singular/plural. Regular pluralization in English consists of adding *s* or *es*, but other forms are common, such as *criterion – criteria* or *goose – geese*.
- Verb conjugations. These include markings for number and tense.
- Nominalizations. Often marked by the *–ation* suffix, as in *union – unionize – unionization*.
- Alternate spellings. The most common set of variations I saw were British spellings, e.g., *labor – labour*.
- Hyphenation. This is can especially be seen as new words become more common, as in *e-commerce - ecommerce*.

For the most part, these sorts of variations have no effect on meaning. However if users will be using a search engine of some type, it should be one that supports stemming and the other variants.

5.2.2 Open and closed class words

In linguistics, parts of speech are often divided into "open-class" and "closed-class". Closed classes are those in which all members of the class can be listed, and which gain or lose members slowly, if at all. Examples of closed classes in English are determiners, pronouns, and prepositions. Open classes are those whose members cannot be listed, and where new words appear regularly. In English, these include nouns, verbs, adverbs, and adjectives.

We can use this same idea to look at classes of words in the LSC. Closed classes would be those clusters where most, if not all, the synonyms that people are likely to use can be listed, and where new words are unlikely to appear. Examples would likely include gender, education, family status, marital status, time, and units. (Race and ethnicity could be included here, except there are many hyphenated terms, e.g., *Chinese-American*, that would be difficult to predict. This structure is *productive* in the linguistic sense, meaning that it invites new combinations.) Similarly, some clusters are very volatile indeed, and could be considered open classes, both because of the changes in work and employment themselves, and because the words that people use for them change over time. Examples of open classes might be work arrangements, benefits, descriptions of industries, or legal concepts. It is interesting to note that in the words collected for the LSC, these clusters seem to contain many words that refer to concepts outside the scope of BLS information.

We could also define a third kind, enumerated classes. Classes of this sort would be those that may not be listed exhaustively, but which can be linked to reference sources that contain finite lists of members. Examples would be geographical area, occupation titles, and industry types. Like the open classes, end users can come up an almost infinite number of words or names in the class. These names are more likely to be in the scope of the BLS, even if they are at too fine a level of granularity. For example, there is an

almost infinite list of occupation titles. In theory, however, they can all be placed within the existing classification of occupations, at least at some level. Therefore, these clusters are similar to the closed classes in that there exists an exhaustive list of authoritative terms.

This notion of open, closed, and enumerated classes is of more than just theoretical interest. It also affects maintenance of the LSC - how new words should be added to the collection. The closed class clusters will need little in the way of additions. (There are only two sexes, after all, and these are unlikely to change!) Race and ethnicity perhaps deserve some mention here. In the world at large, these could be considered closed class, except for the kinds of productive formations noted above I am reluctant to call these enumerated classes for this reason. Membership in these clusters is quite limited from the BLS perspective; however this could change in the future.

Update for the other classes is more problematic. For any linguistically productive class, the speed at which new words appear would far outpace a reasonable level of effort devoted to keeping up. Establishing a well-organized structure that will support non-experts' browsing is a way of working around this problem. If a user has a particular word or phrase in mind that cannot be found by direct search, her or she could still find a category into which it could fall, and browse that to find something that is close in meaning. This does not completely eliminate the task of adding new words, however. "Trendy" words, which come and go quickly, should not be added because they are so ephemeral. *Webmistress* is a good example of this. It was used for a year or so, and then was subsumed by *webmaster*. Other words, however, are more permanent additions to the vocabulary. *Webmaster* has been around for a while, and seems to have a well-established meaning; for example, it is commonly used in job advertisements. So there needs to be some form of environmental monitoring to identify words that have been around for a while.

5.2.3 Templates

The LSC contains mostly single words, or phrases that name a single concept (e.g., *airline industry*). But people frequently combine words to express a more complex concept, or more precisely, a combination of concepts. Indeed, some of the clusters in the LSC rarely occur unmodified, such as *Time* or *Data Product*. Consider these examples of descriptions of groups of people.

black middle-aged women
white teenage boys
65-year old African American males

These all consist of the same three concepts: race, age, and gender. We could then think of a pattern, or template for describing people as

GROUP OF PEOPLE = RACE + AGE + GENDER

This starts to look remarkably like a grammar rule. An ordinary linguistic grammar consists of rules identifying combinations of parts of speech, such as

noun phrase = determiner + noun.

A semantic grammar uses the names of semantic categories, such as the LSC clusters, in place of parts of speech. Semantic grammars have been used to describe language in limited domains from airplane flights to medical lab tests (Hendrix, 1978; Sager, 1981). The general language in the LCS is much broader, of course, but it maps to the restricted "language" that is based on the tables and other data products. In this sense, a "well-formed phrase" would be one that could reasonably describe a way of breaking out the data. I have collected phrases from the LSC and from the original sources for further analysis.

I see at least three uses for a collection of templates modeled after a semantic grammar.

1. Templates could identify common concept combinations with LABSTAT, e.g., in choosing values for variables. Associating the sequence of choices with a template and example phrase could clarify what information can (and cannot be) combined.
2. Templates could serve as examples of the tables that are available. In this sense, they would serve as "abstracts" of what could be found there.
3. Templates could also help identify questions that call for information from more than one agency. Such phrases would reflect ideas that the general public has, but which do not match agency boundaries. Knowing more about what information commonly want to combine could help efforts such as MAPSTATS and other "hot reports" that combine agencies' information.

It is possible that some of the clusters would need to be broken into smaller groupings to be useful at the template level. Preliminary investigation into the demographic clusters look promising, however. I hope to explore this idea further in next year's work.

5.3 Summary

The LSC as it exists now is ready for additions and refinement by BLS experts.

- Some of the larger, less well-formed clusters (e.g., *Work Arrangements*) could be subdivided.
- BLS term mappings need to be entered into Column 2.
- BLS publications, tables, series, and other data products need to be entered into Column 4.
- For words and phrases that could lead to information outside the scope of BLS information, links or pointers could be provided to help users find the appropriate source. This is not feasible for all words and phrases, but certainly could be done for those that BLS consultants see most often.

Some consideration should be given to policies and procedures for maintenance, both to reflect changes in BLS information, and to enlarge the collection of general words and phrases.

Designs for the tools that will be based on the LSC can be developed. This includes

- Determining the functions of the tools, and their priority.
- Looking at the best way for end users to "enter" the crosswalk, for example, via semantic categories that are designed for browsing. This could entail investigating existing organizational structures that are familiar in everyday life. Any browsing structure would then need to be tested by a sample of non-expert users.

The idea of templates, or a semantic grammar for BLS information based on the LSC should be explored.

6. Bibliography

Cabré, M. (1999). *Terminology: Theory, Methods and Applications.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Haas, S. W. (1999a). *"What Everyone Calls It": Mapping Users' Words and Terminologies*. Open Forum on Metadata Registries, February 16-19, 1999, Washington, D. C. Available at http://ils.unc.edu/~stephani/bls/metadatareg-99.ppt.

Haas, S. W. (1999b). *Knowledge Representation, Concepts, and Terminology: Toward a Metadata Registry for the Bureau of Labor Statistics.* Final report to the Federal Bureau of Labor Statistics. Available at http://ils.unc.edu/~stephani/bls/fin-rept-99.pdf.

Haas, S. W. (2000a) *Terminology for Statistics: How Can End Users Connect?* Open Forum on Metadata Registries, January 17-21, 2000, Santa Fe, NM. Available at http://ils.unc.edu/~stephani/bls/metadatareg-00.ppt.

Haas, S. W. (2000b). The role of knowledge representation in managing statistical information. *Proceedings of International Conference of Establishment Surveys II* (to appear).

Haas, S. W. & Hert, C. A. (2000). Terminology development and organization in multi-community environments: The case of statistical information. Proceedings of the 11th ASIS SIG/CR Classification Workshop (to appear).

Hendrix, G., Sacerdoti, E., Sagalowicz, D., & Slocum, J. (1978). Developing a natural language interface to complex data. *ACM Transactions on Database Systems*, 3, 2, 105-147.

Hert, C. A. (1999). *Federal Statistical Website Users and Their Tasks: Investigations of Avenues to Facilitate Access: Final Report to the United States Bureau of Labor Statistics*. Available at: http://istweb.syr.edu/~hert/BLSphase3.pdf.

Pearson, J. (1998). *Terms in Context*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Riggs, F. (1993). Social science terminology: Basic problems and proposed solutions. In H. Sonneveld & K. Loening, (eds.) *Terminology: Applications in Interdisciplinary Communication*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 195-222.

Sager, N. (1981). *Natural Language Information Processing: A Computer Grammar of English and its Applications*. Reading, MA: Addison-Wesley.

7. Appendix: Information about Word Cluster Tables in LSC

Absent from Work

Approximate number of words and phrases: 25
Brief description: words concerning reasons for absence, how lost time is counted.
Notes:
<none>

Age

Approximate number of words and phrases: 113
Brief description: words describing age or age ranges of individuals.
Notes:
1. There is a series of age terms that identifies ages or age ranges, e.g., *teenager*. An individual passes from one to another of these ranges during his/her lifetime. A second type of age terms identifies a specific cohort of people based on the years in which they were born, e.g., *baby boomers* or *gen-x*. An individual stays in this cohort during his/her lifetime.
2. Most age terms other than specific age or age range are imprecise as to who would or wouldn't be included.
3. I've included some terms here that have a strong implication of age, but don't necessarily march with age.

- *Student* and related terms. A person could be a student at any time of life, but the stereotype is of a younger person, high school or college.
- *Retirement* and related terms. A person could retire from a job or profession at just about any age, but the stereotype is of an older person. 55-60 may be considered early for retirement. A person who has never worked doesn't retire, but for some queries using the term *retired*, older people should be included. For example, *retired couple*, both members of the couple would be included, even if only one of them had worked. Finally, even if someone has retired from his/her life's major profession, he/she may still be working (e.g., part-time).

Benefits

Approximate number of words and phrases: 117
Brief description: words describing types of worker benefits and non-monetary compensation.
Notes:
1. In many sections, finer or coarser distinctions could be made, according to the disambiguating power of BLS data resources. See *retirement benefits*, with the various flavors of pensions, for an example.
2. The bottom of the tables lists "perceived benefits-related" terms. These include phrases such as *workers compensation* and *FICA*, which aren't employer-provided benefits in the same way as *health insurance* (for example), but which may be perceived as being in roughly the same class.

Consumer costs

Approximate number of words and phrases: 166

Brief description: words describing consumer costs, buying power, Consumer Price Index (CPI), and its components.

Notes:

1. *lodging* is ambiguous between housing and places to stay while traveling. I've put it in both places.
2. I've combined travel, business travel, and commuting.
3. The last section, class = ???. Unclear if this belongs here or in producer costs. These phrases could be used in either context.
4. This cluster includes CPI and cost of living terms. These are *not* repeated in the economy cluster.

Data Products

Approximate number of words and phrases: 70

Brief description: words describing publications, surveys, and various kinds of data, statistics, and "numbers".

Notes:

1. This cluster is a little different from others. It is possible that the general data product terms can be appended to any other cluster (requires template investigation). Some of the examples are very specific, such as *SIC Manual*. Others are extremely vague or ambiguous, such as *workplace statistics*.

Disability

Approximate number of words and phrases: 36

Brief description: words describing types of disability, disabled employment, workplace accommodations.

Notes:

1. BLS has limited information on disability and disabled workers. It is somewhat related to work-related injuries, also the ADA is work-related. Many of these words and phrases may therefore be out of scope, but perhaps referable.

Diversity

Approximate number of words and phrases: 74

Brief description: words describing diversity and discrimination issues (age, gender, race, ethnicity)

Notes:

<none>

Economy

Approximate number of words and phrases: 116

Brief description: words describing types of economic measures other than CPI and PPI.

Notes:

1. CPI and cost of living related terms are in the consumer costs table. PPI, producer profits, and productivity are in the producer costs table. This cluster is

various miscellaneous measures of the economy. (Some of the grouping may be a little arbitrary here.)
2. Terms in the miscellaneous index group are sometimes ambiguous or vague. Some may be components of CPI or PPI.

Education

Approximate number of words and phrases: 110
Brief description: words describing degrees, level of education, and employer-related training.
Notes:
1. In general, *education* refers to secondary, college, graduate, and technical school education. *Training* often refers to on-the-job training, or other targeted, often employer-related training. But this isn't an entirely uniform division in use. I have included both here.

Employment

Approximate number of words and phrases: 208
Brief description: words describing measures of employment, job seeking, hiring.
Notes:
1. There is some overlap between employment and labor force, and there is a possibility for confusion by naïve users.
2. There are lots of demographic modifications of these terms.
3. There is also some overlap with the work-arrangements cluster.

Ethnicity

Approximate number of words and phrases: 56
Brief description: words describing ethnicity. See notes.
Notes:
1. This cluster contains terms that are not ethnicity according to BLS definitions, but which may be considered so by the non-expert. Further, many of these are not tracked by the BLS. These may be used as examples of information that is *not* available for end users.

Family status

Approximate number of words and phrases: 43
Brief description: words describing parents, size of family, head of household.
Notes:
1. This category contains phrases referring to family members, parenthood, and who in a family is employed. I've also put phrases referring to *head of household* status here.

Gender

Approximate number of words and phrases: 83
Brief description: words describing gender.
Notes:

1. Gender terms, like ethnicity and race, almost always occur in combination with other semantic clusters. Sometimes the combinations are intrinsic to the word, such as *Latina*, combining ethnicity and gender. This also occurs with occupation, as in *waitress* or *actress*.

Geographical area

Approximate number of words and phrases: 77
Brief description: words describing various kinds of geographical and political divisions.
Notes:
1. I haven't included exhaustive examples of categories such as the states or all metropolitan regions. I've included some such as *New York*, that could be ambiguous between state and city, or *Washington D.C.*, which people may consider either a city or a state. Similarly with *Puerto Rico*.
2. *Ghetto, ghetto dweller,* and *inner city* have common connotations of income level/poverty, as well as identifying an urban setting.
3. In the BLS context, do *national* and *American* always refer to the same regions or data sets?
4. Similarly, in the BLS context, do *foreign*, *international* and *non-US* mean the same thing?

Hours

Approximate number of words and phrases: 52
Brief description: words describing when and how much people work, including shifts.
Notes:
<none>

Income

Approximate number of words and phrases: 50
Brief description: words describing socio-economic class.
Notes:
1. *Income* here refers to general income categories or levels such as *middle-class* or *poverty*, regardless of the source of income. Wages, salaries, and other sources of income are in other clusters.

Industry description

Approximate number of words and phrases: 182
Brief description: words that can describe industries, including e-commerce.
Notes:
1. This cluster covers ownership, sector, size, and similar concepts.
2. I've included more notational variants in the e-commerce group than usual, because spelling and punctuation is still in fluctuation.

Injury

Approximate number of words and phrases: 236

Brief description: words describing types of injuries, causes, and risk evaluation and prevention.
Notes:
1.I did not attempt to find an exhaustive list of injury types. The first 2 groups are those that I happened to find.
2. I found many terms related to stress, so I separated them out. Otherwise, types are in no particular order.
3. I have not strictly followed the BLS classification of injury, event, source.

Job categories

Approximate number of words and phrases: 603
Brief description: words describing occupation titles as well as broader classifications.
Notes:
1. A large number of these terms identify fairly specific job titles. Naturally, this is just a sample, and cannot in any way be considered exhaustive. Two things are of interest, however. One is the range of titles for essentially the same type of job, often some formal, some more casual (e.g., *bureaucrat* and *desk jockey*) The other is the frequent ambiguity between industry and occupation; this is probably a distinction that end users often can't make clearly. This should be remembered as one considers linking occupation titles to the Occupational Handbook and industries to NAICS codes – sometimes that may lead people in the wrong direction.
2. This collection starts with some general descriptions of occupation types. The second major grouping is slightly more specific kinds of jobs, e.g., *manufacturing jobs*. The third group contains terms related to job titles and descriptions. The last group contains more specific job titles, very loosely grouped by profession.
3. What is an *invisible career*?
4. Many of the specific job titles could be categorized in a couple of ways. For example, *postal police* could be in among the postal employees or in among the other police officers. Similarly, *college administrators* could be with other academic jobs, or with other management jobs.

Job tenure

Approximate number of words and phrases: 45
Brief description: words describing retention, turnover, time in job.
Notes:
1. Some terms refer employee retention, others to job retention, reflecting two different perspectives. From a BLS data-availability standpoint, is this an important difference?

Labor force

Approximate number of words and phrases: 99
Brief description: words describing labor force, pool of available workers.
Notes:
1. The first grouping may need refinement, please check.

2. I've included *retirement* terms here. There is an implicit assumption that if someone is retired, then he/she was working (and therefore part of the labor force) at some time. Retirement is also often conflated with age.

Legal

Approximate number of words and phrases: 64
Brief description: words naming various laws and related actions.
Notes:
1. This cluster has various laws, acts, and actions related to labor and somewhat related laws.
2. The last row is a list of miscellaneous acts.

Marital status

Approximate number of words and phrases: 30
Brief description: words describing marital status of individuals.
Notes:
<none>

Other demographic

Approximate number of words and phrases: 35
Brief description: miscellaneous demographic words, not enough in any group to justify their own table.
Notes:
1. Oh well, at this point I can't do without this pseudo-cluster. Some of these may move into other clusters,
either as I gain a better understanding, or as the experts look at where they fit with the BLS data. I've made some suggestions where I could.
2. For *foreign-born* and the related terms, the contexts suggest that these refer to individuals in the U.S., so these are not geographical area terms.

Population

Approximate number of words and phrases: 47
Brief description: words describing ways of measuring or organizing population.
Notes:
1. This mostly uses words such as *demographics* or *population*, although some notions of comparison to population are intrinsic in the phrase, such as *per capita*. These terms almost always occur in combination with other clusters.
2. These are frequently associated with census information

Producer costs

Approximate number of words and phrases: 102
Brief description: words describing a variety of costs (and some profits) of industries.
Notes:
1. This cluster includes some terms for establishment income, profits and assets, costs, productivity, and PPI.

Race

Approximate number of words and phrases: 64
Brief description: words describing race.
Notes:
1. As with ethnicity, there are some terms here that don't follow BLS definition of race, or go beyond the distinctions that BLS collects data on, but which may be confused by end users.
2. Many racial terms are closely related to diversity terms.
3. I've include *ghetto dwellers* and *ghetto youths* here, because there is a strong association of race in those terms, although it is not absolute.

Statistics

Approximate number of words and phrases: 40
Brief description: words describing various kinds of statistical data and data collections.
Notes:
1. These are various terms that can be used to describe a statistical product or used to request information. (I didn't include *information* in this set). Grouping is "naive".
2. The *specific results* group refers to general numbers produced by some statistical function, such as *average* or *rate*. The *general results* group refers to information without specifying its exact form. For example, an *amount* can be used to refer to a count, a total, a ratio, etc. BLS experts REALLY need to look at this.

Time

Approximate number of words and phrases: 27
Brief description: words describing time spans and periodicity.
Notes:
1. The *chronology* group is sometimes used to express the kind of data product the user wants.
2. The time frequencies almost always occur only as a modifier of information – an example is included
3. The vague time-spans don't really have a definite length or period – probably require clarification.

Type of industry

Approximate number of words and phrases: 721
Brief description: words naming types of industries.
Notes:
1. This cluster shares many of the same characteristics as the job category cluster. There is often the potential for confusion between occupation and industry.
2. This was a difficult cluster to group for several reasons. Almost any noun that names an object could be considered as the name of an industry. Further, many objects have a manufacturing phrase and a retail phase, which would fall into

different NAICS categories. I took a different approach, trying to think about "where it is used". Another way of thinking of these is in "Yahoo-like" categories, based on the object or service, rather than its stage of production. This would be something to consider in presenting end-users with an alternate way of finding industry information. Or a telephone directory type organization.
3. Warning: these are VERY broad categories, and certainly non-exhaustive.
4. The first groupings label major sectors, and also some "extra-sector" pseudo industries, such as *Peace Corps* and *military*. For most of these (e.g., *manufacturing*), I have many phrases which illustrate modification.

Unemployment

Approximate number of words and phrases: 80
Brief description: words describing unemployment and job loss.
Notes:
1. I've put *federal unemployment* under industry modification, which would assume that someone wanted unemployment among federal workers. However, this could also be interpreted as *national unemployment*.

Unions

Approximate number of words and phrases: 65
Brief description: words describing union membership and activity.
Notes:
1. I have included just a couple of examples of types of unions.

Units

Approximate number of words and phrases: 26
Brief description: words describing how data is gathered and grouped at the point of collection and/or reporting.
Notes:
1. The idea behind this cluster is to represent the units of survey on which data is "distributed". This includes individuals, households, and businesses, therefore there is some overlap with family status and types of industries. This cluster does not include any geographical area terms, although they can also be thought of as units in some contexts.
2. I've included *employer* in the individual and the business grouping; it could be used either way.

Wage, salary & income

Approximate number of words and phrases: 337
Brief description: words describing monetary compensation and other sources of income.
Notes:
1. I thought about separating this into to clusters, wage & salary, and income, using the definition that income could come from sources other than employment. On the other hand, it is clear that end users don't have a clear understanding of the

difference. Notice, for example, the group of terms *job category + income*, indicating that they may be thinking of income as coming from a job.
Given this decision, I then thought about the organization within the table. I saw two obvious possibilities. I could treat wage, salary, and income as one concept, and group by modification (occupation, demographic, etc.) Or, I could separate these three base terms, and within these three major groupings include the modifications. There are advantages/disadvantages to each. (If we weren't so tied to 2 dimensions, there would be ways to capture all of these groupings!) But I finally decided to use the latter strategy.
2. With all the job categories + wage salary, the entries should just be considered examples, not exhaustive.
3. This list should also be merged with the outcome of Carol Hert's and mine experiments last year – I'll get to it.

Work arrangements

Approximate number of words and phrases: 214
Brief description: words describing how, where, under what conditions people work.
Notes:
1. This cluster contains a variety of terms that describe working conditions. Some of these, such as *shift work* also appear in other clusters. The terms here capture both tangible and intangible aspects of work, from *drug testing* to *employee morale*. It's likely you'll want to reclassify/reorganize these.