Evaluating extracted phrases and extending thesauri

Gordon W. Paynter, Sally Jo Cunningham, and Ian H. Witten

Department of Computer Science, University of Waikato, New Zealand {gwp, sallyjo, ihw}@cs.waikato.ac.nz

ABSTRACT

We describe an interface that uses the phrases occurring in a document collection as a basis for browsing the collection and accessing its contents. Phrases are automatically extracted from the document text to represent the subject matter of the collection. Clearly, the interface's utility depends on how good these phrases are. We evaluate the system by comparing the phrases extracted from a large Web site to those in a thesaurus used by the organization responsible for the site. This analysis serves two purposes: it aids the user by verifying that the phrases extracted are relevant to, and provide good coverage of, the subject areas of the Web site and thesaurus; and it aids the thesaurus compiler by identifying phrases in widespread use that do not appear in the thesaurus.

INTRODUCTION

Browsing is a significant human information seeking activity [1], and encompasses a variety of behaviors. Users of a digital library may wish to browse as a form of search—for example, browsing a list of authors to find that document written by, oh, it's John something, I think his last name starts with an "H" or maybe it's "Wh". Digital libraries frequently provide support for browsing document metadata such as titles or authors (e.g. [10]); however, the utility of current browsing schemes reduces as the size of the collection increases, and the metadata itself grows too large to efficiently scan.

Another form of browsing involves exploring documents grouped by subject matter. This type of browsing is frequently supported in physical libraries—which generally group documents on shelves according to a subject classification scheme—but is less common in digital libraries. Manual classification is expensive and rarely available for documents in digital libraries or Web-based document collections; automated classification is still very much a topic for ongoing research [4]. Similarly, a subject thesaurus can be an invaluable searching and browsing tool for topically exploring a collection, although again documents in digital libraries are rarely tagged with thesaurus metadata.

Another approach to providing a topic-oriented tool for collection browsing is to support exploration of keyphrases extracted from the collection's documents [5,6]. These phrases are best viewed as a supplement to, rather than a substitute for, a subject classification scheme or thesaurus. An alphabetically sorted list of keyphrases lacks structure, and cannot be browsed to learn the underlying structure and semantic organization of a given domain—as can a thesaurus or classification scheme. However, a phrase based browsing tool allows the user to examine the terminology actually present in the collection, and to view individual terms in phrase context. Context is particularly effective for a user in distinguishing between relevant and irrelevant usage of a given term; for example, in distinguishing between bank as a financial institution (e.g. "World Bank") and as a geographic entity (e.g. "river bank").

The phrase-based browsing interface described in this paper is applicable to any collection of HTML documents (and is currently being integrated into an established digital library system [10]), but this paper focuses on a specific interface constructed for the Web site of the United Nations Food and Agriculture Organization (FAO, *www.fao.org*). This Web site is of particular interest in studying phrase extraction and browsing because of its relationship to the AGROVOC thesaurus. AGROVOC is a multilingual thesaurus for agricultural information systems, developed by the FAO to provide a controlled vocabulary for indexing bibliographic records and research projects [2]. The subject matter of the thesaurus and Web site is very similar—the FAO ultimately intend to index their Web site with AGROVOC terms—allowing comparisons between the vocabularies of the thesaurus, the Web site, and the phrase browsing interface. These comparisons form the body of this paper, and can be used to evaluate the browsing interface and extend the thesaurus itself.

The next section of the paper describes the phrase browsing interface. We then examine the quality of the extracted phrases by comparing them, and several other pertinent phrase sets, with the phrases that appear in AGROVOC. Finally we consider the ways in which the extracted phrases can aid the thesaurus editor, and discuss the implications of this analysis and our future work.

THE PHRASE-BASED BROWSING INTERFACE

The phrase-based browser is an interactive interface to a phrase hierarchy that has been extracted automatically from the full text of the Web site. It is designed to resemble a paper-based subject index or

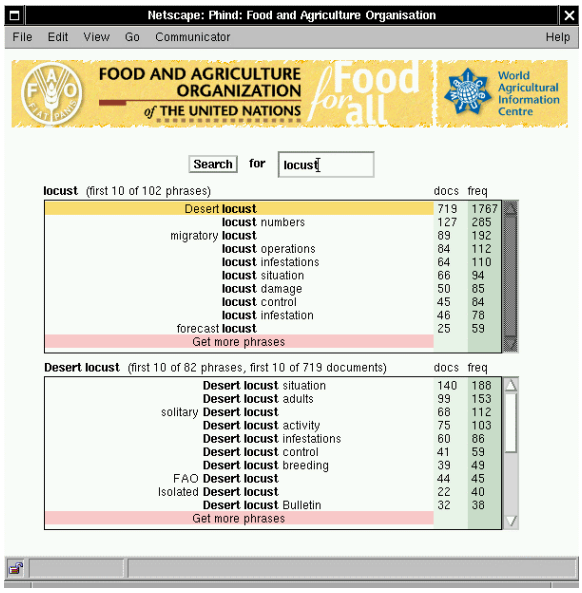

Figure 1: Browsing for information about *locusts*

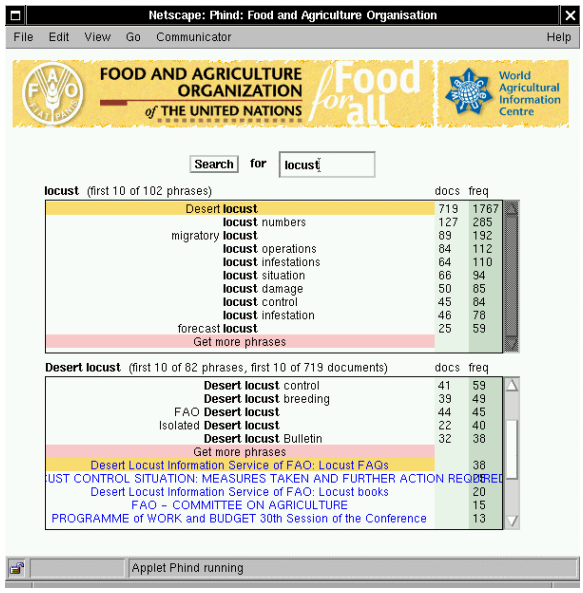

Figure 2: Expanding on *Desert Locust*

thesaurus. Figure 1 shows the interface in use. The user enters an initial word in the search box at the top. On pressing the *Search* button the upper panel appears. This shows the phrases at the top level in the hierarchy that contain the search word—in this case the word *locust*. The list is sorted by phrase frequency; on the right is the number of times the phrase appears, and to the left of that is the number of documents in which the phrase appears.

Only the first ten phrases are shown, because it is impractical with a Web interface to download a large number of phrases, and many of the phrase lists are very large. At the end of the list is an item that reads *Get more phrases* (displayed in a distinctive color); clicking this will download another ten phrases, and so on. The interface accumulates downloaded phrases: a scroll bar appears to the right for use when more than ten phrases are displayed. The number of phrases appears above the list: in this case there are 102 top-level phrases that contain the term *locust*.

So far we have only described the upper of the two panels in Figure 1. The lower one appears as soon as the user clicks one of the phrases in the upper list. In this case the user has clicked *Desert locust* (which is why that line is highlighted in the upper panel), causing the lower panel to display phrases containing the text *Desert locust*.

If one continues to descend through the phrase hierarchy, eventually the leaves will be reached. A leaf corresponds to a phrase that occurs in only one document of the collection (though the phrase may appear several times in that document). In this case, the text above the lower panel shows that the phrase *Desert locust* appears in 82 phrases in 719 documents. The first ten documents are visible when the list is scrolled down, as is shown in Figure 2. In effect, the panel shows a phrase list followed by a document list. Either of these lists may be null (in fact the document list is null in the upper panel). The document list displays the titles of the documents.

It is possible, in both panels of Figures 1 and 2, to click *Get more phrases* to increase the number of phrases that are shown in the list of phrases. It is also possible, in the lower panels, to *Get more documents* (again it is displayed at the end of the list in a distinctive color, but to see it that entry is necessary to scroll the panel down a little more) to increase the number of documents that are shown.

Clicking on a phrase will expand that phrase. The page holds only two panels, and if a phrase in the lower panel is clicked the contents of that panel will move up into the top one to make space for the phrase's expansion. Alternatively, clicking on a document will open that document in a new window. In fact, the user in Figure 2 has clicked on *Desert Locust Information Service of FAO: Locust FAQs*, and this brings up the page shown in Figure 3. As Figure 2 indicates, that document contains 38 occurrences of the phrase *Desert Locust*.

Figure 4 shows another example of the interface in use. In this case, a French user has entered the word *poisson*, exposing a weakness of the phrase extraction algorithm. The FAO site contains documents in French, but our phrase extraction system is tailored for English. The French phrases are displayed are of much lower quality than the English ones in Figures 1 and 2; the list of ten phrases in the upper panel of Figure 4 contains only four useful ones. Phrases like *du poisson* (usually meaning *of fish*) are not useful, and can even obscure more interesting material. However, the system is still usable. Here, the user has expanded *commercialisation du poisson* and, in the lower panel, has clicked on a document titled *INFOPESCA*.

Figure 3: Example Web page

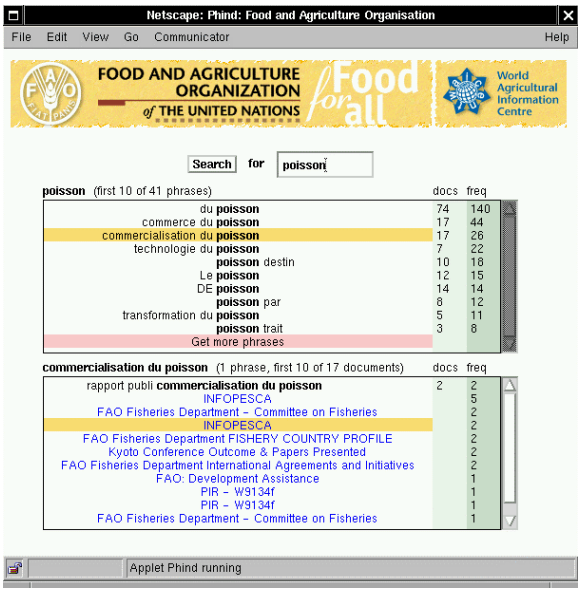

Figure 4: Browsing for information on *poisson*

EVALUATING PHRASES WITH A THESAURUS

The phrases in the browsing interface represent the topics present in the Web site using the terminology employed by the document authors. But how well does this set of phrases match the standard terminology of the discipline? We investigate this question by comparing extracted phrases with those in the AGROVOC agricultural thesaurus. The degree of overlap between the two sets of phrases provides a rough indication of the quality of the extracted phrases as subject descriptors, and the extent to which they cover the range of agricultural topics. Conversely, the applicability of the AGROVOC thesaurus to the FAO site can be assessed by measuring the extent to which the AGROVOC phrases appear in the natural text of the documents.

The *AGROVOC* thesaurus

AGROVOC is a multilingual thesaurus for agricultural information systems, developed by the FAO to support subject control for the AGRIS agricultural bibliographic database and the CARIS database of agricultural research projects [2]. The thesaurus supports the three working languages of the FAO—English, French, and Spanish—and versions in Arabic, German, Italian, and Portuguese are under construction. AGROVOC is actively supported by the FAO and its international community of users, and is periodically updated to reflect changing terminology or shifts in the boundaries of the research field. A searchable version appears at *www.fao.org/AGROVOC*.

The thesaurus is of a significant size—each language version includes more than 15,700 descriptors, and approximately 10,000 non-descriptors (which are synonyms, or otherwise related terms, that are linked to a preferred descriptor that should be used in its place). Thesaurus terms are nouns or noun phrases, and all—including non-descriptors—were selected for inclusion on the basis of their common usage in the agricultural research literature. The AGROVOC vocabulary forms a rich semantic network describing the agricultural domain, and the thesaurus provides links between terms describing hierarchical relationships (*broader term*, *narrower term*), associative relations (*related terms*), and synonym links between descriptors and non-descriptors (*use*, *use for*).

AGROVOC phrases were deliberately designed to be brief (three or fewer words if possible) and compact (at most 35 characters). These limitations were imposed by the original thesaurus software [2]. The strict upper limit on characters has proven problematic, in that lengthy terms (such as the names of organizations, enzymes, chemical compounds, etc.) have had to be abbreviated—sometimes in arbitrary or non-standard ways. This can make querying more difficult for users, who have to guess when and how a phrase has been abbreviated. The overlap between the extracted and AGROVOC phrases is also reduced, though only slightly.

The AGROVOC phrases used in these analyses are taken from the English version only, and include both descriptors and non-descriptors. Despite their name, non-descriptors are useful when searching, since they are meaningful domain terms that searchers might use in a query.

Four phrase sets

We compare the thesaurus to four phrase sets, three based on the FAO Web site and one on an unrelated document collection. A version of the FAO Web site was distributed on CD-ROM in 1998 and contains

	AGROVOC thesaurus		Extracted phrases
1	*forest canopy*	1	forest Academy
2	*forest decline*	2	forest access
3	*forest dieback*	3	forest Act
4	*forest ecology*	4	forest activities
5	forest establishment	5	forest administration
6	*forest fires*	6	forest agencies
7	forest floor vegetation	7	forest agenda
8	*forest grazing*	8	forest animals
9	*forest health*	9	forest area
10	*forest industry*	10	forest assessment
11	*forest inventories*	11	forest authorities
12	*forest land*	12	forest authority
13	forest litter	13	forest base
14	*forest management*	14	forest benefits
15	forest measurement	15	forest biodiversity
16	*forest mensuration*	16	forest biology
17	*forest meteorology*	17	forest biomass
18	*forest nurseries*	18	forest Botany
19	*forest pathology*	19	forest boundaries
20	forest pests	20	*forest canopy*
…	…	…	…
41	*forest workers*	236	forest zoology

1	coppice forest	1	actual forest
2	duff (forest litter)	2	aggregate forest
3	*high forest*	3	Amazon forest
4	*minor forest products**	4	amenity forest
5	mixed forest stands	5	American forest
6	monsoon forest	6	artificial forest
7	nontimber forest products	7	available forest
8	*nonwood forest products**	8	Bangladesh forest
9	*secondary forest products**	9	bavarian forest
10	semliki forest virus	10	Black forest
11	slash (forest litter)	11	boreal forest
12	thorn forest	12	Chimanes forest
		…	…
		206	young forest

Table 1: Phrases beginning with (above) and containing (below) the word *forest*.

21,700 Web pages, as well as around 13,700 associated files (image files, PDFs, etc). This corresponds to a medium-sized collection of approximately 140 million words of text. The Web site has since grown to many times this size, but we have used the 1998 version because it was selected by editors at the FAO, and contains no dynamic content.

The first phrase set, *fao-all*, was formed by extracting every phrase of one to four words from the FAO Web site. This represents an upper bound on performance, as no technique can extract more phrases than appear on the Web site.

The second phrase set, *fao-browser*, consists of every phrase used in the browsing interface. We have experimented with several different ways of creating a phrase hierarchy from a document collection. The phrases used in the browsing interface described here are produced by a combination of rudimentary syntactic processing and sequential grammar induction techniques, described by Paynter *et al.* [9]. The phrases displayed by the interface have a minimum length of two words; for the purposes of this analysis all the words appearing in them are added to *fao-browser* as single-word phrases.

The third phrase set, *fao-keyphrase*, was formed by extracting six phrases from each of the FAO web pages with the KEA keyphrase extraction algorithm [3]. These phrases approximate the keywords that many authors assign to technical documents. This emphasizes precision over recall: because few keyphrases are associated with each document they are more likely to be true indicators of the focus of the document (and therefore closer to the intent of AGROVOC thesaurus entries).

The fourth and final phrase set, *cstr-all*, is intended as a control, and is not based on the FAO Web site at all. Instead, it consists of all the phrases appearing in 500 documents selected randomly from the Computer Science Technical Reports collection of the New Zealand Digital Library. If the AGROVOC thesaurus is indeed well-suited to use with the FAO Web site, we would expect the CSTR phrases to perform poorly in comparison to the first three phrase sets.

Overlap with AGROVOC phrases

We begin our analysis with an example to illustrate the degree and type of overlap found between the sets of phrases. Table 1 shows the phrases beginning with or containing the word *forest* in AGROVOC and at the top level of the browser hierarchy. Italics indicate that the AGROVOC phrase occurs amongst the extracted phrases (and vice versa). Phrases marked with a single asterisk appear in the phrase hierarchy, but not at the top level. Some phrases appear in the plural and only coincide with extracted phrases when they are stemmed, though none appear in the excerpt in Table 1.

The overlap between the AGROVOC thesaurus and the extracted phrase sets is quantified in Tables 2 and 3. The first column of Table 2 (labeled *Unstemmed*) shows that 48% of the words appearing in the AGROVOC thesaurus phrases are also present in the FAO documents (*fao-all*). However, only 19% of the words in the computer science documents (*cstr-all*) occur in the AGROVOC vocabulary. This supports our assumption that AGROVOC is a suitable thesaurus to use with the FAO Web pages. The proportion of AGROVOC words contained in phrases in the browser hierarchy (*fao-browser*) is smaller, but still represents a respectable 30% of the AGROVOC terms. As expected, the KEA keyphrases (*fao-keyphrase*) cover an even smaller proportion of AGROVOC terms.

The proportion of full AGROVOC phrases that are included in the FAO site and the browser hierarchy is high—36% and 22% respectively (Table 3). This is particularly encouraging, as it indicates that a significant number of links exist between

	Unstemmed	Lovins stemmer	Iterated Lovins
Number of unique words			
Agrovoc	20574	17293	15670
fao-all	169209	123975	107870
fao-browser	44226	30441	25013
fao-keyphrase	7886	5913	5284
cstr-all	105054	79628	73855
Number of Agrovoc words covered by words in...			
fao-all	9945	8685	8210
fao-browser	6186	5599	5384
fao-keyphrase	2483	2356	2294
cstr-all	3974	3908	3961
Proportion of Agrovoc words covered by words in...			
fao-all	48.3%	50.2%	52.4%
fao-browser	30.1%	32.4%	34.4%
fao-keyphrase	12.1%	13.6%	14.6%
cstr-all	19.3%	22.6%	25.3%

Table 2: Vocabulary overlap between AGROVOC and other phrase sets

	Unstemmed	Lovins stemmer	Iterated Lovins
Number of phrases			
Agrovoc phrases	27466	26701	25901
fao-all	19071445	18098815	17764015
fao-browser	278091	245374	233095
fao-keyphrase	13855	12183	11655
cstr-all	7364864	6945038	6859435
Number of Agrovoc phrases covered by...			
fao-all	9835	10750	10855
fao-browser	6166	6913	7014
fao-keyphrase	1447	1793	1874
cstr-all	1765	2731	3033
Proportion of Agrovoc phrases covered by...			
fao-all	35.8%	40.3%	41.9%
fao-browser	22.4%	25.9%	27.1%
fao-keyphrase	5.3%	6.7%	7.2%
cstr-all	6.4%	10.2%	11.7%

Table 3: Phrase overlap between AGROVOC and other phrase sets

AGROVOC terms, the FAO Web site, and the extracted hierarchy.

As was previously noted, stemming can affect the degree of match. We examine this effect by comparing the overlap between unstemmed phrases and phrases stemmed using the Lovins and Iterated Lovins algorithms [7]. The Lovins algorithm stems words to their root form; for example, *dictionary* is reduced to *diction*. The iterated algorithm repeatedly applies the Lovins stemmer until the stem no longer changes; *dictionary* is thus stemmed to *dict*. Tables 2 and 3 show that when words are stemmed more severely, the number of unique entries decreases because similar phrases are stemmed to equivalent root terms (Table 3, top row). At the same time, the proportion of matches increases as near-misses become equivalent. Although this effect may be significant, below we consider only the unstemmed version of each phrase, as these are phrases presented to the user in the browsing interface.

Recall

The proportion of the AGROVOC phrases that are covered by a phrase set, described in Table 3, is a measure of recall. For example, Table 3 shows that the browser phrases cover 22% of the AGROVOC phrases. In fact, we can calculate the AGROVOC recall of any set of phrases, and of any subset of the sets described above.

Figure 5 graphs the change in recall as the number of phrases in each of the four phrase sets is varied. To create this graph, the phrase sets were first sorted by frequency. Recall was calculated for the subset containing only the single most frequent phrase, then the subset containing the two most frequent phrases, then the three most frequent, and so on for as many phrases as are in each set. (Note that the horizontal axis uses a logarithmic scale.)

The resulting curves display the characteristics we would expect. The steepest curve is formed by *fao-keyphrases*, indicating they are generally of a high quality, but the curve terminates when 13,900 phrases and a recall of 5% are reached. The next curve, derived from *fao-browser*, is only slightly less steep but continues to a much higher recall value, and the third curve, *fao-all*, exhibits the greatest recall but also the largest number of phrases. Finally, the *cstr-all* curve is very shallow—even after seven million phrases are considered, the recall is only slightly higher than the recall for 13,000 keyphrases.

To form these curves, the phrase sets are sorted by "frequency", but this measure has a slightly different meaning in each case. The *fao-all* and *fao-browser* sets are sorted by the number of times they appear in the FAO Web pages. The *fao-keyphrases* are sorted by the number of times each keyphrase has been assigned to a document. When they are sorted by the number of times they occur on the Web site the curve is very similar, and terminates at the same point. The *cstr-all* phrases are sorted by the frequency with which they appear in the one thousand CSTR documents.

Precision

We now turn to the precision of each set of phrases. This is the proportion of the phrases that occur in AGROVOC—i.e. the number of phrases in the set that occur in AGROVOC divided by the total number of phrases in the set.

Figure 6 graphs the precision as the number of phrases selected increases, and again the curves

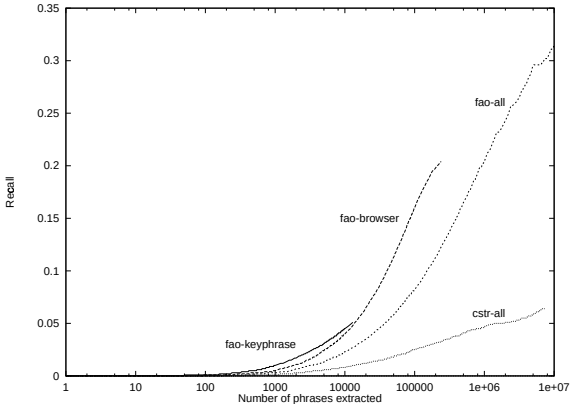

Figure 5: Recall of each phrase set

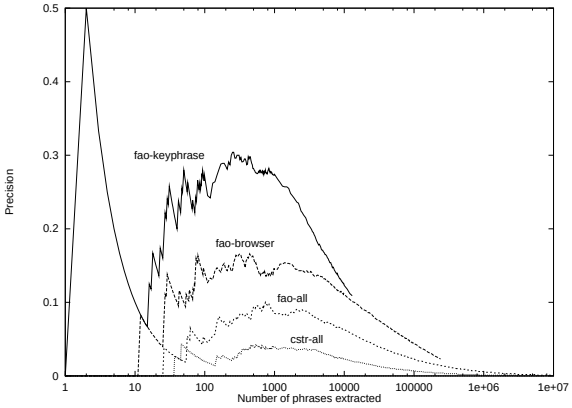

Figure 6: Precision of each phrase set

match the expectations outlined above. The *fao-keyphrases* have the greatest precision, followed by the *fao-browser* phrases and then the *fao-all* phrases. The highest precision achieved by the *cstr-all* phrases is less than 0.06, which means that even under the best of conditions only 6% of the most frequent computer science phrases appear in AGROVOC.

Precision-recall curve

Figure 7 shows a precision-recall curve for each of the four phrase sets. For most of their lengths, the *fao-keyphrase* dominates *fao-browser*, which dominates the *fao-all*, which dominates *cstr-all*. We conclude that the keyphrases are generally closest to the AGROVOC phrases because they are much more selective—only 14,000 keyphrases are extracted, compared to 278,000 browser phrases (Table 3). Consequently the recall attained by the browser phrases is much greater.

The consistently poor performance of *cstr-all* supports the hypothesis that the AGROVOC phrases have more in common with phrase sets based on the FAO Web site than with documents chosen from other fields of study.

Phrases not covered

It is interesting to consider which AGROVOC phrases were not found on the web site. The English language version of AGROVOC contains 27,000 terms. The FAO Web site contains 9800 (36%) of these phrases at least once; the remaining 17,600 (64%) do not appear in the collection. Our phrase browsing interface would be improved if it could recall more of these phrases.

Many of the phrases that do not appear might be characterized as "unusual" ones—for example, precise scientific names that are rarely used in normal discourse (examples appear in the third column of Table 5). Others are "ordinary" terms that are simply not used in the collection. We can get a rough feel for the number of "common" and "unusual" phrases by assuming that any word which appears in the Oxford English Dictionary is "ordinary", and dividing AGROVOC into three classes of phrases: those that comprise ordinary words, those that comprise specialist words, and those that contain at least one of each.

Table 4 breaks down the 27,000 phrases in AGROVOC into these groups, and shows how many of each type occur in the text of the FAO Web site. Not surprisingly, a far greater proportion of the ordinary phrases are detected than of phrases containing specialist words. However, the proportion of ordinary phrases detected is still only slightly over half (55%). Table 5 shows a selection of randomly chosen AGROVOC phrases that do not appear on the FAO web site.

EXTENDING THE THESAURUS

This analysis demonstrates that AGROVOC and the FAO web pages cover similar subject matter. We can exploit this relationship to attempt to extend the thesaurus. Our goal is to suggest new terms to the thesaurus maintainer; these new terms will be chosen from the phrases that appear in the phrase sets but are not already in the thesaurus.

We cannot tell which suggestions are good, and neither can our program. What we can do is bring phrases that are likely to be useful to the attention of the thesaurus maintainer.

Our aim in making suggestions is to identify those phrases that are most likely to be good thesaurus phrases but are not already in the thesaurus. Figure 6 shows the precision curve for each of the phrase sets; this tells us what proportion of the phrases in a set are actually in AGROVOC. If we assume that good thesaurus candidates are likely to be interspersed with the actual candidate phrases, then we can infer from Figure 6 that the best candidates will come from *fao-keyphrases* and *fao-browser*, and will appear among the most frequent phrases.

Table 6a lists the ten most frequent phrases that appear in *fao-keyphrase* and *fao-browser* in order of decreasing frequency. The phrases are not useful: instead of the hoped for list of meaningful English nouns, we have a list of foreign stopwords. (This is

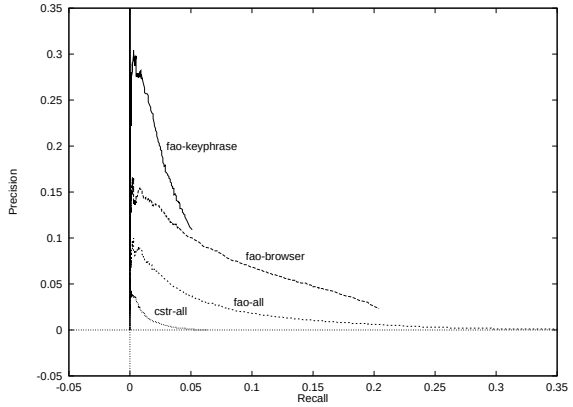

Figure 7: Precision-recall curve

because the system uses a part-of-speech tagger that cannot identify foreign words and assumes they are nouns.) In order to identify the best phrases in this list, we need to filter out undesirable phrases. This involves several steps.

First we eliminate all single-word phrases because they are seldom likely to be unambiguous. We then eliminate all phrases that contain one or more foreign words. The new list is shown in Table 6b.

The list is still not useful: it contains a variety of web page artifacts, trivial phrases, and the names of FAO organizations. We therefore eliminate any phrase that: is a Web page navigation instruction (e.g. *previous*, *next*); contains stopwords (e.g. *of*, *the*, *for*); is based on a faulty text parsing (e.g. *conf rence* for *conference*); contains the name of an FAO or other organization (e.g. *copyright fao*, *david lubin memorial library*); contain numbers or values that have received unusually high importance through repetition in tables (e.g. *700f*, *100f*); or are measurements or (e.g. *million tonnes*, *years yield*).

Table 6c shows the first twenty suggestions in the list after post-processing. It still contains phrases that are unlikely to be useful to thesaurus maintainers, but does contain what appear to be pertinent suggestions. It is the task of the thesaurus editor to evaluate these phrases and identify those that are useful, but we can draw some conclusions about the suggestions.

Some of the remaining phrases already appear in AGROVOC in a different form. For example, *agricultural census* and *agricultural production* appear as *agricultural censuses* and *agricultural product* respectively. Such phrases could be detected by comparing their stemmed forms to a stemmed version of AGROVOC.

Some phrases identify geographical areas, like *rome italy*, *asia pacific*, and *united states*, many of which already appear in AGROVOC. While the first example is probably not a good phrase, the second is potentially useful, and the third suggests that *united states* is a frequent contraction of *united states of america*, which appears in AGROVOC.

Class of phrase	AGROVOC	Number covered by *fao-all*	Percent covered by *fao-all*
Phrases comprising only ordinary words	11043	6048	54.8%
Phrases comprising both kinds of word	3055	622	20.4%
Phrases comprising only specialist words	13368	3165	23.7%

Table 4: Number of ordinary and specialist phrases in AGROVOC

Other phrases represent subtopics and variations of AGROVOC entries, and the thesaurus editor must decide whether they should be included. For example, *desert locusts*—the phrase explored in Figure 1—is similar to *locusts*, which appears in AGROVOC as a non-descriptor: *acrididae* is the preferred descriptors for locusts and grasshoppers. On this basis, it would seem reasonable to include *desert locust* as a non-descriptor equivalent to *schistocerca gregaria*, which appears in AGROVOC as a narrower term of *acrididae*.

The phrase *aquaculture production* is another potential thesaurus term. Several similar terms, including *aquaculture*, *aquaculture equipment*, and *aquaculture techniques*, already appear in AGROVOC. The phrases *forest genetic resources*, *plant genetic*, and *forest genetic* are all similar to the AGROVOC non-descriptor *plant genetic resources*, which is a synonym for the descriptor *genetic resources*, and for *forest resources*. The phrase *crop prospects* is similar to *crop management*.

Other phrases are clearly not suitable for a thesaurus. *Non wood* is the negation of an existing AGROVOC term, *wood*. Arguably, *non* is a stopword and this phrase should have been filtered out. *Socio economic* appears in the text as *socio-economic*; an equivalent term, *socioeconomic*, appears in AGROVOC. The intriguing phrase *sin embargo* possibly has a similar origin. Several organizational types are included, like *fisheries department*, *working group*, and *regional office*. *Explanatory notes*, *case study*, and *press releases* are all types of documents that occur on the FAO web site.

DISCUSSION

This study illustrates two facets of our work: aiding the browser with browsing interfaces that cover most topic areas, and aiding the thesaurus maintainer by highlighting domain-specific terms and phrases they may have overlooked.

Our analysis of the overlap between the AGROVOC and Web site vocabularies indicates that the two are similar enough that a tool linking the two hierarchies is likely to be useful. For example, a user might begin an interaction like that depicted in Figures 1 and 2 by entering the search term *forest* into the phrase-based browser. The most frequent phrase,

Ordinary phrases	Mixed phrases	Specialist phrases
air front	acacia aulacarpa	agrostis capillaris
artificial satellites	acacia mellifera	capitulum
christmas roses	acaricidal properties	dolichos hosei
dandelions	agave lecheguilla	entamoeba coli
duck plague virus	antidiuretic hormones	flexibacter
forced sale	austral islands	gortyna xanthenes
grub felling	diorite soils	hyperthermia
human offspring	eastern spruce budworm	lachnolaimus
metabolic sink	fe symbol	leishmania tropica
platypus mammal	gum kinos	lobeliaceae
sepals	hibiscus rostellatus	naegleria
shears	lactobacillus sake	orthosia
sour orange	queen bee excluders	pseudaletia separata
spruce gum	salvia aethiopis	teflon
stags	scophthalmus rhombus	urtica
tegument	slender wheatgrass	vigna vexillata
terrace cropping	soil microaggregates	xiphosura

Table 5: Ordinary and specialist phrases in AGROVOC but not in *fao-all*

which is *forest products*, might then be selected from those in Table 1. But this term is also represented in the AGROVOC thesaurus; access to the thesaurus would also have brought to the user's attention 44 specific types of forest product (for example, *Christmas trees*, *charcoal*, and *particle boards*), and 10 related topics (such as *logging wastes*, *cellulose products*, and *tanning agents*). These AGROVOC terms could then be browsed in the interactive interface. Interestingly, in the AGROVOC entry for *forest product*, three of the 54 narrower/related phrase links contain the word *forest*, one contains *forestry*, and six contain *products*. The majority of the AGROVOC links bring in new search or browsing terms for the user to consider.

We want to formalize the post-processing of suggestions. Several steps that are currently semi-automated, like identifying foreign terms and stopwords, could be performed automatically. Then we will generate a list of our best suggestions and send them to the FAO thesaurus maintainer.

We plan to extend our work to other thesauri. In order to make meaningful suggestions, we require an extensive body of text from which to extract phrases and an on-line version of the thesaurus. Our next project will examine the *Alcohol and Other Drug Thesaurus*.

	fao-keyphrases	*fao-browser*
a	de	de
	la	in
	000f	la
	et	en
	de la	for
	les	et
	500f	des
	food	les
	details	is
	forest	on
b	previous previous	rome italy
	title title page	such as
	title page contents	index terms
	page contents contents	fao rome
	million tonnes	copyright fao
	toc next page	for example
	agricultural census	director general
	index terms	area ha
	previous page toc	in many
	years carcass wt	home page
c	agricultural census	rome italy
	desert locust	holdings reporting
	forest genetic resources	united states
	commodity problems	main entry
	peoples participation	plant genetic
	non wood	socio economic
	explanatory notes	non wood
	aquaculture production	member countries
	asia pacific	agricultural production
	women and population	sin embargo
	environmental information	agricultural census
	intergovernmental group	coarse grain
	protected areas	asia pacific
	production of main	technical assistance
	aquaculture production trends	food summit
	fisheries department	natural resource
	regional office	working group
	press releases	fisheries management
	crop prospects	uruguay round
	case study	forest genetic

Table 6: Suggestions

REFERENCES

1. Chang S. J. and Rice R. E. (1993) "Browsing: a multidimensional framework." *Annual Review of Information Science and Technology 28*, pp. 231-276.
2. Food and Agriculture Organization of the United Nations (1995) *AGROVOC: multilingual agricultural thesaurus*. FAO, Rome.
3. Frank E., Paynter G. W., Witten I. H., Gutwin C. and Nevill-Manning C.G. (1999) "Domain-specific keyphrase extraction." *Proc Int Joint Conf on Artificial Intelligence*, pp. 668-673, Stockholm, Sweden.
4. Giles C. L., Bollacker K. & Lawrence S. (1998) "CiteSeer: An automatic citation indexing system." In *Proc ACM Digital Libraries*, pp. 89-98.
5. Gutwin C., Paynter G. W., Witten I. H., Nevill-Manning C., and Frank E. (2000) "Improving Browsing in Digital Libraries with Keyphrase Indexes." *J. Decision Support Systems*.
6. Jones S. and Paynter G. W. (1999) "Topic-based browsing within a digital library using keyphrases." *Proc ACM Digital Libraries*, pp. 114–121.
7. Lovins J. B. (1968) "Development of a Stemming Algorithm." *Mechanical Translation and Computational Linguistics*, Vol 11, pp. 2231.
8. Nevill-Manning C. G., Witten I. H. and Paynter G. W. (1999) "Lexically-generated subject hierarchies for browsing large collections." *Int J on Digital Libraries*, Vol. 2, No. 2/3, pp. 111-123; September.
9. Paynter G. W., Witten I. H., Cunningham S. J., Buchanan G. (2000) "Scalable browsing for large collections." *Proc ACM Digital Libraries*, pp. 215-223.
10. Witten I. H., McNab R. J., Boddie S., Bainbridge D. (1999) "Greenstone: a comprehensive open-source digital library software system." *Proc ACM Digital Libraries*, pp. 113-121.