Simpson, Jane. “Representing information about words digitally”. *Researchers, Communities, Institutions, Sound Recordings*, eds. Linda Barwick, Allan Marett, Jane Simpson and Amanda Harris. Sydney: University of Sydney, 2003.

This paper was originally published online as a refereed conference paper through *Open Conference Systems* by the University of Sydney, Faculty of Arts in 2004.

Representing information about words digitally

Jane Simpson

An HTML version of this paper can be found at:
http://www.paradisec.org.au/Simpson_paper_rev1.html

Introduction

Dictionary makers have four main tasks: to collect data, to analyse and organise it, to store it, and to display it for users. The advent of computers has changed our understanding of all four tasks, and has made new things possible in the areas of representing signs (spoken or signed), access to information, and indicating the denotations of words. However, as Bird and Simons (2003) note, new problems have arisen as a result of the new technology, in particular the tensions between the desire for longevity of format, the technical skills required to implement some of the new functions, and the small market for dictionaries of endangered languages. Sharp divisions are rising between what is possible for large commercial dictionaries of world languages, what is possible for dictionaries of endangered languages, and what is possible for one-off displays of word-lists. I illustrate these points mostly by reference to 'talking' dictionaries. The causes of these divisions lie not only in the relative lack of resources, but also in the difference in demands for longevity for dictionaries. I shall show how the new technology has affected the tasks of dictionary-making, before turning to the rise of multimedia dictionaries.

Collecting and interpreting data

The first task for a dictionary-maker is the collection of data. The main data collected about words for dictionaries concerns information about sound, meaning, use, history and relations to other words. Before computers became widely available, the main source of words was from written material, and to a much lesser extent transcription of spoken words. Thus dictionaries tended to represent the words used in writing, and not words or phrases used mainly when speaking. In the days before sound recording made it possible to retrieve and verify spoken words, the fact that a word appeared in written work, and was accepted by readers as a real word, was the main guarantee of the authenticity of the word. But lexicographers still rely on searching written transcriptions for new words, and then going from the transcript to the audio file to check for the pronunciation, since there are as yet no easy tools for searching streaming sound from, say, radio or television, for new words, or even for collecting many examples of the pronunciation of the same word.

Once a word has been identified as needing an entry, lexicographers interpret the material surrounding the word in the light of as many examples as they can collect and examine in the time available. The reliability of the interpretation depends in part on how much material the lexicographer can assess in preparing the entry. Gathering material has been made much easier by the arrival of computers, the development of large-scale digital corpora, the growing amounts of material on the internet, and the possibilities for cheap solicitation of data on the internet, (for example the Australian Broadcasting Commission and the Macquarie Dictionary's online collaboration at "Australian Word Map" http://www.abc.net.au/wordmap/ (Figure 1)). This has meant that the number of examples to examine has increased greatly, even for endangered languages where the body of texts is usually small. What can be done with the examples is limited by the lexicographers' time and ability to take in so many examples. Techniques for automatic sorting and classification of the information are needed (Fillmore and Atkins 1994), in order to reduce the burden on the lexicographer. Some are implemented in large commercial dictionaries. For example, lexicographers and computational linguists from the Information Technology Research Institute (Brighton) developed automated "word sketches" which showed the collocations of a given word, grouped into grammatical categories based on the British National Corpus. These were used in compiling the entries for Macmillan dictionaries (Kilgarriff and Rundell 2002). However, word sketches require natural language processing tools, such as lemmatizers and part-of-speech taggers, which would need to be developed for small languages. In Australia so far this has not happened, because the body of texts available for most language is too small to justify the labour, and because there are few computational linguists working on endangered languages.

On the basis of the collected material, lexicographers use their own judgment to create lexical entries, including guides to the pronunciation of the word. In the past, sound was particularly difficult to collect, because English spelling does not serve as an unambiguous guide to pronunciation. If the lexicographers had not heard the word, they could be led astray; thus the *Australian National Dictionary* (Ramson 1988) treats 'Nunga' (the name many South Australian Indigenous people call themselves) as though it were the same as 'Nyungar' (the name many southern Western Australian Indigenous people call themselves). The lexicographers did not realise that the orthographic 'u' in the first syllable represents different vowels in each case.

The lexicographers' pursuit of reliability of data and validity of interpretation of data is in part driven by the weight societies often place on dictionaries as reference works. Before computers were used in dictionary-making, it was hard to distribute primary data (sound, video, contemporary notes). The lexicographer's job was to synthesise and analyse that material, and the dictionary entry was a representation of that synthesis. In terms of information, the dictionary came to be seen as semi-primary documentation - a way of making primary data accessible through the filter of the lexicographer. References to published sources of example sentences, as in the *Australian National Dictionary* (Ramson 1988) are references to primary documentation for the existence of a word, and for its use with the meaning proposed by the lexicographer. They act as a guarantee of the lexicographer's proposal. The lexicographer's filter seemed invisible to the normal dictionary user, who often assigned to dictionaries the role of authorities, arbiters of what a word means, or what a reasonable person would understand by such and such a word. Once dictionaries were viewed as authorities, they gained the function of language standardisers. What a word means, or how it should be pronounced, is determined by the dictionary.

Now that digitisation has become widely available for text, sound and image from recordings of speech events, it means that the primary data of field recordings can be made widely available. This has led to the development of the distinction between language documentation and language description (Himmelmann 1998). However, making material available is not the same as making it accessible, and it is generally hard to find the information one wants quickly in raw field notes. So, dictionaries still serve as ways of making the primary documentation accessible. But more and more they are seen as consisting of two parts, a way of linking to primary documentation, and the lexicographers interpretation of that primary documentation. Such primary documentation could consist of sound and text collections, which can be linked to dictionaries (e.g. Baker and Kovacs 2003, Nathan 2000). These types of links provide the user with the means to judge the authenticity of the data. At the same time, computers have caused certain kinds of dictionaries to maintain the language standardising function - the dictionaries associated with spell-checkers help standardise the spelling of words, while the thesauruses associated with some word processing programs reinforce the idea that particular words are partial synonyms of each other.

Digital linking between the primary documentation and the lexical entries requires software, and the technical skills to use that software. Large commercial dictionaries can afford to employ programmers and even software designers to customise this software. Unfortunately for ordinary lexicographers, the most widely available software that is easily used is commercial proprietary software; for example FileMaker Pro allows the insertion of images and sound files (and see Baker and Manning 1998). The ease of use and the existence of documentation and a pool of other users make such products very attractive to small groups attempting to provide dictionaries for endangered languages, (or of other resources, such as image archives and catalogues). But the seductiveness of the ease of installation, data entry and linking blinds the users to the long-term storage problems of proprietary software as well as to the restrictions on display - it is difficult for naive users to produce attractive printed displays from FileMaker Pro. Ideally, dictionaries will be prepared in open source standards such as HTML or XML, which will allow multi-media linking, and have more chance of longevity (Bird and Simons 2003). At the moment, however, there are no widely available lexicography software packages for multimedia dictionaries which implement widely agreed-on standards for linking dictionaries.

Storage and Display

Once materials have been collected and analysed, the next tasks involve storage and display. The two are treated as independent actions, although they do interact. In the pre-computer days, the book was mostly treated as both storage and display; the card indexes from which the dictionary was created were rarely accessed, and often contained early versions of entries, rather than the final copy-edited and proof-read entry. Thus the book was usually the master copy. But the longevity and archiving problems were essentially solved by distribution of books, and the presence of copies of books in different libraries and archives. Once a book was printed and in libraries, the lexicographer could relax, thinking that the work was now in principle indefinitely available.

Having a book as the primary way of storing and displaying information has limitations. In terms of representation of information, the requirement that the information be representable on a page places limits on how helpful the representation of a word is. The appearance of, and information in, a dictionary entry are constrained by tradition as well as by the confines of the printed page, the cost of production of books, and the limits imposed by the need for books to be handled comfortably. For sign language dictionaries these limitations are highly constraining, as pictures and words do not convey adequately the movement of handshapes. The limitations for sounds can also be severe. Most readers have trouble with the International Phonetic Alphabet and the kinds of pronunciation guides used in dictionaries (Fraser 1997). This is a significant problem for endangered languages. Aboriginal participants in languages workshops have told me that it is painful to learn to use a dictionary, learn a word from the dictionary, and pronounce it to a speaker, only to have that speaker reject their pronunciation or fail to understand what they have said. This leads them to distrust dictionaries.

A related problem concerns the difference between the sound denoted by a word and the sound of a word as a sign, as in the following English example:

Tut /tʌt/ **tuts, tutting, tutted** 1. **Tut** is used in writing to represent a clicking sound that you make with your tongue to indicate disapproval, annoyance and sympathy 2. If someone **tuts**, they make a clicking sound with their tongue to indicate disapproval, annoyance or sympathy. *He tutted and shook his head.*
[*Collins Cobuild English Dictionary*. London Harper Collins 1995.]

The dictionary entry does not make it clear that there are now two different lexical items in our mental lexicons, the click /|/ or /!/ and the CVC sequence /tʌt/, which is also used in writing to represent the click sound. They directly result from the difficulty of representing click sounds in the English orthography. In writing we have developed a way of representing the sound /|/, as "tut". But from reading, we have started to pronounce "tut" as it is spelled, /tʌt/. We use this as a verb or interjection to express disapproval, /tʌt//tʌt/, as well as the click /|//|/. This brings up the need (raised in Burke (1998) and Sobkowiak (2003)), to provide audio files to represent the denotation of a word when it involves sound. In the following Warlpiri dictionary entry, the technical definition "retroflex affricate click" is of little use for the average reader of a dictionary. Being able to hear the sound of the young kangaroo would be a more effective way of letting them learn what *kirrkirrmani* denotes.

kirrkirr-ma-ni V. make click Definition: x produce clicking sound, characteristic sound produced by young kangaroo
Kirrkirr-manulparla kurdu-pardu marluku. The joey was going click, click, click to the kangaroo.
Note: Actual sound is like that of retroflex affricate click. *Warlpiri Dictionary*. (Laughren et al in prep.).

The problem of describing the denotation of *kirrkirrmani,* and the lack of clarity about the two pronunciations of "tut" arise because of the difficulty of recreating the sound of a word from a spelling or description of it. These problems can, and are, being solved now, because sounds can now be made available digitally. Commercial dictionaries such as the new Macmillan dictionaries (Rundell 2003) come with CD-ROMs that show both the pronunciation and the denotation of words denoting sounds. For a listener/viewer to go from a dictionary entry to an example of a word in a simple digital sound recording of a spoken text, there must be a link, which could be a link to a particular sound file, or to a time-coded section of a larger sound file, or it could be a link mediated through searching through a transcript of a sound file. However, doing this linking can be very time-consuming. The ordinary linguist or lexicographer does not yet have robust tools for efficient linking of sound and dictionary entry (and see Baker and Kovacs 2003).

Displaying dictionaries digitally has a number of advantages which have been discussed elsewhere (Corris et al 2000), including fast searching of big dictionaries (compare using the twenty volume *Oxford English Dictionary* with the online version http://dictionary.oed.com/), different ways of searching (for example, geographic interfaces (Figure 1), bilingual interfaces and thesaurus interfaces (Austin and Nathan 1998, Figure 2), and the ability to link entries to each other. The comparative lack of constraints on space means that pictures and sounds can be added in. Some of these ideas have been implemented in *Kirrkirr,* http://www-nlp.stanford.edu/kirrkirr/ (Manning et al 2001), a multimedia interface for dictionaries in XML format, which has been used for the *Warlpiri Dictionary* (Laughren et al in prep.) (Figure 3), as well as for Nahuatl. It was designed by Christopher Manning and Kevin Jansz, and is maintained by Manning.

In principle, the display of digital dictionaries can be refreshed and reproduced more easily than paper dictionaries; that is, new information can be imported. This is undoubtedly the case for large commercial dictionaries, with in-house programmers. But for lexicographers working on endangered languages, the time and labour spent in going from stored material to displayed material means that second editions of such dictionaries are rare. Another serious side-effect is that often last-minute changes are made on the displayed version of the dictionary, which then becomes the master, rather than the stored version. This is a particular problem when books are printed from databases. Since the stored version usually is more computationally tractable than the displayed version, this results in a loss of usefulness long-term.

Problems with digital dictionaries

As the preceding discussion has shown, producing a substantial digital dictionary which makes sensible use of the possibilities of digitisation (linking and multimedia) requires more computer skills than the "ordinary working linguist" (Lawler and Dry 1998) or lexicographer generally has. And, whereas when a paper dictionary is produced, it can be shelved in a library and left for hundreds of years, anything other than a plain text digital dictionary requires constant attention to ensure that it still works on the latest computers. This is not something that the ordinary lexicographer or linguist is trained to be conscious of. If they are conscious of it, then this tends to make them nervous about embarking on a project whose longevity is not guaranteed. The same problems arise with scholarly editions of literary or historical works; scholars are wary of spending several years of their lives preparing a work which may not be useable in ten years' time (Berrie 2001).

Distributing dictionaries on the web adds further complications, the internet allows user interaction, comments and feedback on words, as discussed earlier. In first world countries and in cities, access can be fast and cheap. However, in the remote areas and third world countries where most of the world's endangered languages are spoken, access to computers, download charges and phone charges can make online dictionaries inaccessible (and see Bird and Simons 2003). The internet offers widespread distribution, which helps short-term preservation ("lots of copies keep stuff safe" LOCKSS http://lockss.stanford.edu/). But, in the view of many Indigenous language speakers, this advantage is offset by the loss of intellectual property in their languages, once representations of them are freely available on the web. When wordlists of Indigenous languages are on the web, then people will collect those words and use them for purposes the original depositors did not dream of - electronic poetry for instance. Words from the lists will be used for names of houses, farms, boats, businesses, dogs and babies. For example, in 2003 the word *lardili-yan* was used on a sculpture in an exhibition in Gosford, New South Wales. This word is identical with the word for 'bird' in Wagiman, a language spoken in the Northern Territory of Australia, and was probably obtained from the excellent online Wagiman Dictionary http://www.aiatsis.gov.au/lbry/dig_prgm/e_access/digital/a339234/dict/dict.html (Wilson et al 2001, Figure 4) rather than from one of the handful of speakers (Mark Harvey p.c. 9 December 2003). Some speakers and communities will be happy about this, considering it a mark of interest in and respect for their language; others will resent it. The problem with resentment is that it will result in reluctance on the part of communities to have their material available digitally at all, let alone on the web. Thus, distributing a dictionary on the web requires careful thought about the consequences. One way of doing this is acknowledging the ownership of the languages, as in the first online dictionary of an Australian Indigenous language (Austin and Nathan 1998), which starts:

"The Kamilaroi/ Gamilaraay language belongs to the Kamilaroi people and to Kamilaroi country, northern New South Wales, Australia"

But in general, the idea that languages belong to the speakers is a new idea to many users of the Internet, who, as speakers of world languages, are accustomed to the idea that a language is available for anyone to learn and use, and that material on the web is there to be used in whatever way the user feels like.

Multimedia dictionaries: pictures

I turn now to specific concerns in building multimedia dictionaries. More and more dictionaries are making use of multimedia (Macmillan, as already mentioned, but also *Oxford Advanced Learners Dictionary on CD-ROM* http://www.oup.com/elt/global/products/oald/OALD_cdrom/ (Hornby and Wehmeier 2000), and the *American Heritage Dictionary* http://www.houghtonmifflinbooks.com/epub/ahd4.shtml*)*. Pictures have been the first to come, drawing on the old tradition of illustrated dictionaries. They are useful in improving the representation of signs (whether spoken or sign language signs), access to the information, and evocation of the denotation of words. Some electronic picture dictionaries are already available for Australian languages (Hamilton 1996-8) (Figure 5).

It is certainly true that for many words, their denotation can be more quickly grasped by looking at a picture. Words for natural kinds, tools, weather, and topographic terms are fairly readily picturable (of course many other words are not so easily picturable). However, my experience with collecting pictures for the Warlpiri dictionary is that it is much more time-consuming than expected. While there are many images available, ideally these will be tested with speakers to ensure that the speakers recognise what the pictures are supposed to represent, and that they are happy with the pictures being used to illustrate the word concerned (a point made by Joyce Hudson in the pre-computer days). Many pictures of plants, birds and animals, for example, do not give an indication of size. And many pictures of plants and animals do not give the habitat (creek bed, sandhill, rocky hill), although for Indigenous people of Central Australia habitat is often crucial in classifying and identifying plants and animals. A second problem is whether people are in the picture; such pictures need to be used with great caution in indigenous Australian communities, since people may be upset by seeing pictures of close relations who have died. A third problem is rights management. Photographers and illustrators must be acknowledged and their copyright in images protected. This of course can make web distribution problematic. For areas where many languages are spoken, we need "piccybanks", regional picture banks of images that have been agreed to be good representations, and for which the image creator has come to an arrangement over copyright.

Pictures are mostly used to indicate the denotation of a word. But they can be used for access - clicking on a picture to learn what the word for it is. For sign languages, pictures can also be used to represent the sign. However, video is generally more useful for this, since it allows the representation of movement and change of handshape (e.g. Johnston 1998). Video dictionaries of sign language are also useful in providing a medium for the deaf to find more information. Finally, video can be useful also in indicating the denotation of action words, and this is drawn on in existing video dictionaries of, for example, body-building terminology and classical ballet movements.

Multimedia dictionaries: sound

Sound has made its way into digital dictionaries. On 7th December 2003, I typed into Google the phrase "talking dictionary", and turned up about 36,100 hits and seven advertisements (amounting to 59 pages). A look at the results suggested that they fell into several types: handheld dictionaries for learners, dictionaries for the sight impaired, big English dictionaries (*Collins Cobuild Dictionary* http://www.cobuild.collins.co.uk/catalogue/cob4.html, *Oxford Advanced Learners Dictionary on CD-ROM* (Hornby and Wehmeier 2000), the Macmillan dictionaries, and the *American Heritage Dictionary* which claims "70,000 audio word pronunciations"), non-commercial bilingual dictionaries, and special purpose dictionaries.

Talking dictionaries for the sight impaired are an obvious use (Figure 6) - providing a medium for them to find more information. Language learners are an obvious market for talking dictionaries, because pronunciation is hard to replicate from transcriptions. Since there are many people who want to learn languages, the market both for talking versions of the big English dictionaries, and for the handheld talking dictionaries is potentially very large, and was the goal of the seven advertisements found. But, surprisingly, among the handheld dictionaries were some for languages which are rarely the targets of language learners- Albanian (Figure 7) being one such (perhaps the foreign aid workers and reconstruction workers provide a commercially viable market?). Most of these dictionaries provide the sounds of the words, but are accessed by typing. Some however offer voice recognition, 'say a phrase in language X, and receive a spoken translation in language Y' (Figure 8).

There are also a number of non-commercial bilingual dictionaries providing the sounds of words, for example a Hmong talking dictionary (Figure 9). Special interest groups also sometimes put up talking dictionaries, especially where the words are unusual, such as a dictionary of koi fish (Figure 10). These usually only contain a small number of words, reflecting the large amount of time presently needed to add digital sound to dictionaries.

It is quite noticeable however, that these dictionaries use sound mostly for representing the sign, and occasionally for accessing the information. They rarely use sound for indicating the denotation of the word. This is clearly an area for further development. For example, users would find it helpful to have the definitions of words denoting sounds or sound related actions extended with prototypical examples, e.g. for English screech, husky, sigh, opera or Warlpiri *kirrkirr-mani* (clicking), *yawulyu* (women's ceremony). Definitions are still essential - I don't think one could tell just from listening what can be called a 'screech' and what a 'scream'. Definitions of birds and animals could be enhanced by recordings of their characteristic sounds - particularly important for creatures like frogs which may be more often heard than seen. This would also be useful for creatures whose name is onomatopoeic - hearing the call of a crested bellbird would help the user understand why the Warumungu call that bird, *karnparnpalala.*

There are also many decisions to be made in collecting sound for dictionary, which usually need to be made in conjunction with the speakers of the language, and can be quite time-consuming. These include decisions such as:

• ***Whose voice to use?*** This is an important issue in Indigenous communities where language ownership is a living concern. The words should be pronounced by someone who has the authority to do so. This is also the case more generally, if dictionaries are treated as major representations of the language, or as carrying some kind of authority, or some other message. For example, the front page of the "Celebrity Talking Dictionary of breast cancer terms http://www.breastcancer.org/dictionary/welcome.php", (Figure 11), has as a selling point:

> "You'll be able to hear how the terms are pronounced and what they mean in the voices of the fabulous celebrities - from the media, sports, music, and medical worlds -- who have contributed to the Dictionary."

Perhaps the compilers are right. The vocabulary of breast cancer is intimidating, and hearing a term defined by someone who is not a specialist in the area, but who is known to the user, (and who perhaps the user admires) will probably make the definition easier to take in.

• ***How many speakers?*** Ideally, more than one speaker would be recorded; this allows users to abstract away from speaker peculiarities. It also gives them some choice, so that if they find it hard to listen to one voice, they can listen to a different voice.

• ***What variation should be represented?*** The answer depends on what the users and the community want the dictionary to represent - what is acceptable variation, and what is considered unacceptable, ungrammatical, substandard, uneducated etc. It is an important question to discuss, because the interests of the language documenter may clash with those of language standardisers, people who want their language represented by a standard form.

- ***Problems of collecting the sound (quality and content)*** A decision has to be made if the text is to be read or spoken. Natural spoken texts have the advantage of authenticity, but may result in unusable material if the speech is fast. If the text is to be read, there are often problems finding a literate speaker as opposed to a fluent speaker, let alone a literate speaker with authority. This speaker then has to be put in a good situation for recording (free of noise and interruptions). The speaker has also to agree to give up a large amount of time for the recording. There are also problems of content. The dictionary cannot be completed before the speaker records the words and sentences, because the speaker will inevitably make changes and have disagreements. In effect the speaker is acting as an editor for the dictionary; any words or sentences they object to will have to be modified in the accompanying print dictionary.

Making the decisions mentioned above can take months of negotiation and planning. For example, recording fewer than eight hundred words and sentences for the *Warumungu Illustrated Dictionary* has taken more than five months (Samantha Disbray p.c. December 2003). Similarly, a large amount of time is involved in digitising sound to archival quality, selecting the relevant examples of words, and linking the dictionary entries to the sounds. Baker and Kovacs (2003) show how an entry can be efficiently and automatically linked to occurrences of words in pairings of transcribed speech and digital sound, instead of the time-consuming and often inadequate splicing of words from digital recordings and manual linking to the dictionary entries. This looks as though it will be very useful, depending on the factors mentioned above. Of course, the usefulness does depend on how clear the examples in natural speech are, whether all the words required appear in good quality digital sound-transcription pairings, and whether the language owners are happy to have the speaker(s) in the texts represent them.

Conclusion

It is clear that multimedia dictionaries can in principle provide a much better representation of the sign, whether as a record of the sound of a word, or as a record of the movement and facial expression associated with a sign language sign. The potential for access by voice recognition or by keystroke also improves access to information for everyone, including the sight impaired and the hearing impaired. Sound and pictures, together with the definition, can also provide a much better evocation of the denotation of picturable words or sound words. But multimedia dictionaries bring to a head the problems mentioned earlier. Two problems that need to be addressed now are, firstly, that there is currently no easy way for an ordinary lexicographer to make a sophisticated multimedia digital dictionary which conforms to open source standards, without the help of a programmer. And secondly there is no easy way to revise and update such a dictionary. The problems of longevity and archiving are becoming increasingly severe. Now, when standards change, the dictionary maintainer has not just one thing to update, but a package of text, sound, picture and video and the links between them to update. Since standards, software and hardware for archiving text, video, picture and sound material may all change independently, maintaining useable copies of digital dictionaries in archives is now a big undertaking.

Large commercial lexicographers can afford this work, because they expect immediate commercial returns from book sales, they expect continuing demand which will fund new editions, and they know that for major languages the work of making dictionaries will continue. Longevity of format is not generally a major concern for them. People doing pilot projects with small amounts of data, such as the *Koi Fish Encyclopaedia* probably are not interested in longevity, and so can experiment with different kinds of software. People with access to computational linguists can engage in one-off projects which may produce a good result for a particular language. Most lexicographers of endangered languages have few resources, since the number of people who wish to learn endangered languages is usually too small to justify the time and cost involved in making a good digital dictionary. But they must consider carefully the longevity of their work. Theirs may be the final documentation of a language, and so must be stored in the best and most enduring format they can afford. They do not have access to computational linguists and they do not have time to spend learning software packages whose longterm archival survival is doubtful. When they go into partnership with information technology professionals, this frequently results in interfaces for the dictionary at hand (often in proprietary formats), rather than interfaces which can be easily installed and customised by a lexicographer for another language. The problem is exacerbated by funding bodies and Government departments that want to produce glossy CD-ROM products to show that they are doing something about endangered languages. This has resulted, and is resulting, in expensive CD-ROMs with lots of glitzy features that contain little language material, and are not customisable for other languages. They are **infertile plants**, and they are the recipient of most funding support. There has been no effort put into thinking about what users will do with these CD-ROMs. And yet the effort and expense involved in putting together a CD-ROM for endangered languages can only be justified if the CD-ROM will have lots of users who will spend a lot of time looking at it or doing things on it. Thus a CD-ROM of a big dictionary is justified, because people will use it time and again as a reference. A CD-ROM for teaching phonics will be justified if each year a new class of students will use it more than a couple of times. And so on.

In conclusion, time, labour, lack of accepted standards for dictionary software, lack of technical skills on the part of the lexicographer, and the need for longevity of format are major reasons why there are so few talking dictionaries of endangered languages. My 'lexicographic dream' (de Schryver 2003) is of **fertile plants**. A fertile plant would be shell dictionary software which is easy to use, and can be used for many languages. 'Easy to use' means that language workers can enter in data easily, browse it, update it, and access it (that is, not stored on a server accessible only by a decaying phone link). It should allow users to print out hard copies of different views of the data and different subsets of data. A key requirement of the software is that it should meet archival and portability standards (Bird and Simons 2003). This would make it easy to translate when new sound, text and picture standards emerge.

Acknowledgments
This paper draws on ideas developed from working with Miriam Corris, Christopher Manning and Susan Poetsch, as well as from conversations over the years with Peter Austin, Brett Baker, Jenny Green, Robert Hoogenraad, Doug Marmion, David Nash, David Nathan and Nick Thieberger, and with many other people trying to build dictionaries for languages they are working on. It also benefited from discussion at PARADISEC's Digital Audio Archiving Workshop http://conferences.arts.usyd.edu.au/index.php?cf=2 and from suggestions made by an anonymous reviewer and by Steven Bird.

References

ABC Online. 2002. *Australian Word Map*. http://www.abc.net.au/wordmap. Viewed on 10/12/03.

American Heritage Dictionary. 2000. Fourth Edition. Boston: Houghton Mifflin. Advertisement at http://www.houghtonmifflinbooks.com/catalog/titledetail.cfm?titleNumber=H25103 Viewed on 7/3/04.

Anonymous. n. d. *Celebrity Talking Dictionary of breast cancer terms*. http://www.breastcancer.org/dictionary/welcome.php Viewed on 10/12/03.

Anonymous. n.d. *LOCKSS*. http://lockss.stanford.edu/. Viewed on 10/12/03.

Austin, Peter, and Nathan, David. 1998. *Kamilaroi/Gamilaraay Web Dictionary*. http://coombs.anu.edu.au/WWWVLPages/AborigPages/LANG/GAMDICT/GAMDICT.HTM. Viewed on 9/12/03.

Baker, Brett, and Christopher Manning. 1998. "A dictionary database template for Australian Languages." Paper presented at AUSTRALEX. University of Queensland.

Baker, Brett, and Michael Kovacs. 2003. "The AudioText Linking Tool." In DIGITAL AUDIO ARCHIVING WORKSHOP: Researchers, communities, institutions and sound recordings: Research implications and feasibility of distributed digital archives for ethnographic sound material in Australia's geographic region. Sydney: University of Sydney: PARADISEC, 2003.

Berrie, Phillip William. 2001. "Are electronic editions inherently obsolete?" Paper presented at *Computing Arts 2001* Conference, Sydney. http://idun.itsc.adfa.edu.au/ASEC/JITM/Sydney200109PWB.pdf. Viewed 20/2/04

Bird, Steven, and Gary Simons. 2003. "Seven dimensions of portability for language documentation and description." *Language* 79(3): 557-582
Also http://www.language-archives.org/documents/portability.pdf. Viewed 13/4/04.

Burke, Sean Michael. 1998. *The design of online lexicons.* Master's thesis, Northwestern University.
http://www.speech.cs.cmu.edu/~sburke/ma. Viewed on 19/2/04.

Collins COBUILD. 2003. *Advanced Learner's English Dictionary*. With CD-ROM. Fourth Edition. Glasgow: Harper Collins Publishers. Advertisement:
http://www.cobuild.collins.co.uk/catalogue/cob4.html. Viewed on 10/12/03.

Corris, Miriam, Manning, Christopher, Poetsch, Susan, and Simpson, Jane. 2000. "Bilingual dictionaries for Australian Aboriginal languages: user studies on the place of paper and electronic dictionaries." In *Proceedings of the ninth EURALEX International Congress, EURALEX 2000*, eds. Ulrich Heid, Stefan Evert, Egbert Lehmann and Christian Rohrer, 169-181. Stuttgart: EURALEX.

de Schryver, Gilles-Maurice. 2003. "Lexicographers' dreams in the electronic dictionary age." *International Journal of Lexicography* 16:143-199.

Fillmore, Charles J., and Atkins, B.T.S. 1994. "Starting where the dictionaries stop: the challenge of corpus lexicography." In *Computational approaches to the lexicon*, eds. B.T. Sue Atkins and Antonio Zampolli, 349-393. Oxford: Oxford University Press.

Fraser, Helen. 1997. "Dictionary pronunciation guides for English." *International Journal of Lexicography* 10:181-208.

Hamilton, Philip. 1996-1998. *Uw Oykangand and Uw Olkola Multimedia Dictionary*.
http://www.geocities.com/Athens/Delphi/2970/. Viewed on 9/12/03.

Himmelmann, Nikolaus P. (1998). "Documentary and descriptive linguistics." *Linguistics* 36: 161-195

Hornby, A. S. and Sally Wehmeier. 2000. *Oxford Advanced Learner's Dictionary (with CD-ROM).* Sixth Edition. Advertisement and some online access at:
http://www.oup.com/elt/global/products/oald/OALD_cdrom/. Viewed on 10/12/03.

Johnston, Trevor. 1998, c1997. *Signs of Australia on CD-ROM* [a dictionary of Auslan (Australian Sign Language)]. North Rocks, N.S.W.: Royal Institute for Deaf and Blind Children.

Kilgarriff, Adam, and Michael Rundell. 2002. "Lexical Profiling Software and its lexicographic applications : a case study." In *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13-17, 2002*, eds. Anna Braasch and Claus Povlsen, 807-818. Copenhagen: Center for Sprogteknologi, Kbenhavns Universitet.
ftp://ftp.itri.bton.ac.uk/reports/ITRI-02-18.pdf Viewed on 19/2/04.

Laughren, Mary, Hale, Kenneth, and Hoogenraad, Robert. in prep. *Warlpiri dictionary*. Electronic datafiles. 2000 version. Brisbane: Department of English, University of Queensland.

Lawler, John M. , and Dry, Helen Aristar eds. 1998. *Using computers in linguistics : a practical guide*. London ; New York: Routledge.

Manning, Christopher D. , Jansz, Kevin, and Indurkhya, Nitin. 2001. "Kirrkirr: Software for browsing and visual exploration of a structured Warlpiri dictionary." *Literary and Linguistic Computing* 16:123-139. See also http://www-nlp.stanford.edu/kirrkirr/. Viewed on 10/12/03.

Nathan, David. 2000. "The spoken Karaim CD: sound, text, lexicon and "active morphology" for language learning multimedia." In *Studies on Turkish and Turkic Languages (Proceedings of the Ninth International Conference on Turkish Linguistics)*, eds. Asli Goksel and Celia Kerslake. Wiesbaden: Harrassowitz. http://www.dnathan.com/papers/karaimcd2.pdf. Viewed on 9/12/03

Ramson, William S. ed. 1988. *The Australian National Dictionary: a dictionary of Australianisms on historical principles*. Melbourne: Oxford University Press.

Rundell, Michael (Editor-in-chief). 2002. *The Macmillan English dictionary for advanced learners*. London: Macmillan.
http://www.macmillandictionary.com/about.htm. Viewed on 19/2/04.

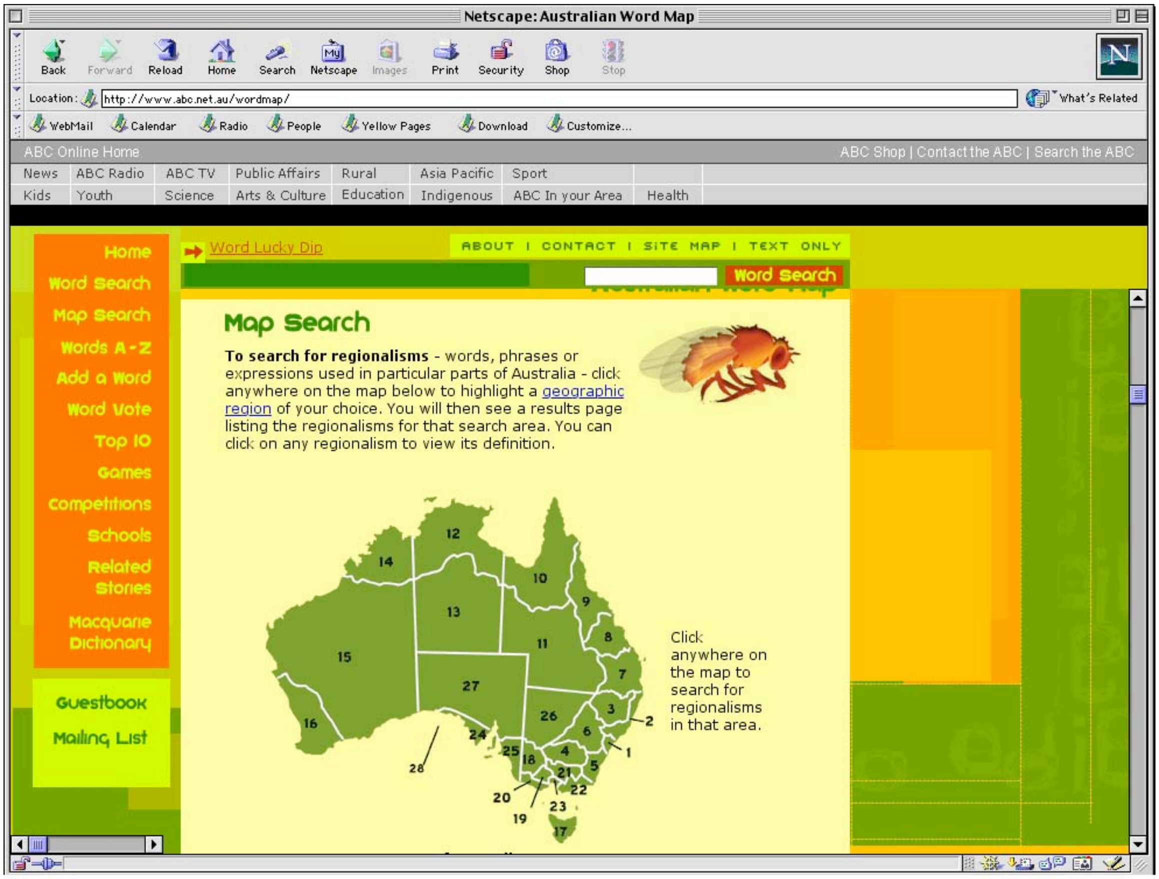
Rundell, Michael (Editor-in-chief). 2003. *The Macmillan essential dictionary*. London: Macmillan. With CD-ROM.

Sobkowiak, Wlodzimierz. 2003. "Review article: Pronunciation in Macmillan English Dictionary for Advanced Learners on CD-ROM." *International Journal of Lexicography* 16, no. 4: 423-441.

Wilson, Stephen, Harvey, Mark, Dalpbalngali, Lulu Martin, Emorrotjba, Helen Liddy, Benbo, Paddy Huddlestone, Gumbirtbirtda, Clara McMahon, and Gapbuya, Lenny Liddy. 2001. *The Wagiman online dictionary*.
http://www.aiatsis.gov.au/lbry/dig_prgm/e_access/digital/a339234/dict/dict.html. Viewed on 9/12/03.

Figures

Figure 1: Wordlist with a geographic interface and the possibility of interactive adding of words: screenshot of ABC Online. 2002. *Australian Word Map*. http://www.abc.net.au/wordmap. Viewed on 10/12/03.



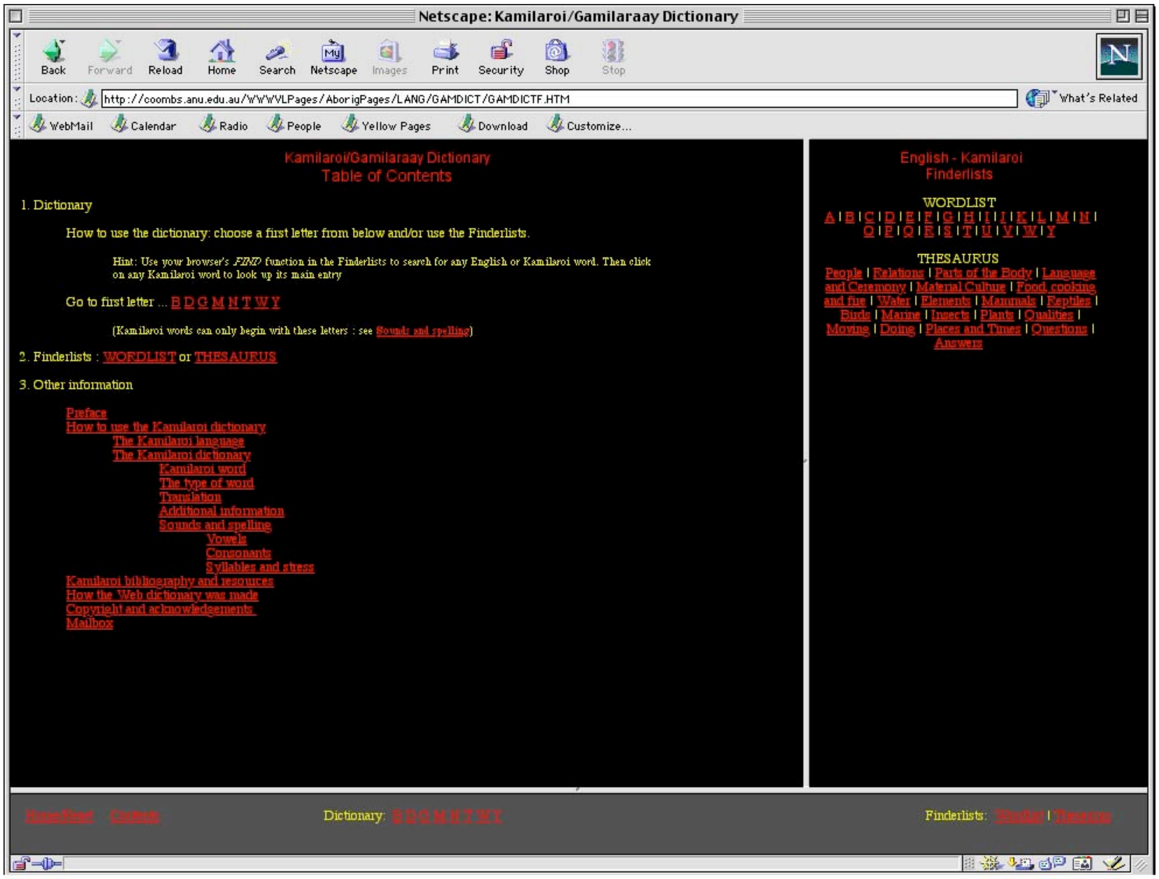
Map Search

To search for regionalisms - words, phrases or expressions used in particular parts of Australia - click anywhere on the map below to highlight a geographic region of your choice. You will then see a results page listing the regionalisms for that search area. You can click on any regionalism to view its definition.

Click anywhere on the map to search for regionalisms in that area.

Back to text

Figure 2: Dictionaries with multiple ways of looking up information: screenshot of Austin, Peter, and Nathan, David. 1998. *Kamilaroi/Gamilaraay Web Dictionary*. http://coombs.anu.edu.au/WWWVLPages/AborigPages/LANG/GAMDICT/GAMDICT.HTM. Viewed on 9/12/03.



Location: http://coombs.anu.edu.au/WWWVLPages/AborigPages/LANG/GAMDICT/GAMDICTF.HTM

Kamilaroi/Gamilaraay Dictionary
Table of Contents

1. Dictionary

How to use the dictionary: choose a first letter from below and/or use the Finderlists.

Hint: Use your browser's *FIND* function in the Finderlists to search for any English or Kamilaroi word. Then click on any Kamilaroi word to look up its main entry

Go to first letter ... B D G M N T W Y

(Kamilaroi words can only begin with these letters : see Sounds and spelling)

2. Finderlists : WORDLIST or THESAURUS

3. Other information

- Preface
- How to use the Kamilaroi dictionary
  - The Kamilaroi language
  - The Kamilaroi dictionary
    - Kamilaroi word
    - The type of word
    - Translation
    - Additional information
    - Sounds and spelling
      - Vowels
      - Consonants
      - Syllables and stress
- Kamilaroi bibliography and resources
- How the Web dictionary was made
- Copyright and acknowledgements
- Mailbox

English - Kamilaroi
Finderlists

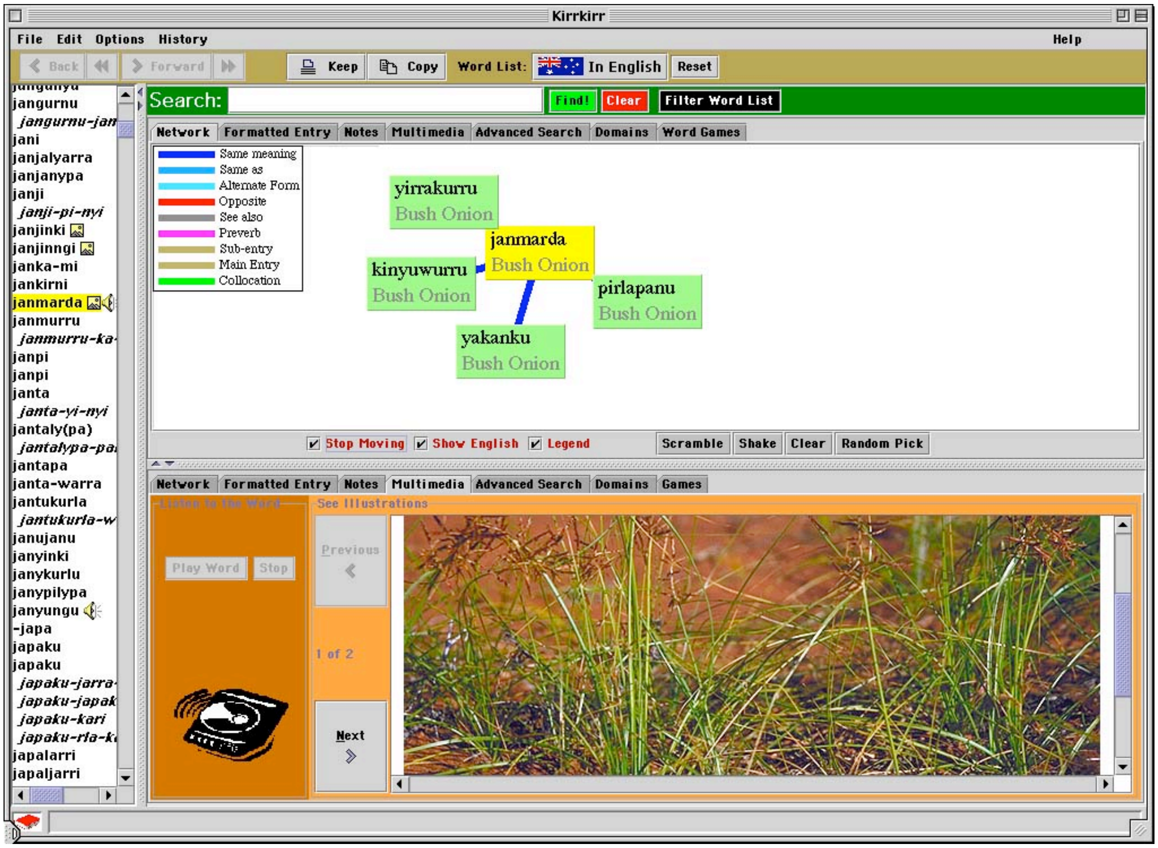
WORDLIST
A | B | C | D | E | F | G | H | I | J | K | L | M | N |
O | P | Q | R | S | T | U | V | W | Y

THESAURUS
People | Relations | Parts of the Body | Language and Ceremony | Material Culture | Food, cooking and fire | Water | Elements | Mammals | Reptiles | Birds | Marine | Insects | Plants | Qualities | Moving | Doing | Places and Times | Questions | Answers

[illegible] [illegible] Dictionary: B D G M N T W Y Finderlists: Wordlist | Thesaurus

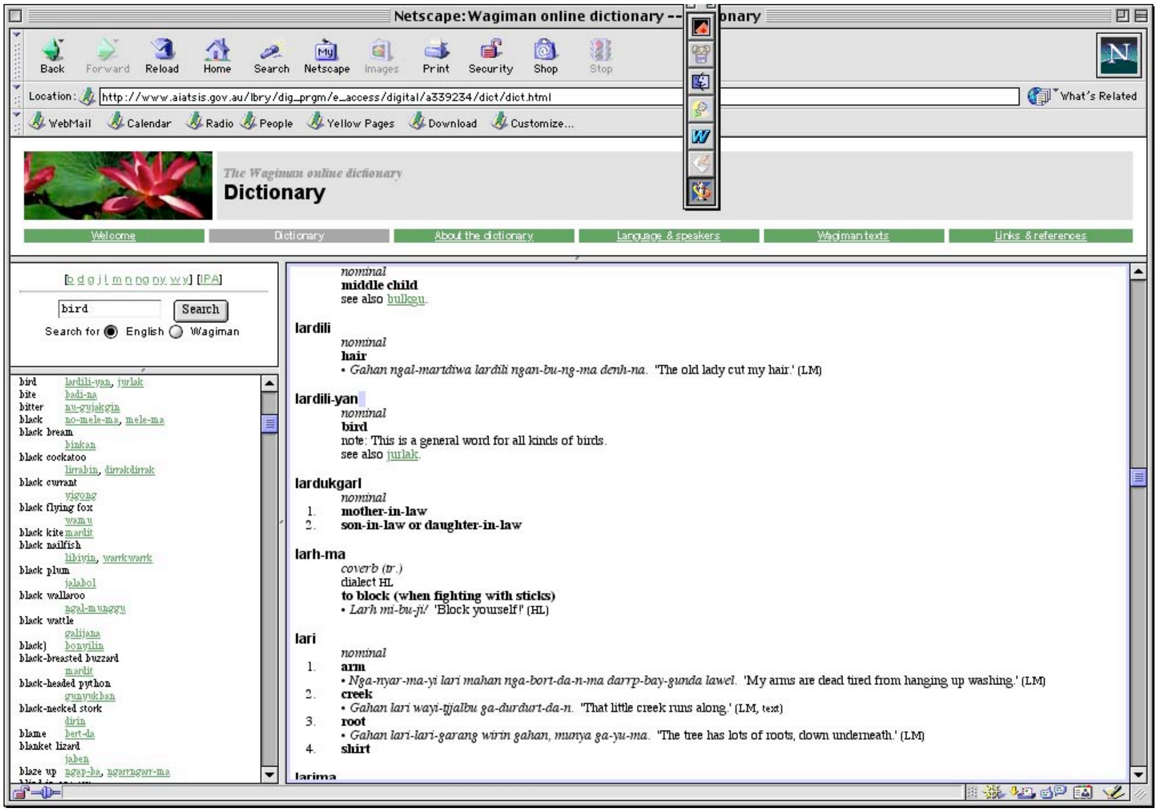
[Back to text]

Figure 3: Multimedia dictionaries: the *Kirrkirr* interface for the *Warlpiri Dictionary*.



Back to text

Figure 4: *The Wagiman online dictionary:* screenshot of: Wilson et al 2001. *The Wagiman online dictionary*. http://www.aiatsis.gov.au/lbry/dig_prgm/e_access/digital/a339234/dict/dict.html. Viewed on 9/12/03.



bird	lardili-yan, jurlak
bite	badi-na
bitter	nu-gujakgin
black	no-mele-ma, mele-ma
black bream	binkan
black cockatoo	lirrabin, dirrakdirrak
black currant	yigong
black flying fox	wamu
black kite	mardit
black nailfish	libiyin, warrkwarrk
black plum	jalabol
black wallaroo	ngal-munggu
black wattle	galijana
black)	bonyilin
black-breasted buzzard	mardit
black-headed python	gunyukban
black-necked stork	dirin
blame	bert-da
blanket lizard	jaben
blaze up	ngap-ba, ngarrngarr-ma

nominal
middle child
see also bulkgu.

lardili
nominal
hair
• *Gahan ngal-martdiwa lardili ngan-bu-ng-ma denh-na.* 'The old lady cut my hair.' (LM)

lardili-yan
nominal
bird
note: This is a general word for all kinds of birds.
see also jurlak.

lardukgarl
nominal
1. **mother-in-law**
2. **son-in-law or daughter-in-law**

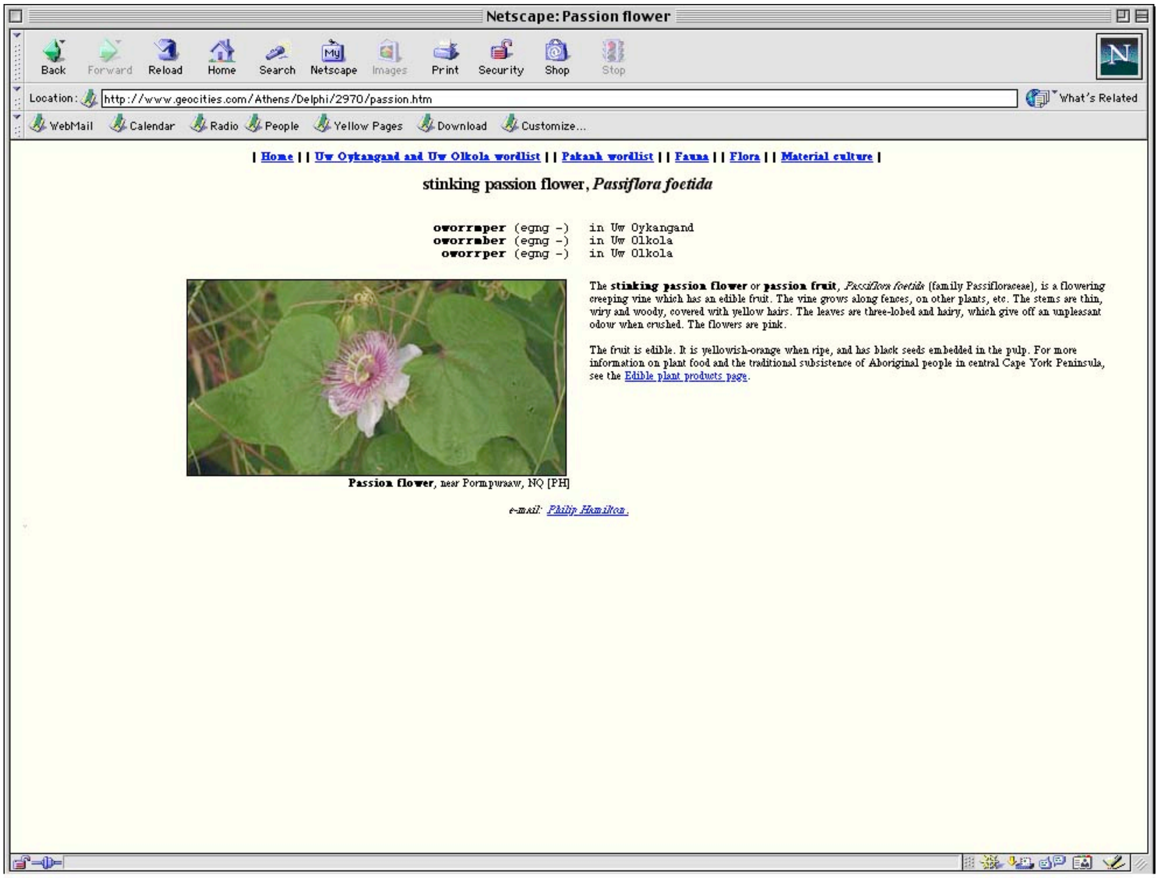
larh-ma
coverb (tr.)
dialect HL
to block (when fighting with sticks)
• *Larh mi-bu-ji!* 'Block yourself!' (HL)

lari
nominal
1. **arm**
• *Nga-nyar-ma-yi lari mahan nga-bort-da-n-ma darrp-bay-gunda lawel.* 'My arms are dead tired from hanging up washing.' (LM)
2. **creek**
• *Gahan lari wayi-tjjalbu ga-durdurt-da-n.* 'That little creek runs along.' (LM, text)
3. **root**
• *Gahan lari-lari-garang wirin gahan, munya ga-yu-ma.* 'The tree has lots of roots, down underneath.' (LM)
4. **shirt**

larima

Back to text

Figure 5: Dictionaries with pictures: Screen shot of Hamilton, Philip. 1996-1998. *Uw Oykangand and Uw Olkola Multimedia Dictionary*. http://www.geocities.com/Athens/Delphi/2970/. Viewed on 9/12/03.



| **Home** | | **Uw Oykangand and Uw Olkola wordlist** | | **Pakanh wordlist** | | **Fauna** | | **Flora** | | **Material culture** |

stinking passion flower, *Passiflora foetida*

```
oworr■per (egng -)   in Uw Oykangand
oworr■ber (egng -)   in Uw Olkola
 oworrper (egng -)   in Uw Olkola
```

The **stinking passion flower** or **passion fruit**, *Passiflora foetida* (family Passifloraceae), is a flowering creeping vine which has an edible fruit. The vine grows along fences, on other plants, etc. The stems are thin, wiry and woody, covered with yellow hairs. The leaves are three-lobed and hairy, which give off an unpleasant odour when crushed. The flowers are pink.

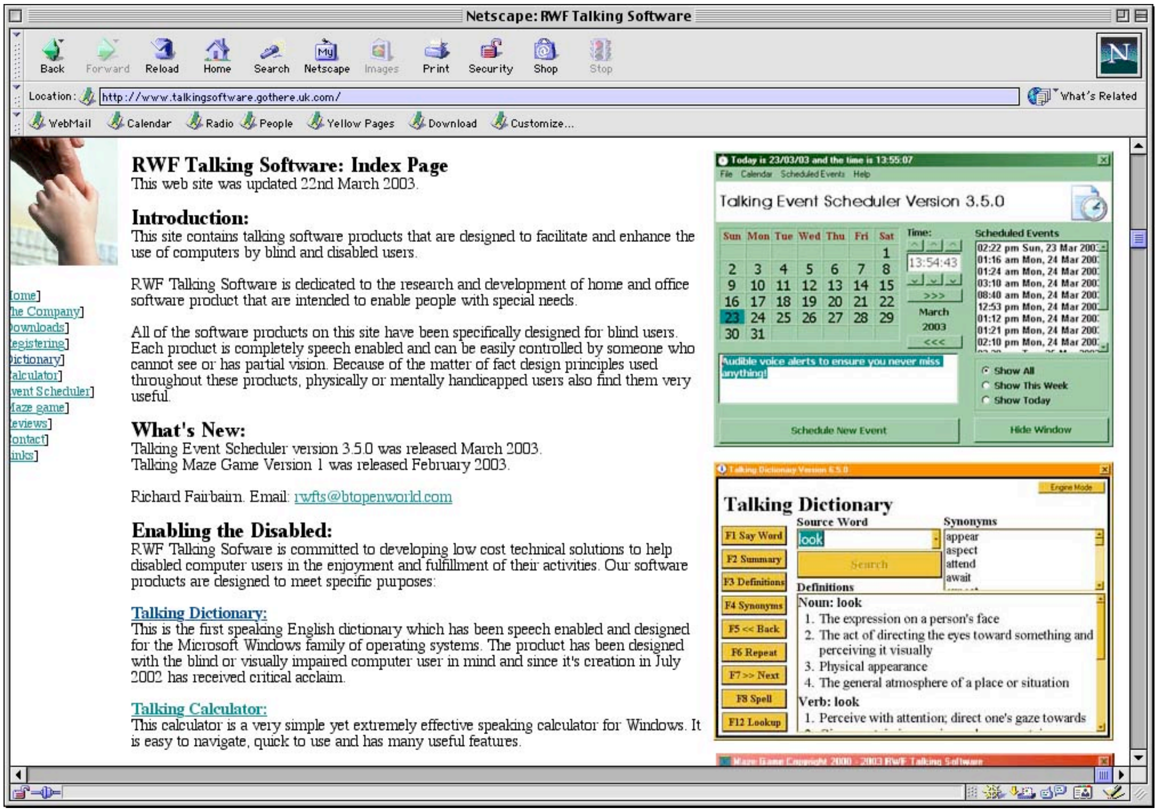
The fruit is edible. It is yellowish-orange when ripe, and has black seeds embedded in the pulp. For more information on plant food and the traditional subsistence of Aboriginal people in central Cape York Peninsula, see the Edible plant products page.

Passion flower, near Pormpuraaw, NQ [PH]

e-mail: *Philip Hamilton*.

Back to text

Figure 6: Talking dictionaries: dictionaries for the sight-impaired: screen shot of http://www.talkingsoftware.gothere.uk.com/. Viewed on 10/12/03.



Location: http://www.talkingsoftware.gothere.uk.com/

RWF Talking Software: Index Page

This web site was updated 22nd March 2003.

Introduction:

This site contains talking software products that are designed to facilitate and enhance the use of computers by blind and disabled users.

RWF Talking Software is dedicated to the research and development of home and office software product that are intended to enable people with special needs.

All of the software products on this site have been specifically designed for blind users. Each product is completely speech enabled and can be easily controlled by someone who cannot see or has partial vision. Because of the matter of fact design principles used throughout these products, physically or mentally handicapped users also find them very useful.

What's New:

Talking Event Scheduler version 3.5.0 was released March 2003.
Talking Maze Game Version 1 was released February 2003.

Richard Fairbairn. Email: rwfts@btopenworld.com

Enabling the Disabled:

RWF Talking Sofware is committed to developing low cost technical solutions to help disabled computer users in the enjoyment and fulfillment of their activities. Our software products are designed to meet specific purposes:

Talking Dictionary:
This is the first speaking English dictionary which has been speech enabled and designed for the Microsoft Windows family of operating systems. The product has been designed with the blind or visually impaired computer user in mind and since it's creation in July 2002 has received critical acclaim.

Talking Calculator:
This calculator is a very simple yet extremely effective speaking calculator for Windows. It is easy to navigate, quick to use and has many useful features.

Today is 23/03/03 and the time is 13:55:07

Talking Event Scheduler Version 3.5.0

Sun	Mon	Tue	Wed	Thu	Fri	Sat
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

Time: 13:54:43
March 2003

Scheduled Events
02:22 pm Sun, 23 Mar 200
01:16 am Mon, 24 Mar 200
01:24 am Mon, 24 Mar 200
03:10 am Mon, 24 Mar 200
08:40 am Mon, 24 Mar 200
12:53 pm Mon, 24 Mar 200
01:12 pm Mon, 24 Mar 200
01:21 pm Mon, 24 Mar 200
02:10 pm Mon, 24 Mar 200

Audible voice alerts to ensure you never miss anything!

● Show All
○ Show This Week
○ Show Today

Talking Dictionary Version 6.5.0

Talking Dictionary

Source Word: look

Synonyms
appear
aspect
attend
await

Definitions

Noun: look

1. The expression on a person's face
2. The act of directing the eyes toward something and perceiving it visually
3. Physical appearance
4. The general atmosphere of a place or situation

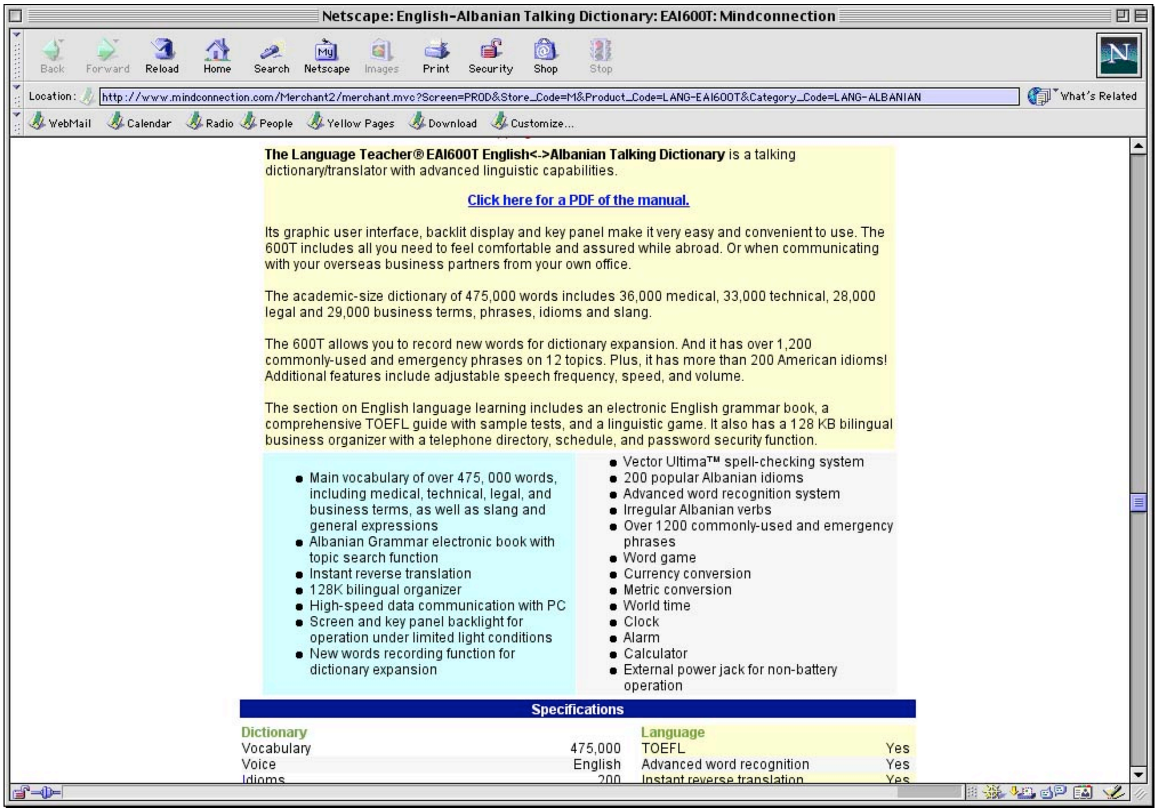
Verb: look

1. Perceive with attention; direct one's gaze towards

Maze Game Copyright 2000 - 2003 RWF Talking Software

Back to text

Figure 7: Talking dictionaries: Albanian handheld dictionary: screenshot of http://www.mindconnection.com/Merchant2/merchant.mvc?Screen=PROD&Store_Code=M&Product_Code=LANG-EAl600T&Category_Code=LANG-ALBANIAN Viewed on 10/12/03.



Netscape: English-Albanian Talking Dictionary: EAI600T: Mindconnection

Location: http://www.mindconnection.com/Merchant2/merchant.mvc?Screen=PROD&Store_Code=M&Product_Code=LANG-EAI600T&Category_Code=LANG-ALBANIAN

The Language Teacher® EAI600T English<->Albanian Talking Dictionary is a talking dictionary/translator with advanced linguistic capabilities.

Click here for a PDF of the manual.

Its graphic user interface, backlit display and key panel make it very easy and convenient to use. The 600T includes all you need to feel comfortable and assured while abroad. Or when communicating with your overseas business partners from your own office.

The academic-size dictionary of 475,000 words includes 36,000 medical, 33,000 technical, 28,000 legal and 29,000 business terms, phrases, idioms and slang.

The 600T allows you to record new words for dictionary expansion. And it has over 1,200 commonly-used and emergency phrases on 12 topics. Plus, it has more than 200 American idioms! Additional features include adjustable speech frequency, speed, and volume.

The section on English language learning includes an electronic English grammar book, a comprehensive TOEFL guide with sample tests, and a linguistic game. It also has a 128 KB bilingual business organizer with a telephone directory, schedule, and password security function.

- Main vocabulary of over 475, 000 words, including medical, technical, legal, and business terms, as well as slang and general expressions
- Albanian Grammar electronic book with topic search function
- Instant reverse translation
- 128K bilingual organizer
- High-speed data communication with PC
- Screen and key panel backlight for operation under limited light conditions
- New words recording function for dictionary expansion

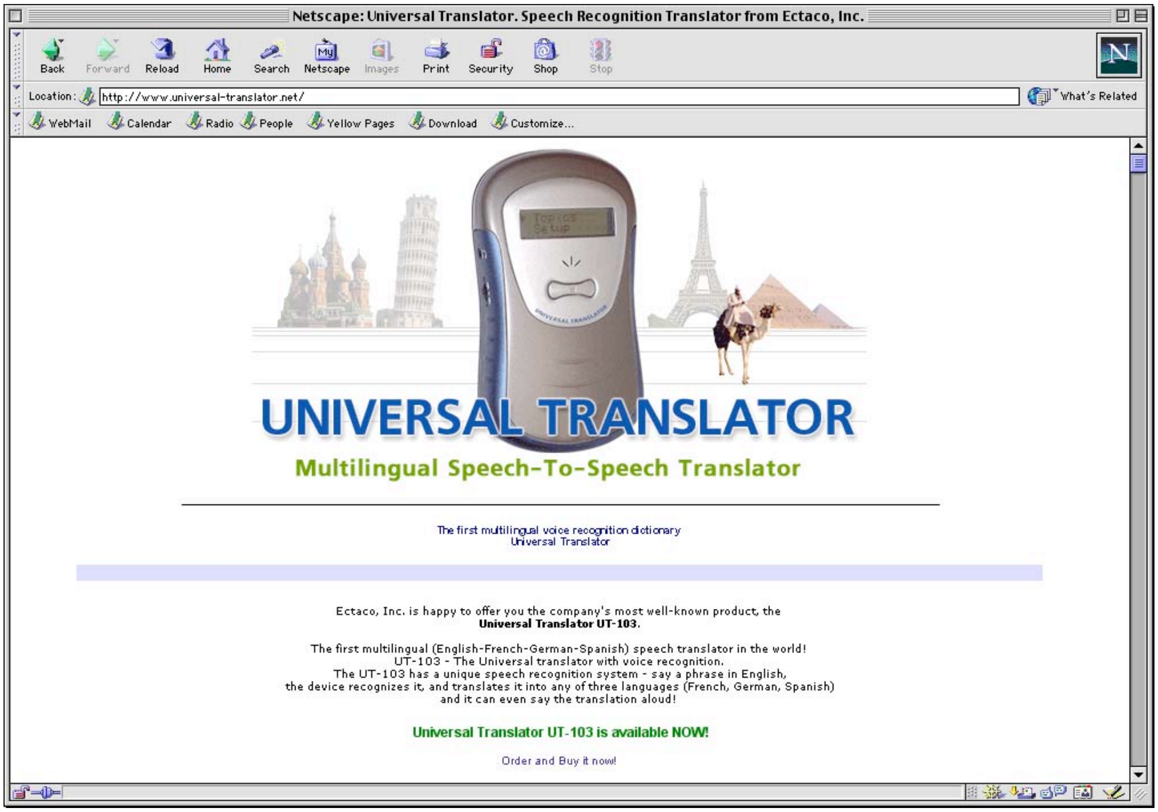
- Vector Ultima™ spell-checking system
- 200 popular Albanian idioms
- Advanced word recognition system
- Irregular Albanian verbs
- Over 1200 commonly-used and emergency phrases
- Word game
- Currency conversion
- Metric conversion
- World time
- Clock
- Alarm
- Calculator
- External power jack for non-battery operation

Specifications

Dictionary		Language	
Vocabulary	475,000	TOEFL	Yes
Voice	English	Advanced word recognition	Yes
Idioms	200	Instant reverse translation	Yes

Back to text

Figure 8: Talking dictionaries: dictionaries with speech recognition: screenshot of http://www.universal-translator.net/. Viewed on 10/12/03.



http://www.universal-translator.net/

UNIVERSAL TRANSLATOR

Multilingual Speech-To-Speech Translator

The first multilingual voice recognition dictionary
Universal Translator

Ectaco, Inc. is happy to offer you the company's most well-known product, the
Universal Translator UT-103.

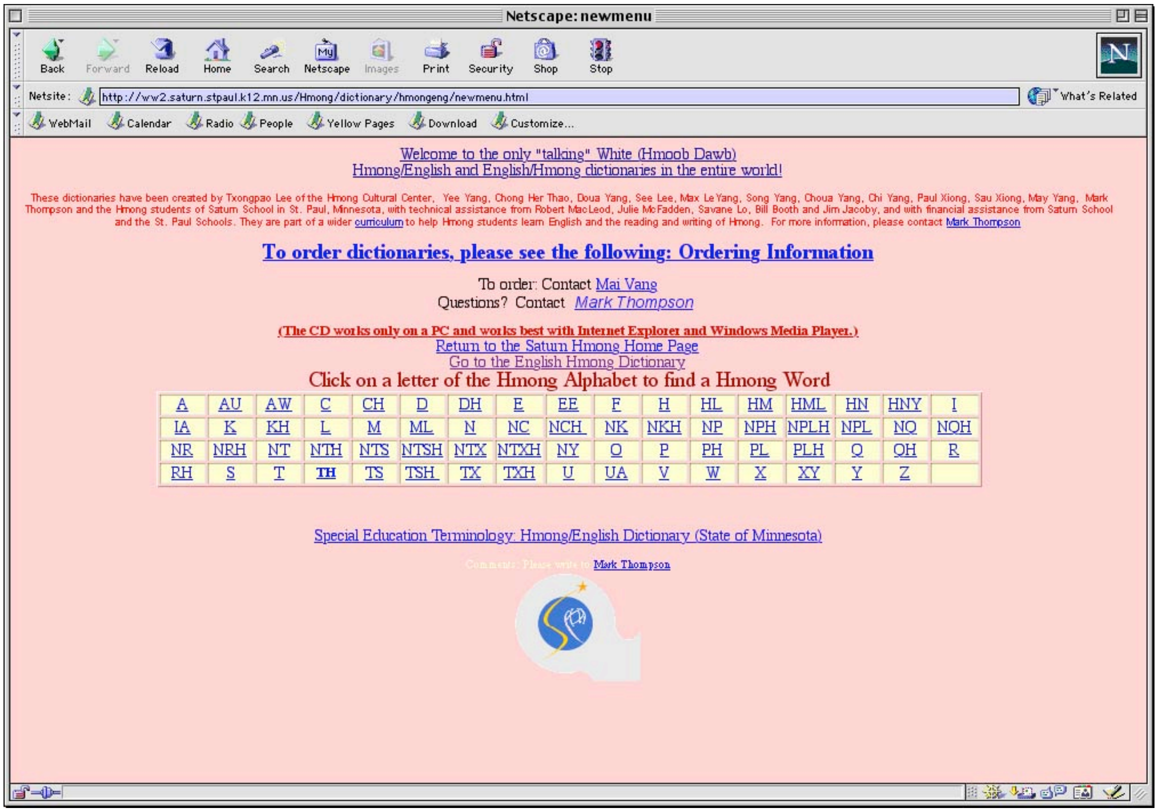
The first multilingual (English-French-German-Spanish) speech translator in the world!
UT-103 - The Universal translator with voice recognition.
The UT-103 has a unique speech recognition system - say a phrase in English,
the device recognizes it, and translates it into any of three languages (French, German, Spanish)
and it can even say the translation aloud!

Universal Translator UT-103 is available NOW!

Order and Buy it now!

Back to text

Figure 9: Talking dictionaries: A Hmong talking dictionary: screenshot of http://ww2.saturn.stpaul.k12.mn.us/Hmong/dictionary/hmongeng/newmenu.html Viewed on 10/12/03.



http://ww2.saturn.stpaul.k12.mn.us/Hmong/dictionary/hmongeng/newmenu.html

Welcome to the only "talking" White (Hmoob Dawb)
Hmong/English and English/Hmong dictionaries in the entire world!

These dictionaries have been created by Txongpao Lee of the Hmong Cultural Center, Yee Yang, Chong Her Thao, Doua Yang, See Lee, Max Le Yang, Song Yang, Choua Yang, Chi Yang, Paul Xiong, Sau Xiong, May Yang, Mark Thompson and the Hmong students of Saturn School in St. Paul, Minnesota, with technical assistance from Robert MacLeod, Julie McFadden, Savane Lo, Bill Booth and Jim Jacoby, and with financial assistance from Saturn School and the St. Paul Schools. They are part of a wider curriculum to help Hmong students learn English and the reading and writing of Hmong. For more information, please contact Mark Thompson

To order dictionaries, please see the following: Ordering Information

To order: Contact Mai Vang
Questions? Contact *Mark Thompson*

(The CD works only on a PC and works best with Internet Explorer and Windows Media Player.)
Return to the Saturn Hmong Home Page
Go to the English Hmong Dictionary

Click on a letter of the Hmong Alphabet to find a Hmong Word

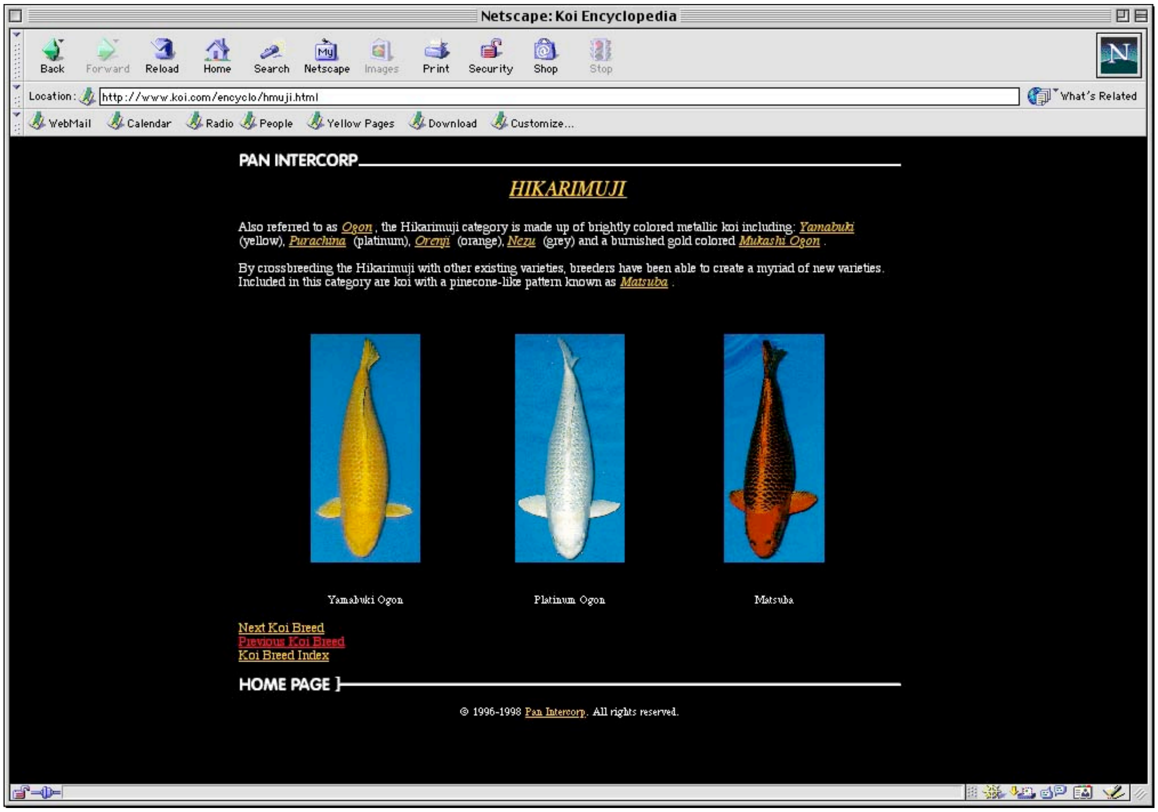
A	AU	AW	C	CH	D	DH	E	EE	F	H	HL	HM	HML	HN	HNY	I
IA	K	KH	L	M	ML	N	NC	NCH	NK	NKH	NP	NPH	NPLH	NPL	NQ	NQH
NR	NRH	NT	NTH	NTS	NTSH	NTX	NTXH	NY	O	P	PH	PL	PLH	Q	QH	R
RH	S	T	TH	TS	TSH	TX	TXH	U	UA	V	W	X	XY	Y	Z	

Special Education Terminology: Hmong/English Dictionary (State of Minnesota)

Comments : Please write to Mark Thompson

Back to text

Figure 10: Talking dictionaries: Koi fish encyclopaedia: screenshot of http://www.koi.com/encyclo/hmuji.html Viewed on 10/12/03.



PAN INTERCORP

HIKARIMUJI

Also referred to as *Ogon* , the Hikarimuji category is made up of brightly colored metallic koi including: *Yamabuki* (yellow), *Purachina* (platinum), *Orenji* (orange), *Nezu* (grey) and a burnished gold colored *Mukashi Ogon* .

By crossbreeding the Hikarimuji with other existing varieties, breeders have been able to create a myriad of new varieties. Included in this category are koi with a pinecone-like pattern known as *Matsuba* .

Yamabuki Ogon

Platinum Ogon

Matsuba

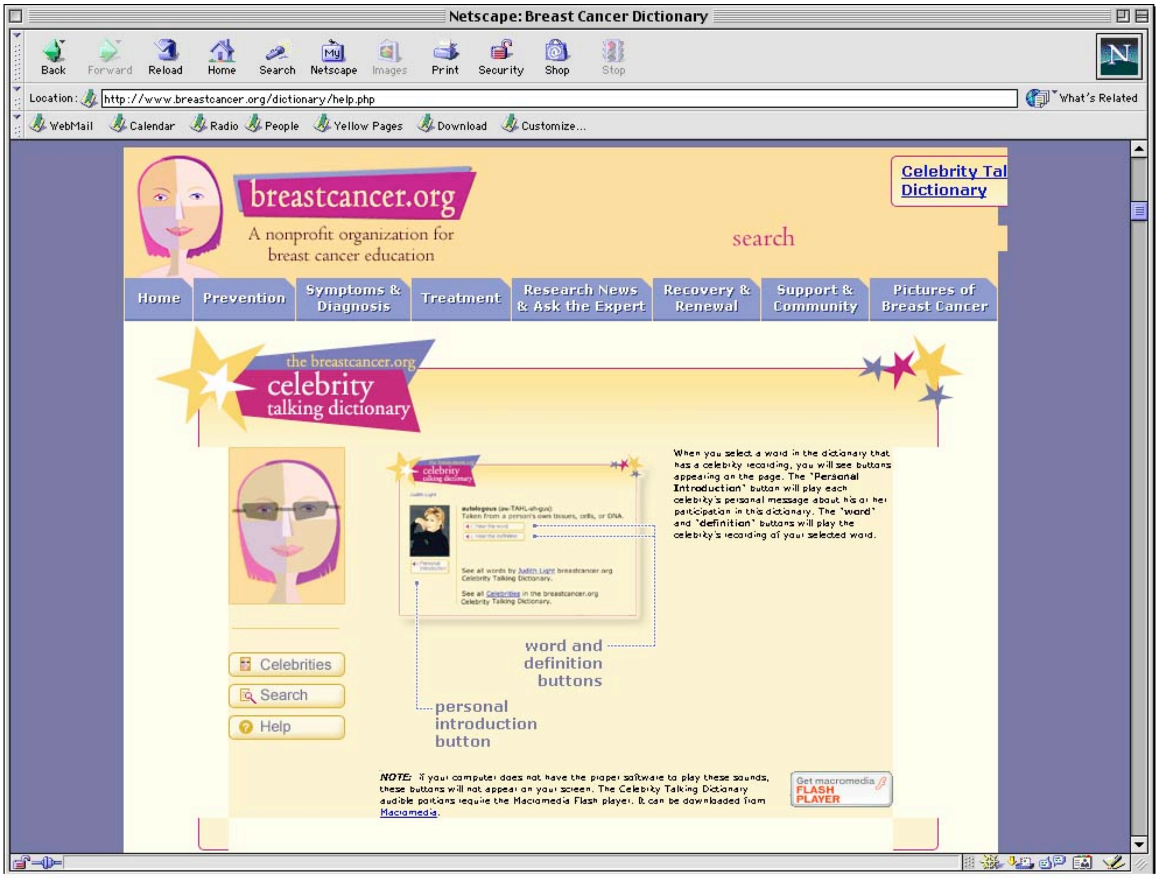
Next Koi Breed
Previous Koi Breed
Koi Breed Index

HOME PAGE]

© 1996-1998 Pan Intercorp. All rights reserved.

Back to text

Figure 11: Talking dictionary: Celebrity Talking Dictionary of breast cancer terms: screenshot of http://www.breastcancer.org/dictionary/welcome.php Viewed on 10/12/03.



autologous (aw-TAHL-uh-gus)
Taken from a person's own tissues, cells, or DNA.

See all words by Judith Light breastcancer.org Celebrity Talking Dictionary.

See all Celebrities in the breastcancer.org Celebrity Talking Dictionary.

When you select a word in the dictionary that has a celebrity recording, you will see buttons appearing on the page. The "Personal Introduction" button will play each celebrity's personal message about his or her participation in this dictionary. The "word" and "definition" buttons will play the celebrity's recording of your selected word.

NOTE: If your computer does not have the proper software to play these sounds, these buttons will not appear on your screen. The Celebrity Talking Dictionary audible portions require the Macromedia Flash player. It can be downloaded from Macromedia.

Back to text