Dictionaries on the Internet: an Overview

Vincent J. DOCHERTY, München, Germany

Abstract

The proliferation of uncontrolled, perhaps uncontrollable dictionary resources on the Internet provokes the question as to whether Samuel Johnson's comment on dictionaries still applies in the age of the Web: "Dictionaries are like watches, the worst is better than none and the best cannot be expected to go quite true" [Piozzi 1925]. Is, by analogy, a free lexicographical resource on the Net better than nothing?

In this paper we shall examine the reactions to our withdrawal of two large bilingual dictionaries from an Internet site and the implications these reactions have for serious publishers with regard to their presence on the Web.

1 Open-source projects

1.1 The tradition of open-source lexicography

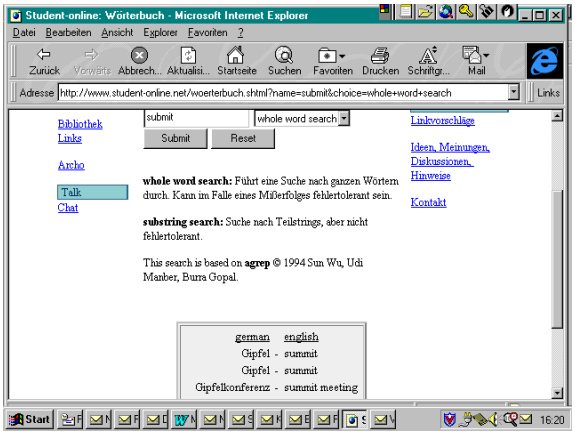
As early as 1857 the publishers of the Oxford English Dictionary initiated a reading programme which saw mainly volunteers provide the in-house lexicographical team with citations of the usage of individual words. Such reading programmes have since become an institution for most serious dictionary publishers, although today the readers are usually paid or have been replaced by text corpora which are analysed automatically.

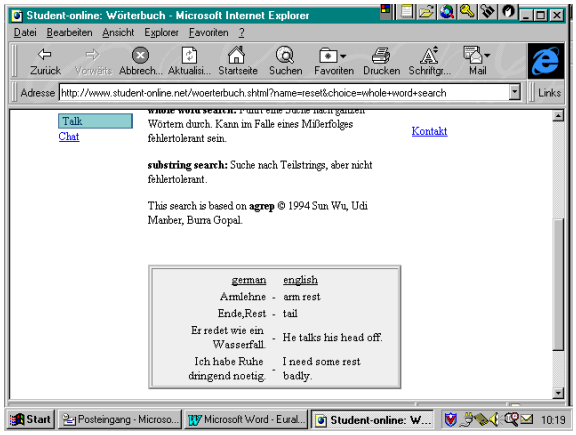
In the field of software development, the advantages of open-source program improvement have clearly been recognized:

> Surprisingly, the individual contributions of thousands of independent online volunteers can lead to better software than the coordinated efforts of a company's paid employees. That, at least, is the lesson from open-source projects such as Linux, an operating system that this week caught Wall Street's imagination when shares in Red Hat, a Linux supplier, surged on news of a big contract. [Economist 1999]

1.2 Open-source on the Web

Open-source projects on the Internet offer Web surfers free access to a large volume of information, including dictionary resources. At the same time, individual users are invited to comment on, add to and improve the existing entries, thus creating a dynamic resource which is wonderfully interactive but, by definition, uncontrolled and perhaps even uncontrollable. In an article published on the Internet appealing for participation by others in "collaborative linguistic knowledge production", [Parks 1999] maintains that "just as the meanings of the words we speak are collaborative creations, so the intelligence in our software is something we must collectively produce and be collectively responsible for" . This so-called "bottom-up collaborative editing" [Carr 1997] can, however, offer no guarantees with regard to the quality of the content, as the following entry from a site consisting of the work of German students shows:

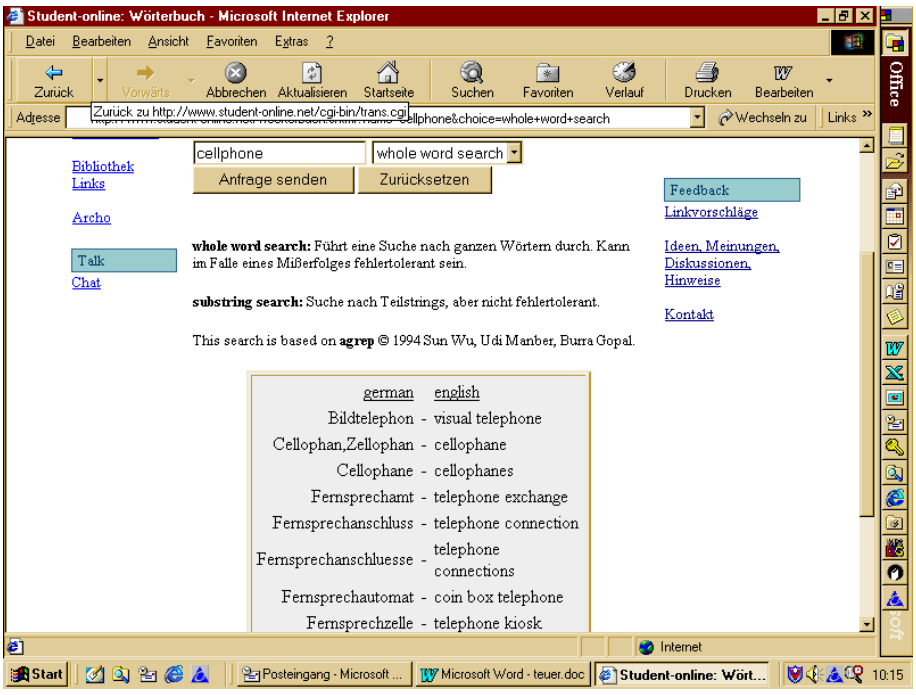




The first prompt word to confront the user is "submit", referring to the search word which should be entered. If the user was unable to understand the meaning of "submit" and entered this word, the answer to the query would be "summit" = "Gipfel", with no more than a cryptic message on spelling tolerance when the required word is not in the database.

A search for the other command – "reset" – results in an equally unusable list of translations, (such as "Armlehne" = "arm rest"), all of which are more or less anagrams of "reset" but none of which are of any relevance to the user.

The following choice of translations is given for the word "cellphone" . Here the user is confronted not only by a mixture of old and new German spelling forms ("Mißerfolges", "Fernsprechanschluss") but also by a number of words with absolutely no connection to the search words. What might at first sight appear to be a useful overview of possible German equivalents of "cellphone" is in effect little more than a list of all near hits from a database with a curious and extremely limited population. Conspicuous by their absence are any indicators reflecting typical usage restrictions:



Again the user is given the same brief indication of why so many variants are offered even though the search word itself is not included: "Kann im Falle eines Mißerfolges fehlertolerant sein" .

While some purists argue that the most important resource on the Internet is the "netizens" themselves, there can be little doubt that uncontrolled authorship can be extremely dangerous if the user is seeking quality and reliability – two of the traditional strengths of professional lexicography.

Even a good resource such as Leo (www.dict.leo.org) offers "das schnurlose Handtelefon" as one of the possible translations of the English word "handy" :

One of the major differences between the open-source projects on the Net and the reading programmes run by serious dictionary publishers is the quality of the "filter" (if indeed there is one). Experienced full-time lexicographers will always be capable of adjusting and improving articles compiled by would-be authors, but this is precisely what is missing from most "bottom-up" projects. It is clear that dictionaries of the sort cited from above are often more of a hindrance than a help to anyone attempting to negotiate the rough seas of the English language on the Internet.

1.3 Ordering the chaos

The suggestion has been put forward that there is a need for some sort of "ordering force" to be set up on the Internet to counterbalance the amorphous mass of material available in unstructured form. In Germany, the information portal Xipolis, a joint venture between Brockhaus and Holtzbrink, aims to achieve this: "Brockhaus/Holtzbrink sind ... auf Vertrauenswürdigkeit eingestellt und setzen ’Wissen mit Herkunft’ gegen ’unstrukturierten Datenmüll’ im Internet" [Buchreport 2000]. It goes almost without saying that this service has its price.

2 The role of serious dictionary publishers on the Internet

2.1 Withdrawing dictionaries from the Internet

A number of traditionally print-driven dictionary publishers have placed quality contents on the Web, sometimes free of charge, but usually for a fee (the method of payment still being one of the major obstacles to this use of a fairly new medium). Long before Encyclopaedia Britannica went online for free, Langenscheidt had two large-scale bilinguals on the Internet. As part of a cooperation with a partner in the field of machine translation, Langenscheidt's Handwörterbuch Englisch (English-German-English) and their New College Spanish dictionary (English-Spanish-English) were available unabridged and at no cost to users for over a year on the partner's Web site. Although not consciously encouraged by Langenscheidt, the presence of these dictionaries was at least tolerated, despite the fact that Langenscheidt could not directly guage the popularity of these works.

Providing a large dictionary online free of charge is a double-edged sword. On the one hand it can offer potential customers the chance to work with the product in real-life situations and encourage them eventually to buy the print product or even other electronic applications, such as CD-ROMs, which might provide a more user-friendly interface. It is also a very positive form of advertising and image improvement. Yet at the same time it creates the expectation among users that they have a right to "freebies", that online services should be of high quality but of no cost to the consumer.

When Langenscheidt finally decided to remove the two dictionaries from their site, the reactions this move provoked among the works' aficionados were both swift and unexpectedly cutting:

> *Betreff: Schwache Idee*
>
> *Das war ja nicht grade eine geistige Meisterleistung, das English Online Dictionary vom Internet zu entfernen. Danke!!!*
> *... wahrscheinlich verkaufen Sie jetzt viieeeel mehr CD-Roms und es gibt keinen Missbrauch mehr und die Direktion hat das so entschieden und ich habe sowieso keinen Anspruch auf ein kostenloses Wörterbuch usw....????*
> *Ich finds einfach schade.*

> *Sehr geehrte Damen und Herren,*
>
> *Soeben stelle ich mit Entsetzen fest, daß es Ihr englisches Online-Wörterbuch unter http://www.lhs-lt.de/products/bt1woerter.htm nicht mehr zu geben scheint.*
> *Was nun?*

Others threatened to boycott the company and stick to free resources:

> *Es ist natürlich Ihre Entscheidung, ein vorhandenes WWW-Angebot zurückzuziehen, ich für meinen Teil werde auf andere Angebote (z.B. http://dict.leo.org/) ausweichen ...*

There was, mixed with the animosity, a good deal of praise for the service Langenscheidt had provided:

> *Das on-line Deutsch-Englisches Worterbuch war das Beste! Haben Sie es gestoppt weil Verkaufzahlen fallen oder ist die URL nur nicht leicht zu finden?*
>
> *Could you tell me what happened to the great electronic Langenscheidt dictionary at http://www.lhs-lt.de/products/t1woerter.htm Is that dictionary available elsewhere on the web now, perhaps on your own web site?? It was a very valuable resource to me, filling a need that my paper copy of your dictionary can't meet. Many thanks,*
>
> *sehr geehrte verantwortliche für das online wörterbuch, welches seit heute nicht mehr verfügbar ist,*
> *ich habe mit großem bedauern die mitteilung heute gelesen. ich werde ihr produkt sehr vermissen. ich habe sowohl den teil deutsch-englisch wie auch englisch-spanisch (meine muttersprache) sehr intensiv benutzt. ich bin der meinung, es handelte sich um ein angebot erster qualität – inhaltlich wie auch aus der sicht der technischen realisierung. ich kann insofern nachvollziehen, daß sie es nicht mehr zur kostenlosen verfügung bereitstellen wollen. ich möchte auf jedem fall mein herzliches dankeschön! sagen.*
>
> *I was a daily user of Langenscheidt New College English – Spanish Dictionary. I'm really disappointed now that it is not longer available online. It is such an excellent dictionary that I'm surprised it was put off line. May I ask what was the reason for doing so? Is it going to be available again in the future?*
>
> *... I want to congratulate the responsible person (or team) for the service provided through that dictionary, it was great while it lasted.*

Interestingly, a number of users at least mentioned the possibility of paying for access to the dictionaries:

> *If you were to put it back on the web and charged a monthly or yearly access fee, I would consider subscribing, depending on the cost.*
>
> *Ich habe überhaupt kein Problem, den Service zu bezahlen, solange ich das Wörterbuch nur wieder bekomme.*

Finally, in a curious pidgin German, one user summed up the linguistic chaos of the Net in a poignantly formulated cry for help:

> *Ich nichts sprechen sie gut deutsch„„„„„Ich habe use die langenscheidt wörterbuch zu helf mich transulate english zu deutsch.....Jetzt nichts kann verbinden nichts mehr„„„„„„„*
>
> *Bitte helfen Mir warum nichts kann verbinden. Ich habe student kommen zu meine haus zu leben fur ein jahr und ich muß use langenscheidt zu helf transulate.*
> *Danke*

This range of reactions reflects the dilemma facing serious publishers of dictionaries, for whom the Internet might seem to be a no-win situation. If they provide quality content at no cost to the user, this service is soon taken for granted and considered a "right" . If the publishers opt out of the Internet and leave it to the more chaotic content providers, the quality of language learning and cross-boundary communication in general could be seriously threatened.

A number of downloadable dictionary resources represent almost as pernicious a threat as the online resources to the established dictionary publishing houses. Start-up companies such as Babylon, whose initial raison d'être was to be bought by Microsoft, put together contents of varying quality and offer them online for downloading free of charge. Some of these companies survive on advertising revenue, while others are funded by venture capitalists looking for long-term returns. Data capture and data control are not subject to the strict criteria that are standard in traditional publishing houses. The problem of lack of quality is at times even reflected in the description of the product, as the following examples from the rather unfortunately named QuickDic range of dictionaries show:

> Lediglich ein kleines Symbol ... zeigt, das [sic] QuickDic aktiv ist... egal ob es sich um eine Textverarbeitung, eine [sic] Internet-Browser, eine Tabellenkalkulation, ein [sic] Editor, ein [sic] Online-Dienst-Software oder eine andere AnwendOung [sic] handelt.[1]

2.2 Copyright and the Internet

The situation is compounded by an increasingly lax attitude on the part of many members of the Web community towards intellectual property rights. One of Langenscheidt's English-Turkish dictionaries was simply keyed in by a small company in Turkey and placed on the Internet. Spot checks by Langenscheidt showed the content to be identical to that of its own dictionary – with the exception of a number of typing errors which obviously occurred when the data was being captured. In another case, a former slaughterhouse butcher whose eyesight had so deteriorated that he was unable to continue in his profession, chose to spend his time keying in large publishers' Japanese dictionaries. He offered a package of five of them on the Internet for a fee, but withdrew them when threatened with legal action.

This problem is only the tip of the iceberg in a small but relatively lucrative branch of publishing where "consulting" the works of competitors has a long and not always glorious tradition anyway. Now so-called dot coms are openly appealing to authors to provide them with dictionaries they just happen to have in their bottom drawers:

> **Sie haben ein interessantes Programm entwickelt oder vielleicht ein Wörterbuch erarbeitet?**
>
> Wenn Sie diese Frage mit ja beantworten können, sollten Sie sich einmal an uns wenden ... Wenn Sie sich in einem Fachbereich bestens auskennen oder eine Fremdsprache perfekt beherrschen, dann könnten Sie auch ein Wörterbuch für eines unserer Produkte erarbeiten! Voraussetzung allerdings wäre das [sic] Sie sich gut mit dem Programm FB-WinTranslator oder FB-AciveTranslator auskennen und sehr sicher die Rechtschreibung beherrschen![2]

At the same time, potential authors are encouraged to provide the publisher with rights to hitherto unknown technologies, which is commonplace in the Anglo-Saxon world but blatantly flaunts German copyright law:

> Dieses Recht zur Veröffentlichung im Internet umfaßt auch das Recht zur Veröffentlichung in anderen lokalen oder globalen Netzen, in Online-Diensten oder über netzbasierte Push-Technologien einschließlich Mailing-Listen oder anderen zukünftigen Technologien der Netzkommunikation.[3]

3 Conclusion

In the face of the many threats to content quality described above, do the established dictionary publishers have a moral obligation to provide some sort of service on the Web free of charge? And what implications does this thought have for future investments on the part of these publishers?

A spokesman for Merriam-Webster has stated that the presence of their complete Collegiate dictionary on the Internet has not had a negative influence on sales of this work but has finally given the company the opportunity to find out exactly what people actually look up, but not all dictionary publishers are convinced that this is the right way forward. In the late 1980s Langenscheidt had a medium-sized bilingual English-German dictionary "smuggled" onto the Internet by a student who had been given the dictionary data for research purposes. When Langenscheidt finally found out about it (almost four months later), an analysis of the data provided showed that - surprise, surprise – the F-word was by far the most popular word looked up, especially by the academic community, which in those days had almost exclusive access to the Net. It was clear that – at least at that point – the resource was not considered a serious linguistic tool, but in many cases used for pleasure or to while away the time. Had access to the dictionary been liable to a fee, this sort of pubertile use would almost certainly not have occurred.

In an early survey of dictionary resources on the Web, [Carr 1997] maintains "Publishers must provide easier access to superior databases if they are to compete with free Internet reference works"[4]. But if high-quality content is to be made available free of charge, who is going to finance this – and ensure that the quality and, more importantly, topicality of the material is guaranteed? Britannica's endeavours to use advertising to solve this problem rather than charge a subscription fee is a model that still has to prove its worth in the real world. And, of course, there is always the underlying risk that this sort of "free offer" will create a sense of expectancy among users that they should get "something for nothing" automatically, which devalues the very contents that are the company's most valuable asset.

Given the anarchic nature of the many alternatives freely available on the Net, it would seem dangerous to leave the development of linguistic resources in the hands of the lexicographically unskilled. In many of these cases, the contents available are in effect worse than nothing (belying Dr Johnson's statement quoted in the Abstract of this paper). It would seem appropriate, then, that serious dictionary producers should seek to find some common ground in their approach to their presence on the Web. Only by forming a concerted front can the traditional publishing houses hope to establish some form of "netiquette" that would apply to the linguistic resources on which so much communication and, ultimately, the future of these companies could depend.

Notes

[1]http://www.brall.com/quickdic/index.html

[2]http://www.brall.de/autorengesucht/autorengesucht.html

[3]http://netlexikon.akademie.de/query; q:Lexikonprojekt%3A%20Autorenvereinbarung

[4]Cf. Carr, 1997, p.211

References

Anon. (1999). Traffic für die Börse? in *Buchreport* No. 43, Oct. 27 1999, p.6.

Anon. (1999). Hacker Journalism, in *The Economist*, Dec. 4 1999, p. 82.

Carr, Michael (1997). Internet Dictionaries and Lexicography, in *International Journal of Lexicography*, Vol 10, pp. 209-230.

Gschwender, Oliver (1999). *Internet für Philologen. Eine Einführung in das Netz der Netze*. Erich Schmidt Verlag, Bielefeld.

Parks, Robert (1999). The Participatory Lexipedia Project: A Collaborative Lexical Database, www.wordsmyth.net/doc-lexipedia.shtml

Piozzi, Hester Lynch (1925). *Anecdotes Of The Late Samuel Johnson*. BCL 1 PR English Literature, No 535.

Tihanyi, Lázló (1999). MoBiGloss: A Virtual Dictionary System on the Internet, in *Papers in Computational Lexicography*, edited by Ferenc Kiefer et al., COMPLEX, Budapest..

Internet Links

The following links provide overviews of the many dictionary resources available on the Net:

www.dictionary.com

www.facstaff.bucknell.eu/rbeard/diction.html

www.informatik.tu.muenchen.de/ğendreyz/dictionaries-d.html

www.laixicon.com/cgi-bin/r.cgi/Static/de/dictionaries/one

www.polyglot.lss.wisc.edu/lss/lang/langlink.html

www.yourdictionary.com