MICHAEL ZOCK
JEAN–PIERRE FOURNIER
LIMSI – CNRS, Orsay, France
E-Mail: zock/fournier@limsi.fr

Proposal for a customizable, psycholinguistically motivated dictionary to enhance word access

1. Introduction

If the fundamental role of a dictionary is obvious for natural language processing, it is less evident according to what principles it should be built. Regardless of the *medium* (paper, computer memory, human brain) and the task, a good dictionary contains *many entries*, a *lot of information* for each one of them, and the relevant information is easily *accessible*. In other words, what counts above all is '*what* is in the dictionary?' and '*how* can the *relevant information* be accessed?'.

We will consider here dictionaries mainly from the language producer's and language learner's point of view, two aspects which are generally ignored in the literature. The problems arising in this context are either on the *dictionary* side, — lack of information (gaps), or overabundance: the user is drowned in information (ease of access), — or on the *user's* side: the person is unable to access the needed information either because it isn't there (ignorance), or because of perfor–mance errors (slips of the tongue : blends, anticipations, etc.). It is this latter case we will focus on, and it is precisely here that electronic dictionaries, enhanced with search facilities, can be of great help.

2. The content of a dictionary

A good dictionary can answer many questions asked by the language user. Here is some of the information a user might want to look for

(a) *Correspondences* between words, or group of words (translation);

(b) *Meaning* (definition), or *domain* of a word (painting => *art* vs. *construction work*);

(c) Meaning *relations* between words (synonyms, antonyms, hyp(er)onyms, etc.);

(d) Lexical category (part of speech: noun, verb, adjective, adverb, etc.)

(e) Grammatical information (e.g. the *gender* of nouns, type of *preposition* and *auxiliary*) ;

(f) *Derivational morphology* (nominalization: sell => selling)

(g) Collocations (e.g. the kind of *verb* required by a given *noun*).

(h) Examples (contexts in which the word is used);

(i) False friends, i.e. known confusions between words (example: the difference between *puisque* (since) and *parce que* (because) in French)

In addition, a user might want to use the dictionary for specific purposes (accessing and learning words). In this case s/he may want *support tools* allowing to *annotate* words with relevant information, or to build exercises for *memorizing* them. The former would allow to enter access keys, i.e. associations (*rose* = flower, gift, red), while the latter could be used to build word lists, which, coupled with a parametrizable flash card system, would allow for word drills.

3. The problem of word access

No matter how rich a database may be, it is of little use if one cannot access the relevant information in time. *Access* is probably THE major problem that we experience when we are trying to produce language in spoken or written form. As we shall see, this is precisely a point where computers can be of considerable help.

Work on memory has shown (Baddeley, 1982) that access depends crucially on the way informa–tion is organized, yet the latter can vary to a great extent. If electronic dictionaries compare favorably to paper dictionaries (weight, ease and speed of access), their potential is still not fully used. Nearly all dictionaries are tailored for a single user. Yet the *needs* of different users, their way to *organize* and to access information vary considerably (think of a tourist, a lawyer, a university professor or a language student). Just as everyone organizes the universe according to their own needs or point of view (e.g. a department store, a library, a house), everybody structures the lexicon according to their own perception of the world. Hence the idea to provide an editor that allows people to build their own dictionary by extending an existing one with comments. Yet one can do more. Actually, the usefulness of electronic dictionaries can be increased considerably by adding efficient search facilities. More here below.

As speech error literature has taught us (Fromkin 1973, 1993; Cutler, 1982) ease of access depends not only on *meaning relations* (word bridges, i.e. associations) or the *structure* of the lexicon, i.e. the way words are *organized* in our mind (see the work on semantic memory Collins & Quillian, 1969; Smith et al. 1974), but also on linguistic form. Researchers collecting speech errors have offered countless examples of phonological errors in which segments (phonemes, syllables or words) are added, deleted, anticipated or exchanged. Reversals like /aminal/ instead of /animal/, or /carpsihord/ instead of /harpsichord/ are not random at all, they are highly systematic and can be explained. Examples like the one below (Fromkin 1973) clearly show that knowing the *meaning* of a word does not guarantee its *access*.

Anticipations take my bike bake my bike

A tank of gas a gas of tank

Installing telephones intelephoning stalls

Perseverations pulled a tantrum pulled a pantrum

Reversals Katz and Fodor Fats and Kodor

Blends grizzly + ghastly grastly

Haplologies Post Toasties Posties

Misderivations an intervening mode an intervenient mode

Word substitutions before the place opens before the place closes

The work on speech errors also reveals that words are *stored* in two modes, by *meaning* and by *form (sound)*, and it is often this latter which inhibits finding the right token: having recombined inadvertently the components of a given word (syllables), one may end up producing a *word*, which either does not exist or is simply different from the one in mind.

This kind of *recombination,* resulting from bookkeeping problems due to time pressure, parallel processing and information overload, may disturb or prevent the access of words. Hence the usefulness of a tool that allows to revert the process. In order to allow for doing so, it is necessary to represent words not only in terms of their meaning, but also in terms of their written and spoken form. The fact that words are indexed both by *meaning* and by *sound* can now be used to our advantage. Suppose you presented the system an unknown word, or a word that does not exactly express the intended meaning. The fact that words are coded *phonetically* allows the recombination of their segments (syllables), hence the presentation of new candidates, among which the user should find the one s/he is looking for. The fact that words are coded *semantically* allows keeping the number of candidates to be presented small. The list of potential candidates will be filtered according to semantic criteria (domain). Hence the phoneme /vin/ would yield *vin*, *vainc*, *vingt* or *vint* depending on whether the domain were "food (wine), battles (to win), digits (twenty), or movement (to come)".

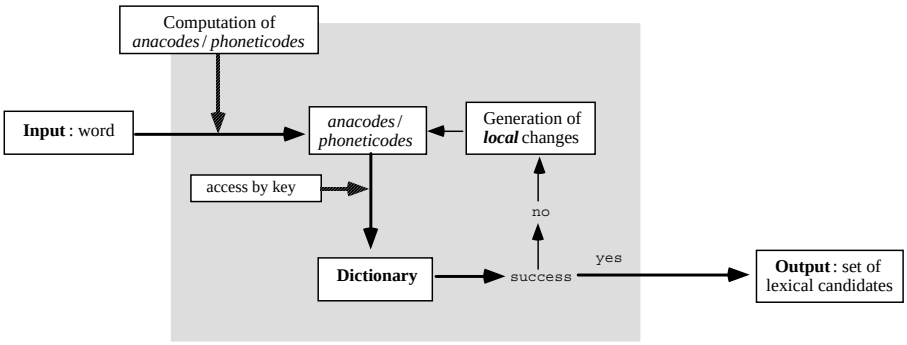
As one can see, if phonological permutations may inhibit the access of a given word, they may also enhance its access: used deliberately and in a controlled way, they may allow the system to override illegal combinations: presenting the user with other alternatives, among which s/he may find the one s/he was looking for. Strange as it may seem, this kind of technique has been used for quite some time in the area of spell checking.

4. Old wine in new bottles: the re–use of *spell–checking techniques* for enhancing *word access*

The component described here below is part of two larger systems PIC (Fournier & Letellier, 1990) and ECLAIR (Letellier), both of which are part of CARAMEL (Fournier et al. 1990). The system has two basic mechanisms for correcting spelling errors: *anacodes* and *phoneticodes*. The former computes an access key for finding the right word. Since an anacode is equivalent to the set of letters composing the word (for example, the anacodes of "calibrer" and "aclibrre" are identical), erroneous *order* of letters is a non–issue. The system would still find the right word, provided that there is such a candidate in the dictionary, and provided that the user didn't omit, add or replace some character(s) with other characters. For example, if the input were aclibrer instead of calibrer, the system would have no difficulty to find the target word (calibrer), since both words are composed of the same set of letters. If the user added letters outside of the anacode, the system would need several runs to check alternative spellings by making local variations (delete or add a character by making systematic permutations). Unlike most systems, this technique allows to deal with spelling errors occurring at the beginning of a word.

The second technique (phoneticodes) consists in converting *graphemes* into *phonemes*, which allows the system to deal with spelling errors due to homophony (two different words having the same spelling), a very frequent phenomenon in French. For example, the system would be able to deal with errors like hippotenuse instead of hypoténuse. Put differently, if the user presented the word

hippotenuse, while he was trying to say hypoténuse, the system, rather than remaining silent, would present the word he was looking for, namely hypoténuse. If the system cannot find directly a candidate, it will perform local changes by performing permutations of phonemes or syllables. Hence it would have no problem to find the word "poteau" (pole) instead of "topo" (topic), both words being composed of the same syllables (/po–to/), the only difference being their order.



The situation is more complex and may even become intractable if extraneous material is added, or if the correction yields an existing word, yet different in terms of meaning from what was intended. Suppose that the target word were "maison" (house), while the user typed /masson/. Relying on the phoneticode, the system might suggest "maçon" (bricklayer), a word that exists, but which is not at all what was intended.

5. Conclusion

We have drawn the readers' attention to the importance of *access* in the context of electronic dictionaries : information must not only be *available*, it must also be *accessible* . Looking at some of the psycholinguistic findings, and looking at the work done on spell checking it seemed that some of the techniques developed in the context of the latter could profitably be used in the domain of the former. While the use of certain spell checking techniques can probably enhance word access, hence the potential of electronic dictionaries, more work is needed in order to keep the search space small.

6. References

Baddeley, A. (1982) Your memory: A user's guide. Penguin.

Collins, A. & Quillian, L (1969) Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 240–247.

Cutler, A. 1982. (Ed) Slips of the Tongue and Language Production. Amsterdam: Mouton.

Fournier J–P., G. Sabah & C. Sauvage–Wintergerst. (1990) A Parallel Architecture for Natural Language Understanding Systems. Pacific Rim International Conference on Artificial Intelligence, Nagoya (Japan), 1990.

Fournier J–P. & S. Letellier. (1990) PIC: a Parallel Intelligent Corrector. Artificial Intelligence application & Neural Networks AINN'90, pp 38–41, Zürich.

Fournier J–P. (1986). Correction automatisée dans les systèmes question–réponse en langage naturel. Actes du 2ème Colloque International d'Intelligence Artificielle, CIIAM'86, pp 659–679, Marseille.

Fromkin, V. (1973) (Ed.) Speech errors as linguistic evidence. The Hague: Mouton Publishers.

Fromkin, V. (1993) Speech Production. In *Psycholinguistics* edited by J. Berko–Gleason & N. Bernstein Ratner. Fort Worth, TX: Harcourt, Brace, Jovanovich.

Smith, E., Shoben, E. & Rips, L. (1974) Structure and process in semantic memory: a featural model for semantic decisions. Psychological Review, 81, 214–241.