If You Care to Find What You Are Looking For, Make an Index: the Case of Lexical Access

Michael Zock, Non-member

ABSTRACT

Obviously, words play a major role in language production, hence finding them is of vital importance, be it for writing or for speaking (spontaneous discourse production, simultaneous translation). Words are stored in a dictionary, and the general belief holds, the more entries the better. Yet, to be truly useful the resource should contain not only many entries and a lot of information concerning each one of them, but also adequate navigational means to reveal the stored in-formation. Information access depends crucially on the organization of the data (words) and the access keys (meaning/form), two factors largely overlooked. We will present here some ideas of how an existing elec-tronic dictionary could be enhanced to support a speaker/writer to find the word s/he is looking for. To this end we suggest to add to an existing electronic dictionary an index based on the notion of association, i.e. words co-occuring in a well balanced corpus, the latter being supposed to represent the average citizen's knowledge of the world. Before describing our ap-proach, we will briefly take a critical look at the work being done by colleagues working on automatic, spon-taneous or deliberate language production, -that is, *computer-generated language*, simulation of the *men-tal lexicon*, or *WordNet* (*WN*),- to see how (in)adequate they are with regard to our goal

Keywords: lexical access, index based on associations

1. INTRODUCTION

We spend a large amount of our lifetime searching : ideas, names, documents, and "you just name it". I will be concerned here with the problem of words, or rather, how to *find* them (word access) in the place where they are stored: the brain, or an external resource, a dictionary.

Obviously, a good dictionary is a well-structured repository with a lot of information concerning words. Yet, what counts is not only the coverage, i.e. number of entries or the quality of the information associated with it, but also access support. Because, what is in-formation good for, if one cannot access it when needed?

I will present here some ideas of how to enhance an existing electronic dictionary, in order to help the user to find the word he is looking for. Before doing so I will take a look at various solutions offered for different production modes, spontaneous, deliberate and automatic language production, to see their qualities and shortcomings. Let me start with the latter.

2. RELATED WORK IN THE AREA OF NATURAL-LANGUAGE GENERATION

A lot of work has been devoted to lexical issues during the last fifteen years. For excellent surveys see [23, 27, 30] or [8] for some earlier work. Two ap-proaches that have been particularly successful were *discrimination nets* [13](figure 1) and *graph-rewriting*, i.e. *pattern-matching* [21, 3] (figures 2a and 2b). The former can be seen as a hierarchically ordered set of tests whose outcome determine the word to be chosen. Since the tests are hierarchically ordered, we have, formally speaking, a tree, whose nodes are the condi-tions (tests) and the leaves the outcome, i.e. words.

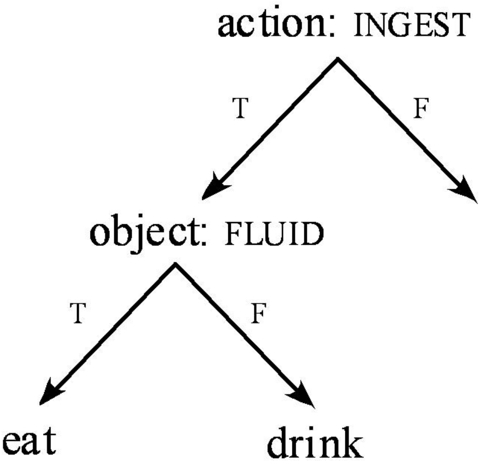


Fig.1: *Discrimination nets, or, the check list approach.*

Yet, words are meant to express meanings, which, taken together form messages. And since the input to the lexicalization component are messages, i.e. mean-

Manuscript received on September 14, 2006.

The author is with Laboratoire d'Informatique Fondamentale (LIF), CNRS, UMR 6166 Parc scientifique et technologique de Luminy, Case 901 - 163 Avenue de Luminy F-13288 Marseille Cedex 9; E-mail: michael.zock@lif.univ-mrs.fr

ings words are supposed to express, it is natural to represent both of them, messages and the words' underlying meaning, by the same formalism, for example, semantic networks or conceptual graphs. According to this view, lexicalization amounts to pattern-matching. Lexical items are selected, provided that their underly-ing content covers parts of the conceptual input (fig-ures 2a and 2b). This being so, the goal is to find suffi-cient, mutually compatible lexical items so as to com-pletely cover the input with minimal unwanted addi-tional information. Unfortunately there are several shortcomings with these two approaches:

- concerning the checklist approach: Apart from the fact that discrimination nets have never been devel-oped at a large scale (i.e. for a subset of a lexicon), it remains to be seen whether this technique is well suited for all types of lexical items. Also, the sum of the information given during the tests does not amount to a full specification of the meaning of the word towards which the tests converge, even if the underlying message is taken into account.

- concerning the pattern matching approach: this ap-proach hinges on the assumption that the message is completely planned in all its details prior to verbalization, which, of course, is hardly ever the case. Yet, what shall we do in case of incomplete conceptual input (figure 2b)?

Of course, one could claim, as I've done elsewhere [32], that the input, i.e. message to be expressed, is incomplete prior to lexicalization, and the role of the lexicon is not only to express the message, but also to help refining its underspecified parts. The speaker (or writer) starts with a skeleton plan (gist, or rough out-line), which he fleshes out with details little by little. For example, instead of saying "x meets y", he pro-vides further information concerning the referents x and y, to produce "(x: The young woman) met (y: her husband)"

It is interesting to note, that in none of these works the issue of word access is addressed at all. As a matter of fact, from a strict computational linguistic point of view, the whole matter may be a non-issue, and as such it is natural that it would not appear neither in Ward's list of problems to be addressed [31], nor in Cahill & Reape's paper 'Lexicalisation in applied NLG systems' [6]. How-ever, if we address the problem of lexicalisation from a psycholinguistic or man-machine interaction point of view (spontaneous discourse or writing on a computer), things are quite different. There is definitely more to lexicalisation than just choosing words: one has to find them to begin with. No mat-ter how rich a lexical 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 are confronted with when trying to produce lan-guage, be it in real-time (oral form) or consecutive mode (written form). As we shall see, this is pre-cisely a point where computers can be of consider-able help. Before doing so, let's take a look at what psychologists have to say.

3. RELATED WORK IN PSYCHOLOGY AND PSY-CHOLINGUISTICS

There is an enormous amount of research in psycholinguistics regarding this issue: a collection of papers edited by [18] and [16], several monographs [28, 1], and an overwhelming amount of empirical studies, to begin with Brown & Mc Neill's land-mark work on the tip-of-the tongue phenomenon [4], but also [15, 24] to name just those. While all these papers take up the issue, they do not consider the use of computers for helping people in their task. Yet this is precisely a point I am particularly interested in. Still, the work being done by psy-chologists and the results obtained are truly impres-sive and very important.

The dominant psycholinguistic theories of word production are all *activation-based, multilayered network models.* Most of them are implemented, and their focus lies on modelling human perform-ance: speech errors or the time course (latencies) as observed during the access of the mental lexicon. The two best-known models are those of Dell [10] and Levelt [17], which take opposing views con-cerning *conceptual input* (conceptual primitives vs. holistic lexicalized concepts) and *activation flow* (one-directional vs. bi-directional).

The Dell model (figure 3) is an interactive-activation-based theory that, starting from a set of features, generates a string of phonemes. Information flow is bi-directional, that is, lower level units can feed back to higher-level components, which may lead to errors. For example, the system might produce *rat* instead of the intended *cat*. Indeed, both words share certain components. Hence, both of them are prone to be activated. At the conceptual level (from the top) they share the feature animal, while at the phonological level (from the bottom) they share two phonemes. When the word node for *cat* is active, any of the following segments /k/, /ae/, and /t/ is co-acvtivated. The latter two phonemes may feed back, leading to *rat*, which may already be primed and be above baseline due to some in-formation coming from a higher-level component. The model can account for various other kinds of speech errors like preservations (e.g., beef needle soup), anticipations (e.g., cuff of coffee), etc.

Based on the distribution of word errors, Dell argues that some aspects of speech generation rely on retrieval (phrases, phonological features, etc), while many others (word/phoneme and possibly morpheme combinations) rely on synthesis. Since generation is a productive task, it is prone to swap-ping or reuse of elements.

WEAVER++ (Word Encoding by Activation and VERification) is also a computational model. It has

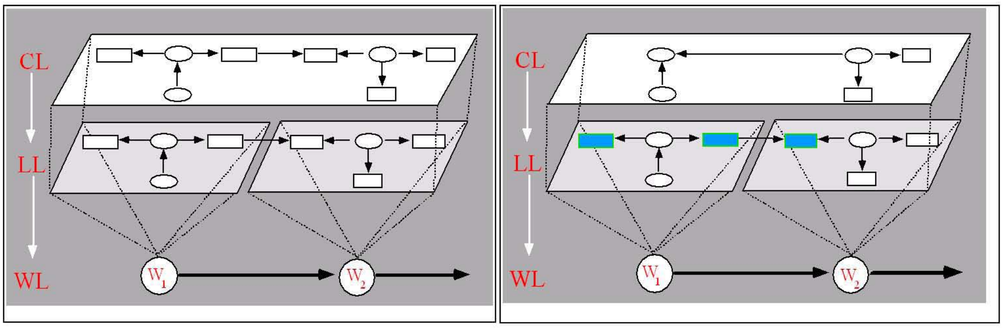


Fig.2: *(2a) Fully specified conceptual input and (2b) Partially specified conceptual input.*

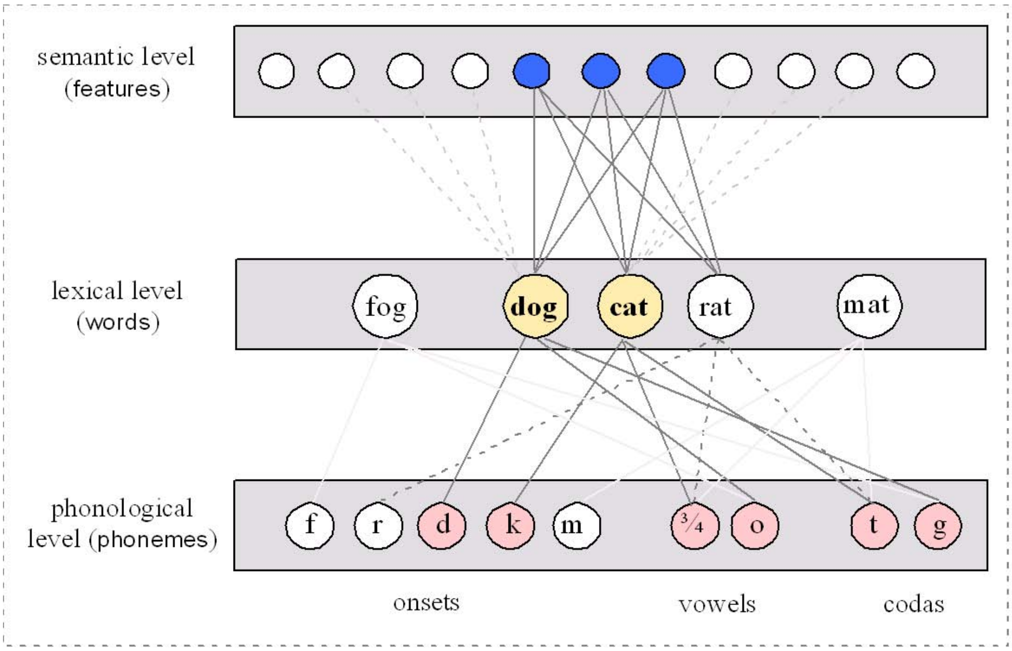


Fig.3: *Dell's interactive model.*

been designed to explain how speakers plan and control the production of spoken words [17]. The model is "hybrid" as it combines a *declarative associative network* and *procedural* rules with *spreading activation* and *activation-based rule triggering*. Words are synthesized by weaving together various kinds of information.

While WEAVER++ is also activation-based, informa-tion flow is only one-directional, top-down. Processing is staged in a strictly feed-forward fashion. Starting from lexicalized concepts (concepts for which a lan-guage has words) it proceeds sequentially to lemma selection, morphological, phonological and phonetic encoding, to finish off with a motor plan, necessary for articulation (see figures 4 and 5). Unlike the previous model, WEAVER++ accounts primarily for reaction time data. Actually, it was developed on the basis of such data collected during the task of picture naming. How-ever, more recently the program managed to parallel a number of findings obtained in psycholinguistics where other techniques (chronometry, eye tracking, electrophysiological and neuro-imaging) have been used.

•Apart from work on the time course of lexical ac-cess, there is a large body of work on memory and speech errors, providing the foundations for the above described models. Work on memory has shown that access depends crucially on the way in-formation is organized [7, 26, 2]. From speech error literature [12, 9] we learn, that ease of access de-pends not only on ***meaning relations***, - (word bridges, i.e. associations) or the structure of the lexi-con, i.e. the way words are *organized* in our mind, - but also on **linguistic form** (similarity at the dif-ferent levels). Researchers collecting speech errors have offered countless examples of phonological er-rors 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 ex-plained. Examples like the one below [12] clearly show that knowing

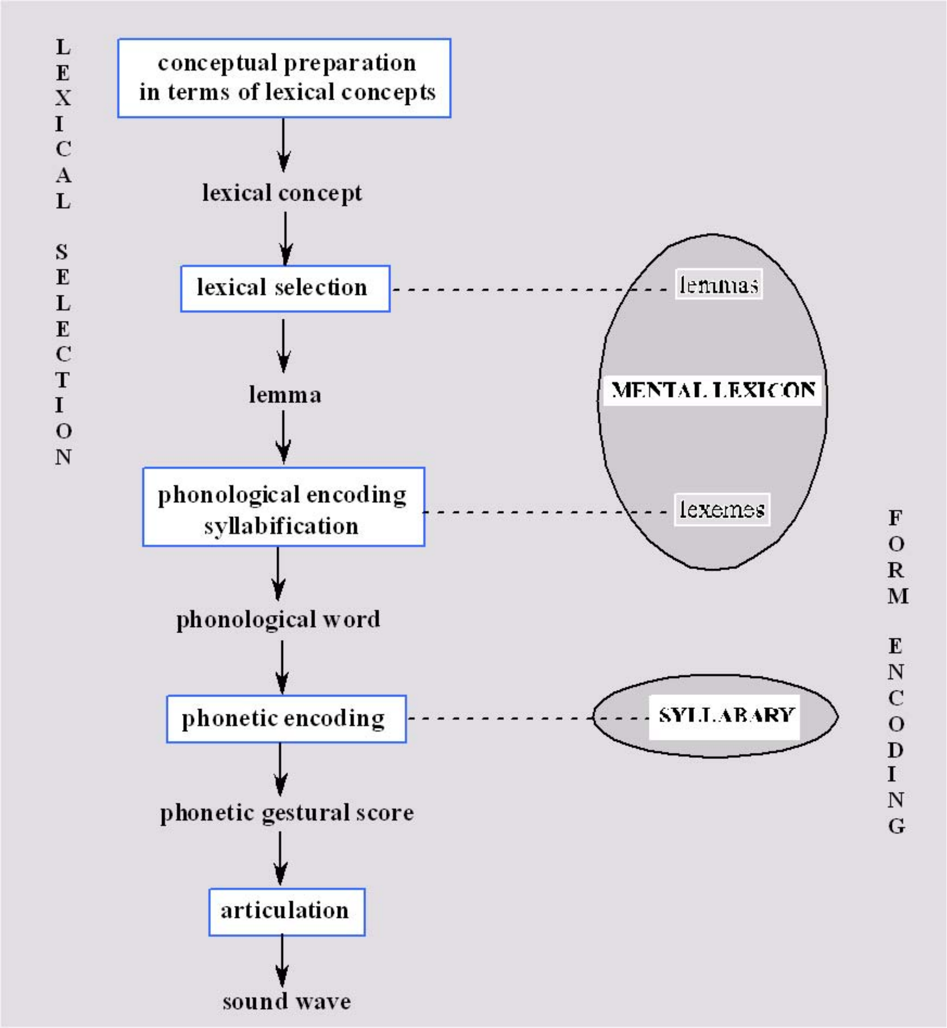


Fig.4: *Stages in the production of words.*

the *meaning* of a word does not guarantee its *access* (table 1).

Table 1: *Various kinds of speech errors*

Error type	intended output	produced output
Anticipations	take my bike	bake my bike
Preservations	pulled a tantrum	pulled a pantrum
Reversals	Katz and Fodor	Fats and Kodor
Misderivations	an intervening mode	an intervenient mode
Word substitutions	before the place opens	before the place closes
Blends	grisly + ghastly	grastly

While all the work discussed so far started from a conceptual input, let's take a look at a tool, designed to contact (enter) the dictionary and to navigate in it by using words. Actually, this kind of information retrieval or access is the one we are most familiar with. Yet, WordNet (WN) the resource we will discuss in the next section is quite different from conventional dictionaries, and as such, it is a great step forward in the right direction: rather than multiplying the number of dictionaries (one for each use or link: definition, synonyms, antonyms, etc.), WN has been built as a single resource (a database) allowing for multiple accesses by following different links [20, 11].

4. RELATED WORK IN THE AREA OF ELECTRONIC DICTIONARIES: FROM WORD TO WORD

Despite the fact that there are many lexical resources available in electronic form (http://www.ilc.cnr.it/EAGLES96/rep2/node1.html), I will discuss here only one, WordNet. WN has been built on the basis of psy-chological mechanisms and organization principles like association, hierarchies, and semantic fields...

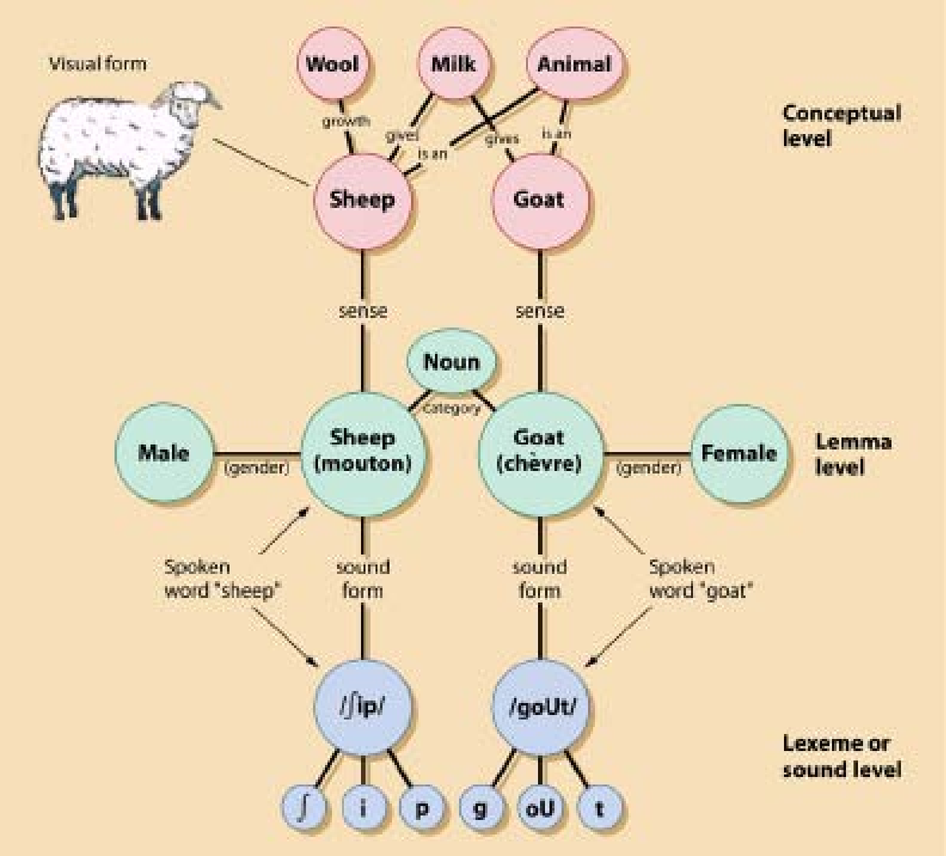


Fig.5: *Fragment of lexical network.*

The way information is structured is quite different from conventional dictionaries. Lexical entries are organized around linked synonym sets. There are basically two kinds of links: lexical and semantic. The former hold between word forms, whereas the latter connect word meanings. Set inclusion, i.e. hypernymy/ hyponymy (general/specific), antonymy (opposite), entailment, and meronymy/ holonymy (part of) are typical links. Different parts of speech are organized differently. For example, nouns and verbs are organ-ized into hierarchies, whereas adjectives are arranged in clusters, each cluster being organized around an-tonymous pairs. Since adverbs are often derived from adjectives, they may have antonyms.

While there is no doubt that WN is a major step in the right direction, it is not perfect, and its authors are very well aware of it. Let me mention just some of its shortcomings.

(1) The '*tennis-problem*': words typically occurring together, hence naturally associated (tennis, umpire, racket, court, backhand), are not linked in WN; (2) *the lack of syntagmatic relations*: "WordNet provides a good account of paradigmatic associations, but con-tains very few syntagmatic links.".... If we knew how to add to each noun a distinctive representation of the contexts in which it is used... WordNet would be much more useful." (Miller, in [11: 33-34]. One can't but agree more. For a proposal going in this direction see [34]. (3) *Incompleteness of links.* For a given synset there is no link between its elements apart from the synonym link. Yet, each element might trigger a dif-ferent association. Take for example 'rubbish' and 'garbage'. While the former may remind you of a 'rabbit' or (horse)-'radish', the latter may evoke the word 'cabbage'. (4) *The problem of meaning.* WN's underlying structure is a lexical matrix whose columns and rows are respectively *meanings* and *words.* The idea sounds perfect, as it seems to model the two major access- or communication modes (meaning/forms, i.e. comprehension/expression). Unfortunately, the column reserved for meanings is not fully functional, i.e. usable by a human user, as meanings are equated with synsets rather than semantic primes (atomic meaning elements). WN expects words rather than its atomic meaning elements as input. Yet this is somehow unrea-sonable, both for producing language in your mother tongue, and even more so when speaking a foreign language.

5. DISCUSSION

I have presented and commented on three approaches dealing with the problem of the lexicon.

One would expect complementarities in the quest of achieving a unified view, yet this is far from obvious. The goals and the methods being simply too different. All of them capture something relevant, but none of them gives us a unified view.

Concerning the work done in the domain of "natural language generation", next to nothing can be used in the context of electronic dictionaries: the issue of word access simply does not arise. The assumption being that what is stored in the machine can naturally be accessed. In addition, most of this work is based on very small dictionaries, tailored to the engineer's specific needs, and the issue of macrostructure (organization) is not addressed at all.

As for the work carried out by psychologists, there are several problems: (a) the size of their dictionaries is extremely small (about 100 entries); (b) the specificities of the macrostructure are not spelled out at all; (c) the models being connectionist, the links cannot be interpreted by human beings: all we get are weighted links; (d) the notion of lemma is problematic as in computational linguistics, a lemma is a concrete form for a given meaning (let say "walk", in order to ex-press some kind of movement), we are nearly empty handed in the case of the mental lexicon. A *lemma* in this framework means nothing more than a semantic-syntactic specification (part of speech, and a set of features), but nothing coming close to a concrete word form, as this is being taken care of by the phonological component, which determines the lexeme.

By looking at this work one gets the impression that people don't have words at all in their mind. No-tions like "words, dictionary, memory" etc. are but metaphors. What we seem to have in our brains is a set of highly abstract information, distributed all over. By propagating energy rather than data or information (as there is no message passing, transformation or accumulation of information, there is only activation spreading, that is, changes of energy levels, call it weights, electronic impulses, or whatever), we propagate signals, activating ulti-mately certain peripherical organs (larynx, tongue, mouth, lips) in such a way as to produce sounds, that, not knowing better, we call words. Another way of putting things is to say that our mind is a word fabric rather than a storehouse, words being synthesized rather than retrieved.

Yet, we are concerned here with *word access.* In this respect, WN has the best potential among the presented candidates. Even though it does not have the *power* or *flexibility* of a mental lexicon- for one it lacks too many of the links known to exist in our mind (see all the work done on "word association"), and secondly, the links are not quantified and context-sensitive. - it could be improved in such a way as to get close to our ideal lexicon. I will show in the remainder of this paper a line of research I am pursing in order to remedy some of the shortcomings mentioned here above. The guide-lines of this work are the natural conditions and practi-cal needs of a speaker or writer looking for a word. Before doing so, let us take a look at the speaker's goals and knowledge at the onset of initiating search.

6. WORD ACCESS ON THE BASIS OF ASSOCIATIONS

There are at least two things that people usually know before opening a dictionary : the word's **meaning**, or at least part of it (i.e. part of the *definition*) and its relation to other words or concepts: x is *more general* than y, x is the *equivalent* of y, x is the *opposite* of y (in other words, x being the hypernyme/synonyme or antonym of y), etc. where x could be the *source word* (the one coming to one's mind) and y the *target word* (the word one is looking for). This is basically concep-tual knowledge. Yet, people seem also to know a lot of things concerning the lexical **form** (lexeme): number of *syllables*, beginning/ending of the target word, its *part* of *speech* (noun, verb, adjective, etc.), and some-times even the *gender* [4, 5, 29]. While, in principal all this information could be used to constrain the search search space, hence, the ideal would be multiple in-dexes, I will deal here only with the conceptual part (meaning, i.e. partial definition, and the words' rela-tions to other concepts or words).

The yet to-be-built (or to-be-enhanced) resource is based on the age-old notion of association: every idea, *concept* or *word* is connected. In other words, I assume that people have a highly connected conceptual-lexical network in their mind. Finding a word amounts thus to entering the network at any point by giving the word or concept coming to their mind (*source word*) and to follow then the links (associations) leading to the word they are looking for (*target word*). In other words, look-up amounts to navigation in a huge lexical-conceptual space and is not necessarily a one-shot process.

Suppose, you were looking for a word expressing the following ideas: *superior dark coffee made from beans from Arabia*, and that you knew that the target word was neither *espresso* nor *cappuccino.* While none of this would lead you directly to the intended word, mocha, the information at hand, i.e. the word's definition or some of its elements, could certainly be used. In addition, people draw on knowledge concerning the role a concept (or word) plays in language and in real world, i.e. the associations it evokes. For exam-ple, they may know that they are looking for a noun standing for a beverage that people take under certain circumstances, that the liquid has certain properties, etc. In sum, people have in their mind an encyclopedia: all words, concepts or ideas being highly connected. Hence, any one of them has the potential to evoke the others. The likelihood for this to happen depends, of course, on factors such as *frequency* (associative strength), *distance* (direct vs. indirect access), *prominence* (saliency), etc.

How is this supposed to work for a dictionary user? Suppose you wanted to find some word (target word: t_w), yet the only token coming to your mind were a somehow related word (source word: s_w). Starting from this input the system would build internally a graph with the s_w at the center and all the words con-nected to it at the periphery. The graph would be built dynamically depending on the demand. If the list con-tains the t_w, search stops, otherwise navigation contin-ues, taking either one of the proposed candidates as the new starting point or a completely new token.

Let's take an example. Suppose you were looking for the word ***mocha*** (t_w), yet the only token coming to your mind were ***computer*** (s_w). Taking this latter as starting point, the system would show all the con-nected words, for example, Java, *Perl, Prolog* (pro-graming languages), *mouse, printer* (*hardware*), *Mac, PC* (*type of machines*), etc. querying the user to decide on the direction of search by choosing one of these words. After all, he knows best which of them comes closest to the t_w. Having started from the s_w *computer*, and knowing that the t_w is neither some *kind of soft-ware* nor a *type of computer*, he would probably choose *Java*, which is not only a *programming language* but also an *island*. Taking this latter as the new starting point he might choose *coffee* (since he is look-ing for some kind of beverage, possibly made from an ingredient produced in Java, coffee), and finally *mocha*, a type of beverage made from these beans. Of course, the word *Java* might just as well trigger *Kawa* which not only rhymes with the s_w , but also evokes Kawa *Igen*, a javanese volcano, or the argotic word of coffee in French.

As one can see, this approach allows word access via multiple routes (there are many ways leading to Rome). In addition, it takes very few steps to make quite substantial leaps, finding a link (or way) between apparently completely unrelated terms. In sum, this is approach is both fast and flexible, at least way more flexible than navigation in a conceptual tree (type hierarchy, ontology) where terms are organized via ISA links, that is hierarchically. In this latter case, navigational mistakes can only be repaired via back-tracking.

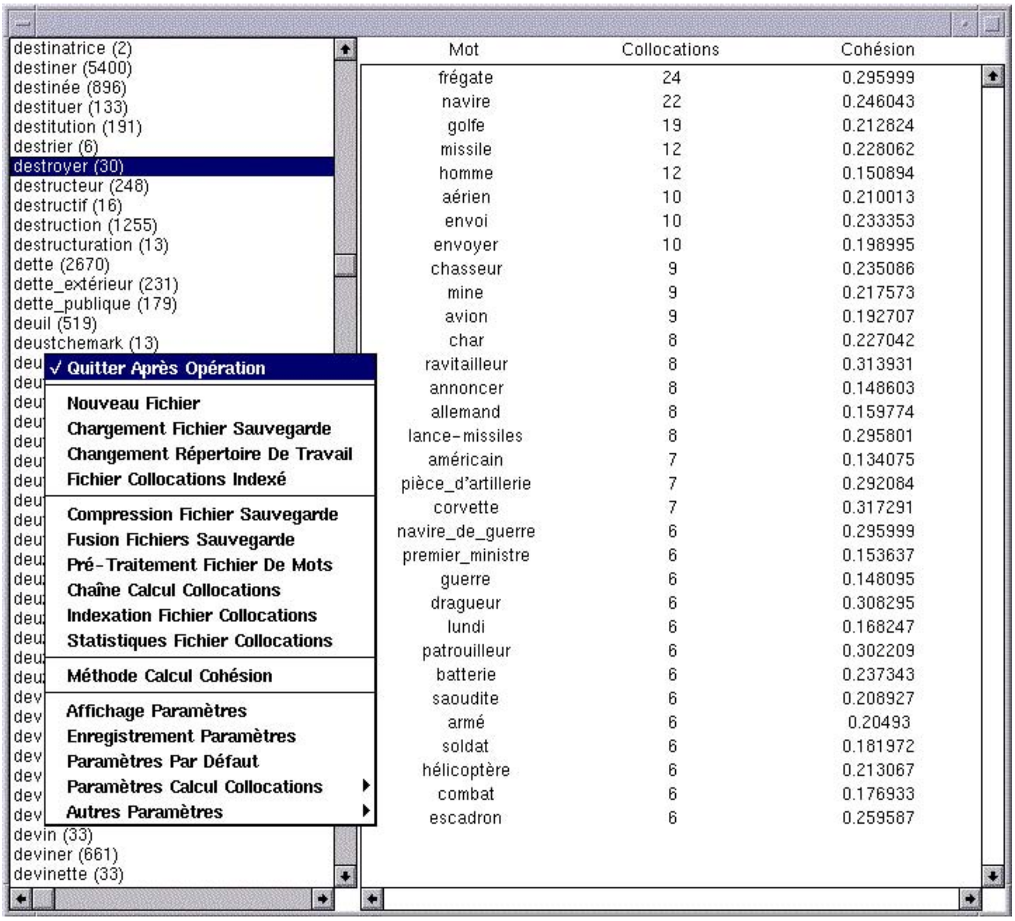
Of course, one could also have several associations (quasi) simultaneously, e.g., 'black, delicious, strong, coffee, beverage, cappuccino, espresso, Vienna, Starbucks, espresso...' in which case the system would build a graph representing the intersection of the asso-ciations (at distance 1) of the mentioned words.

Obviously, the greater the number of words entered and associated to a sw, the more complex the graph will be. As graphs tend to become complex, they are not optimal for navigation. There are at least two factors impeding readability: *high connectivity* (great number of links or associations emanating from each word), and *distribution* (conceptually related nodes, that is, nodes activated by the same kind of association, do not necessarily occur next to each other, which is quite confusing for the user). This being so, I suggest to display by category (chunks) all the words linked to the source word. Hence, rather than displaying all the connected words as a huge flat list, I suggest to present the words in hierarchically organized clusters, the links of the graph, becoming the nodes of the tree. This kind of presentation seems clearer and less overwhelming for the user, allowing for categorical search, which is a lot faster than search in a huge bag of words, provided that the user knows which category a word belongs to.

7. DISCUSSION AND CONCLUSION

Obviously, in order to allow for this kind of access, the resource has to be built accordingly. To this end at least four problems have to be solved. (1) A corpus must be chosen. Since the corpus is supposed to repre-sent the user's world knowledge, it must reflect in the corpus. In other word, the corpus must contain a little bit of everything: some politics, some sports, some geography, etc. (2) Words have to be indexed in terms of the associations they evoke. This means that we need to discover which word evokes which other words, i.e. association. This can be done via a colloca-tion extractor (see figure 6, Ferret's collocation extrac-tor [35]). (3) The weight of the link needs to be deter-mined (relative frequency). This is important for rank-ing the associated words. Ideally, the weight is (re-)computed on the fly to take into account contextual variations. A given word (Java) is likely to evoke quite different associations depending on the context (*coffee* vs. *programming*). (4) Associations must not only be identified, but also be labeled. This is vital for naviga-tion. Typing the links is the hardest task, yet it is very important for navigation. Frequency alone is not only of limited use (people cannot interprete properly nu-merical values in a context like this), it is even mis-leading: two terms of very similar weight (let's say, 'mouse' and 'PC') may belong to entirely different categories (*computer device* vs. *type of computer*), hence choosing one instead of the other may have important consequences for the remainder of naviga-tion, i.e. finding or failing to find the desired word.

The focus of the current work is to deal with the first three problems, that is the building of an index of an existing electronic dictionary (network consisting mainly of syntagmatic associations). All words are indexed in terms of the associations they evoke, and the latter are ranked in terms of frequency (see the right hand column, called coh?sion, in figure 6). For initial results, concerning automatic link extraction (problem number four), see [36].



destinatrice (2)
destiner (5400)
destinée (896)
destituer (133)
destitution (191)
destrier (6)
destroyer (30)
destructeur (248)
destructif (16)
destruction (1255)
destructuration (13)
dette (2670)
dette_extérieur (231)
dette_publique (179)
deuil (519)
deustchemark (13)
devin (33)
deviner (661)
devinette (33)

Mot	Collocations	Cohésion
frégate	24	0.295999
navire	22	0.246043
golfe	19	0.212824
missile	12	0.228062
homme	12	0.150894
aérien	10	0.210013
envoi	10	0.233353
envoyer	10	0.198995
chasseur	9	0.235086
mine	9	0.217573
avion	9	0.192707
char	8	0.227042
ravitailleur	8	0.313931
annoncer	8	0.148603
allemand	8	0.159774
lance-missiles	8	0.295801
américain	7	0.134075
pièce_d'artillerie	7	0.292084
corvette	7	0.317291
navire_de_guerre	6	0.295999
premier_ministre	6	0.153637
guerre	6	0.148095
dragueur	6	0.308295
lundi	6	0.168247
patrouilleur	6	0.302209
batterie	6	0.237343
saoudite	6	0.208927
armé	6	0.20493
soldat	6	0.181972
hélicoptère	6	0.213067
combat	6	0.176933
escadron	6	0.259587

✓ Quitter Après Opération
Nouveau Fichier
Chargement Fichier Sauvegarde
Changement Répertoire De Travail
Fichier Collocations Indexé
Compression Fichier Sauvegarde
Fusion Fichiers Sauvegarde
Pré-Traitement Fichier De Mots
Chaîne Calcul Collocations
Indexation Fichier Collocations
Statistiques Fichier Collocations
Méthode Calcul Cohésion
Affichage Paramètres
Enregistrement Paramètres
Paramètres Par Défaut
Paramètres Calcul Collocations
Autres Paramètres

Fig.6: *Ferret's collocation extractor (1998).*

As we can see, associations are a very general and powerful mechanisms, and if the very notion is age-old, its use to support word access via computer is clearly new. While we have shown elsewhere [33] how words can be accessed on the basis of their spoken or written form, we have tried to deal here with word access based on syntagmatic links, a neglected feature in WN. Even FrameNet or a normal thesaurus cannot provide the kind of navigation described here. More details concerning the envisaged strategies and the problems likely to arise when building semi-automatically the associative network can be found in [34]. Even though the work presented here is still at a very early stage and confined to a very specific task, it has the potential to go well beyond word access: in-formation access by and large, brainstorming and sub-liminal communication, to name just these.

A dictionary is a vital component for any system processing language, be it natural or artificial. There is hardly any task that we can do without it. Yet, what for an outsider seems to be one and the same object, turns out to be something very different viewed by an in-sider. Indeed, the content, structure (organization) and navigational properties of the resource, i.e. dictionary, vary considerably with the task (analysis vs. synthe-sis), time constraints (spontaneous, speech production; written text production), the nature of the information processor (man vs. machine) and the material support of the data (brain, paper, computer). The goals of this paper were twofold: (a) to show how different the resource (and its usage) can be depending on the way language is produced: automatic generation by com-puter, spontaneous discourse production (the speaker relying on his mental lexicon), or planned text-production, i.e. writing: an author making use of an existing electronic resource; (b) to illustrate how an existing electronic dictionary could be enhanced by adding an association-based index to assist the lan-guage producer (writer). There are two reasons why I have looked at the dictionary only from the language producer's point of view: space constraints and practi-cal considerations. People searching for information regarding meanings or spelling are generally quite well served with alphabetically organized dictionaries (dic-tionaries built for the language receiver), in particular if the resource is electronic, as it alleviates the problem of misspelling. The way a dictionary is built depends, of course, crucially on its ultimate usage; yet, this latter has to be anticipated. It seems to me that we have missed a chance by not taking care to look over the dictionary users' shoulders for insights. Doing so would certainly be beneficial not only for the builders of the resource, but also for psychologists, as it shows them in slow motion what people are doing and look-ing for. What can we ask for more?

References

[1] Aitchinson, J.; (2003): Words in the Mind: an Intro-duction to the Mental Lexicon. Oxford, Blackwell.

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

[3] Bateman, J. ; Zock ; M. (2003) Natural Language Generation. In: R. Mitkov (Ed.) Handbook of Com-putational Linguistics, Oxford University Press. 284-304

[4] Brown R.; McNeill, D. (1996). The tip of the tonuge phenomenon. In: Journal of Verbal Learning and Verbal Behaviour, 5:325-337.

[5] Burke, D.M.; MacKay, D.G.; Worthley, J. S.; Wade, E. (1991): "On the Tip of the Tongue: What Causes Word Finding Failures in Young and Older Adults?". In: *Journal of Memory and Language* 30, 542-579.

[6] Cahill, L.; Reape, M. (1999). Lexicalisation in applied NLG systems. Brighton, ITRI: 9.

[7] Collins, A.; Quillian, L. (1969) Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 240-247

[8] Cumming, S. (1986). The Lexicon in Text Generation, ISI: 86-168.

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

[10] Dell, G.S. (1986): A spreading-activation theory of retrieval in sentence production. Psychological Re-view, 93, 283-321.

[11] Fellbaum, C. (1998). WordNet: An Electronic Lexi-cal Database and some of its Applications. MIT Press.

[12] Fromkin, V. (ed.) (1973): Speech errors as linguistic evidence. The Hague: Mouton Publishers.

[13] Goldman N. (1975). Conceptual Generation, in Schank, R. (ed.): Conceptual Information Processing, North Holland.

[14] Hanks, P.; J. Pustejovsky. (2005). 'A Pattern Diction-ary for Natural Language Processing' in *Revue fran-caise de linguistique applique* 10 (2)

[15] Kempen, G.; Huijbers, P. (1983): The lexicalization process in sentence production and naming: Indirect election of words. Cognition, 14, 185-209.

[16] Levelt, W. (1992). Accessing Words in Speech Pro-duction: Stages, Processes and Representations. Cog-nition 42: 1-22.

[17] Levelt, W.J.M., Roelofs, A., Meyer, A.S. (1999): A theory of lexical access in speech production. Behav-ioral and Brain Sciences, 22, 1-75.

[18] Marslen-Taylor, W. (Ed.) (1979) Lexical Representa-tion and Process, Bradford book, MIT Press, Cam-bridge, Mass.

[19] Meluk I., Clas A., Polgure A. (1995) : Introduction la lexicologie explicative et combinatoire. Louvain, Duculot.

[20] Miller, G.A. (ed.) (1990): WordNet: An On-Line Lexical Database. International Journal of Lexicogra-phy, 3(4), 235-244.

[21] Nogier, J.F.; Zock, M. (1992). Lexical choice by pattern matching. In Knowledge Based Systems, 5 (3), pp. 200 - 212

[22] Quillian, R. (1968). Semantic memory. In M. Minsky (ed.) Semantic Information Processing, The MIT Press. Cambridge, MA., 216-270.

[23] Robin, J. (1990). A Survey of Lexical Choice in Natural Language Generation, Technical Report CUCS 040-90, Dept. of Computer Science, Univer-sity of Columbia

[24] Roelofs, A. (1992). "A spreading-activation theory of lemma retrieval in speaking." Levelt, W. (ed.) Special issue on the lexicon, Cognition, 42: 107-142.

[25] Sharoff, S. (2005). The communicative potential of verbs of 'away-from' motion in English and Russian. *Functions of Language*, 12:2, 203-238.

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

[27] Stede, M. (1995). Lexicalization in Natural Language Generation: a survey. Artificial Intelligence Review 8. pp. 309-336

[28] Stemberger, N. (1985) The Lexicon in a Model of Speech Production. Garland, New York.

[29] Vigliocco, G.; Antonini, T.; Garrett, M. F. (1997): Grammatical gender is on the tip of Italian tongues. *Psychological Science*, 8, 314-317.

[30] Wanner, L. (1996). Lexical Choice in Text Genera-tion and Machine Translation. Special Issue on Lexi-cal Choice. Machine Translation. L. W. (ed.). Dordrecht, Kluwer Academic Publishers. 11: 3-35.

[31] Ward, N. (1988). Issues in Word Choice. COLING-88, Budapest.

[32] Zock, M. (1996): The Power of Words in Message Planning, COLING, Copenhagen, 990-5, http://acl.ldc. upenn.edu/C/C96/C96-2167.pdf

[33] Zock, M.; Fournier, J.P. (2001). How can computers help the writer/speaker expe-riencing the tip-of-the-tongue problem ? In: Proc. of RANLP, 300-302.

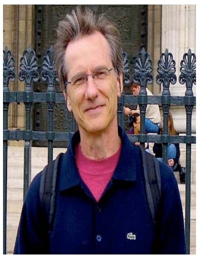
[34] Zock, M; Bilac, S. (2004). Word lookup on the basis of associations: from an idea to a roadmap. Proc. of Coling workshop: Enhancing and using dictionaries, Geneva, 29-35.

[35] Ferret, O.. (1998) ANTHAPSI : un syst?me d'analyse thmatique et d'apprentissage de con-

naissances pragmatiques fond sur l'amorcage, th?se de $3^{ème}$ cy-cle, Limsi, Orsay

[36] Ferret, O. Zock, M. Enhancing electronic dictionaries with an index based on associations. Coling/ACL, Sidney

[37] Goddard, Cliff. 1998. 'Bad arguments against seman-tic primitives'. *Theoretical Linguistics* 24, 2-3, pp. 129-156.



Michael Zock was born 1948 in Germany where lived until 1970. Ever since then he has been living in France where he got his PhD 1980 in experimental psychology (psycholinguistics). He entered the CNRS 1989, working at Limsi, an AI lab close to Paris (Orsay). 2003 he has been promoted to become research director, and 2005 he has joined the NLP group of the Laboratoire d'Informatique Fondamentale (LIF) at Luminy, Marseille. His research interests are geared towards language production by people and machines, with a special emphasis on the simulation of the process and the building of tools to help people to perform this task or to acquire the needed skills. His current research foci comprise lexical access, acquisition of verbal fluency, ontology-driven message- and outline planning.