Lexical Knowledge and Natural Language Processing

Thierry Fontenelle

(Microsoft Natural Language Group)

Introduction

A few days ago, I had a dream. I dreamt of an intelligent computer which would be able to extract knowledge from unstructured text and would be able to answer questions concerning this text. More specifically, I remembered the time when, as a language teacher in my home university, I had to examine scores of students, mark examinations and participate in a sometimes cruel selection process. The scenario which came to my mind when I had this wonderful famous dream was made of a couple of sentences such as:

> At the beginning of this year, I had 100 students and 90 eventually took my exams. I marked their assignments and flunked a third of the undergraduates.

In my dream, the smart machine I had imagined was able to process and answer the following natural language questions, very much like any reasonably intelligent human being:

- How many people failed?
- How many people passed?
- How many students passed the tests?
- How many undergraduates flunked?
- How many students fluffed the tests?
- How many students did the teacher pass?

All these questions may seem naïve insofar as any 10-year-old should be able to answer them without any difficulty. It is interesting to examine the psycholinguistic mechanisms which are activated to analyse these questions and map them onto information extracted from the basic scenario. How do we manage to infer that 30 students fluffed the tests from a statement that the teacher flunked one third of the 90 undergraduates who took the exams? What kind of lexical knowledge do we need to activate and what kind of cognitive mechanisms do we have to trigger in order to make it possible for a computer to

imitate human behaviour? These issues will be addressed in the remainder of this paper.

Computational lexicography and natural language processing

For over two decades, researchers have tried to tap a variety of lexical and textual resources to populate the lexical components of their natural language processing systems. Commercial dictionaries produced by established publishing houses have been found to contain a lot of syntactic, semantic and pragmatic information condensed into compact lexical entries. Over the years, methods were developed to acquire this crucial knowledge from the electronic versions of these commercial dictionaries (Amsler 1979, Michiels 1982, Boguraev & Briscoe 1989, Wilks et al. 1996). The initial attempts to reuse existing dictionaries focused on monolingual reference works, mainly English learner's dictionaries, whose systems of grammatical codes and simplified definitions had been found to house the very syntactic information required to drive a parser (although other researchers have shown that relying too much on dictionaries is dangerous and may not reflect the evidence found in large corpora, Atkins & Levin (1991) being a case in point). The automatic identification of genus terms made it possible to construct partial taxonomies of is_a relations. Such hierarchies of hyperonyms, hyponyms and co-hyponyms are indeed a sine qua non in information retrieval, where a question rarely exactly matches the vocabulary used in the answers. Such a requirement probably accounts for the widespread use of the WordNet database (Fellbaum 1998), which, despite its limitations, has the undeniable merit of being freely accessible and of offering a very wide lexical coverage and a whole gamut of lexical-semantic relations.

In the scenario described at the beginning of this paper, one of the main problems is to make it possible for a computer to compute the similarity between the word *test*, used in some of the questions, and the word *exam*, used in the source text containing the information to be exploited. The following entries, from a variety of well-known dictionaries available in electronic form, ranging from WordNet to LDOCE (Procter 1978) or Cobuild (Sinclair 1987), show that this similarity can be discovered and computed by relying on existing resources, although the level of preparatory work is different from one resource to another. While WordNet makes it possible to go from *test* to *exam* simply because they belong to the same synset (set of synonyms, in WordNet parlance), the exploitation of LDOCE or Cobuild requires a more elaborate analysis of definitions.

WordNet:

test: Sense 4 examination, exam, test -- (a set of questions or exercises evaluating skill or knowledge; "when the test was stolen the professor had to make a new set of questions") => communication, communicating -- (the activity of communicating) => act, human action, human activity -- (something that people do or cause to happen)

LDOCE:

exam *n infml* EXAMINATION (1) **examination** *n* **1** [C(in or on)] a spoken or written test of knowledge

Cobuild:

exam. An **exam** is an official and formal test that you take to show your knowledge or ability in a particular subject or to obtain a qualification	N COUNT = examination

What is clearly needed here is, in any case, a thesauric approach to the organization of the lexicon in order to capture semantically similar items which, in a traditional thesaurus such as Roget's, appear under the same class (see also Calzolari 1988).

Transitivity alternations

In traditional grammar, a transitive verb is defined as a verb which takes a direct object while an intransitive verb occurs without any such direct object (I watched a film on TV last night *vs.* It rained for two hours). Atkins et al. (1986) have shown that this distinction, which is very often used to identify seemingly different senses in dictionaries, is much too superficial and that the linguistic description of the syntactic behaviour of verbs needs to rely on further classifications which are unfortunately much too implicit in dictionaries. In our scenario, it is clear that the question

How many undergraduates flunked?

can only be answered if one realizes that the subject of the intransitive verb *flunk* (= undergraduates) appears as and corresponds to a direct object of the same verb used transitively in the source text of our initial scenario:

I flunked a third of the undergraduates.

This property is not typical of the verb *flunk*, of course. In the same context, this alternation may be found with similar verbs, as is shown in the following examples:

(a) The teacher failed 10 students.
(b) 10 students failed.
(c) The teacher passed 10 students.
(d) 10 students passed.

These verbs are frequently referred to as ergative verbs, i.e. verbs which display the so-called causative/inchoative alternation. This alternation is only one among a much larger set of transitivity alternations (see Levin 1993 for an in-depth study of this fascinating area of the lexicon), but it concerns a sizeable number of English verbs which, like *boil*, *open* or *increase*, can be used transitively with a direct object corresponding to a patient argument undergoing a change of state, or intransitively, with the same patient argument realized as a subject (John opened the door *vs.* The door opened; He boiled the water *vs.* The water boiled; The government increased the price of oil *vs.* The price of oil increased). Several attempts have been made to extract this class of verbs from machine-readable dictionaries, drawing upon a variety of techniques ranging from a careful analysis of definitions and defining formulae ((cause to) V...; make or become + N/Adj; (allow to) V....) to the identification of combinations of grammar codes describing transitivity or intransitivity (see *inter alia* Fontenelle & Vanandroye (1989) or Boguraev (1991) for work on monolingual English dictionaries, Antelmi & Roventini (1992) for work on Italian dictionaries, _ikra(1992) on the Czech language or Fontenelle (1997a, Chapter 5) for work on bilingual English-French dictionaries).

Collocations

The tendency for words to co-occur in prefabricated chunks of language has attracted a lot of attention over the last decade (Sinclair 1991, Cowie 1998). The availability of very large corpora has made it possible to shed some new light onto the concept of collocation and statistical tools are now the norm rather than the exception in many publishing houses, whose lexicographers are confronted with the seemingly insurmountable task of having to sift through thousands of concordances to extract the most relevant facts about the behaviour of the lexical item they are analysing (see Church et al. (1994) for very useful examples of statistical techniques such as mutual information, t-scores, z-scores, etc. applied to dictionary compiling). Research in applied linguistics and language learning has shown that words are best learnt and retained if they are

presented in context and more specifically together with the other items with which they are most likely to appear. On the other hand, native and non-native speakers are very often faced with the tip-of-the-tongue phenomenon, which causes them to look, sometimes in vain, for the appropriate word expressing a given meaning in a given context. These observations have resulted in the creation of a new generation of dictionaries which take the collocational dimension as a central axis and are specifically designed to meet the requirements of those who wish to encode text. The BBI dictionary of collocations (Benson et al. 1986) has remained unrivalled, by and large, despite all the criticism that has been levelled against it. The Explanatory Combinatory Dictionaries produced by Igor Mel'čuk's team have come up with remarkable descriptions of a few hundred French lexemes, focusing on the revolutionary concept of lexical function to capture a wide range of paradigmatic and syntagmatic relations. Indeed, it has been observed that a combination such as *confirmed bachelor* usually only poses a problem when the user starts from the noun (*bachelor*) and tries to find out which adjective can collocate with it in order to express an intensifying meaning. In an encoding perspective, we therefore expect to find this combination under the noun entry since any user would most probably try to discover which adjective to use in the vicinity of *bachelor*. Yet a dictionary such as the Collins-Robert English-French dictionary (Atkins & Duval 1994) lists this collocation under *confirmed* only:

> **confirmed** *adj smoker, drunkard, liar* invétéré; *bachelor, sinner* endurci; *habit* incorrigible, invétéré

The decision to include the collocations under the collocator is justified in a decoding perspective since the dictionary can be used to find out how to translate *confirmed* in various contexts. But the user who wishes to make use of the same dictionary in an encoding perspective is left in the lurch. This is where the concepts of lexical function and of electronic dictionary come in handy. In a combinatory dictionary à la Mel'čuk, such collocational information is indeed to be found under the base of the collocation, i.e. the keyword, which is related to other words by means of lexical functions, i.e. lexical-semantic relations expressed in the traditional mathematical form f(x)=y. In the example above, the collocation *confirmed bachelor* may be represented as Magn (bachelor) = confirmed, which can be read as "the Magn (magnification = intensifying) meaning of the keyword may be expressed with the adjective *confirmed*". About 60 standard lexical functions make it possible to formalize a wide range of paradigmatic and syntagmatic relations (see Mel'čuk 1984, Fontenelle 1997a):

Paradigmatic LFs	A_0 (sun) = solar A_0 (law) = legal A_0 (lexicon) = lexical	A_0: adjective derived from the keyword
	$Able_1$ (read) = literate $Able_2$ (read) = legible	Able: adjective denoting a capability of the 1st, 2nd actant to perform an action inherent in a keyword
	V_0 (advice) = advise V_0 ($promise_n$) = $promise_{vt}$	V_0= verbal form
Syntagmatic LFs	Sing (dust) = grain Sing (grass) = blade	Sing: regular portion
	Mult (fish) = school, shoal Mult (abuse) = spate, storm	Mult: regular group/set
	Son (elephant) = trumpet Son (clock) = tick	Son: typical verb denoting a sound or cry
	$Oper_1$ (attention) = pay $Oper_1$ (pressure) = exert	Oper: "support" verb (make/do)
	$Real_1$ (promise) = keep $Real_2$ (advice) = follow	Real: comply with the requirements, demands of
	Liqu (law) = abolish Liqu (disease) = eradicate	Liqu: liquidate, destroy

If we turn back to the original scenario alluded to at the beginning of this paper, we are able to describe some interesting collocations in terms of lexical functions. Starting from the keyword *exam*, which can act as a central node for our investigation, we may postulate the existence of the following triples which are stored in our mental lexicon:

$Oper_2$ (exam) = sit, take [sit / take an exam]
$Real_2$ (exam) = pass [pass an exam]
$AntiReal_2$ (exam) = fail [fail an exam]

In Meaning-Text Theory (MTT) parlance, the second actant, i.e. the person who is being examined or tested, is the subject of the verbs listed above, hence the use of the subscript $_2$. One sees that some of these verbs may be described as semantically impoverished and correspond to what some linguists have called "support verbs" (one sits or takes an exam = one is being examined or tested). The other verbs carry more semantic weight, however, as a function of the outcome of the test (success → Real; failure → AntiReal).

One of the questions which arise is how to extract such collocational combinations. This has been a hot topic for over a decade now and one of the main sources is most certainly the huge corpora to which sophisticated statistical techniques are applied in order to identify salient, relevant and typical patterns. The notions of relevance and salience are also crucial issues which fall outside

the scope of this paper, but which should not be neglected when discussing the usefulness of such tests as mutual information, which identifies pairs of items co-occurring more often than chance would predict, or t-scores, which is better at contrasting items in context.

In other publications, I have described another type of resource from which it has been possible to extract useful collocational material. Taking the Collins-Robert English-French dictionary (Atkins & Duval 1978) as a starting point, a database of 70,000+ lexical relations was created, tapping the collocational material appearing in italics in the printed version of the dictionary (see Fontenelle 1997 a & b for more information on the construction of this lexical-semantic database and on the rationale which underlies it).

The following examples illustrate the typographical conventions which are applied throughout the dictionary to list possible subjects of verbs (between square brackets), direct objects (unbracketed), noun complements (bracketed) or nouns modified by adjectives.

dart 1 *n* ... **c** *(weapon)* trait, javelot; *(liter) [serpent, bee]* dard
drone 1 *n* **a** *(bee)* abeille m le, faux bourdon ... **b** *(sound) [bees]* bourdonnement; *[engine, aircraft]* ronronnement, *(louder)* vrombissement ... **2** *vi [bee]* bourdonner; *[engine, aircraft]* ronronner, *(louder)* vrombir...
fail 1 *vi* **a** *(be unsuccessful) [candidate]* chouer, tre coll * *or* recal * *(in an exam* un examen, *in Latin* en latin ( ) **2** *vt* **a** *examination* chouer , tre coll * *or* recal * ; *candidate* refuser, coller*, recaler* *(in an exam* un examen)
fluff *vt* **a** *(also* ~ **out)** *feathers* bouriffer; *pillows, hair* faire bouffer. **b** *(* do badly) audition, lines in play, exam* rater, louper*
keep vt **e** *(own; look after) shop, hotel, restaurant* tenir, avoir; *house, servant, dog, car* avoir; *(Agr) cattle, pigs, bees, chickens* lever, faire l' levage de

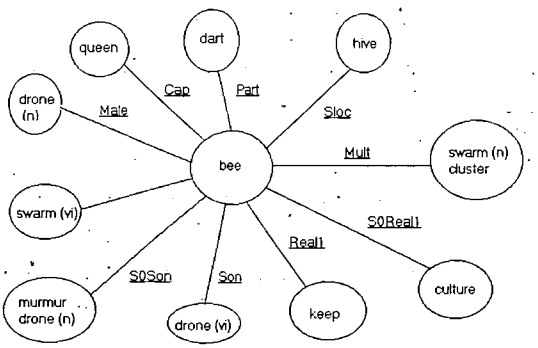
Whenever possible, the relationship between the base of the collocation, which corresponds to the italicized indicator, and the collocator itself (the headword) has been identified and made explicit in terms of a Mel'čukian lexical function. The original contents of the dictionary has then been enriched with a semantic layer, thereby adding a new access path to make it possible to query the data via the base, the collocator or the lexical function, using these criteria as filters in isolation or in combination. The resulting semantic networks take the form of triples which enable the user to reconstruct the lexical web in which a given item may be found.

A quick glance at the entries for *dart*, *drone* and *keep* in the table above reveals that the word *bee* appears 5 times as a metalinguistic indicator in italics, i.e. as a piece of information provided by the lexicographer to guide the user to the appropriate meaning and the correct translation. The status of this

metalinguistic label differs from one entry to the other, however, and the typographic conventions contribute to clarifying this status. *Bee* between parentheses undoubtedly refers to a synonym or a hypernym, s.v. *drone* (1), for instance. When surrounded by square brackets, *bee* can play the part of a collocate or noun complement (drone of bees; dart of a bee) or can refer to a typical subject in a verb entry (bees typically drone). The occurrence of the unbracketed string *bee* in a verbal entry points to a verb-object relation. In addition to these syntagmatic relations (Noun-Noun or Noun-Verb or Verb-Noun collocations), the database includes an explicitation of the lexical-semantic link between a given italicised indicator and the entries under which it appears. In terms of Mel'čuk's lexical functions, the entries containing *bee* above can be rewritten as follows:

Male (bee) = drone (n) (a male bee is a drone).
Son (bee) = drone (vi) (typical verb for the sound made by bees)
S_0Son (bee) = drone (n) (typical noun for the sound made by bees)
Part (bee) = dart (part-whole relationship)
$Real_1$ (bee) = keep (verb denoting the typical action associated with bees)

The semantic network which can be built for all the occurrences of *bee* in the English-French part of the dictionary can be represented diagrammatically as follows:



[Lexical-semantic relations: Cap = head of; Mult = group of; S_{loc} = noun for the typical location of; Son = sound of (verb); S_0Son = sound of (noun); $Real_1$ = typical action (verb); S_0Real_1 = noun for the typical action; note that the term 'semantic network' may be considered an oversimplification since some of the pairs of collocate have not been assigned a standard lexical function - see *swarm*].

One clearly sees that the same scenario may be viewed from different angles altogether and the purpose of this table is to enable a user to select a particular predicate and give the frame elements which revolve around it the appropriate syntactic function. This also provides evidence that the traditional notions of transitivity and intransitivity are insufficient to account for semantic distinctions and role assignment. In the following sentences, for example, the verb *fail* is used transitively, but with different frame element groups (FEG):

(3) The student *failed* the driving test.
(4) The examiner *failed* him because he had gone through a red light.

Fail : FEG: (3) {Examinee, Event}
(4) {Examiner, Examinee}

It is clear that it is not sufficient to state that *fail* subcategorizes for a [+HUMAN] subject since this notion encompasses both examinees (in 3) and examiners (in 4). In a translation perspective, such knowledge will drive the choice of *échouer à* or *faire échouer* in French, for instance and is essential if one wishes to teach a machine to understand the scenario described in the introduction and to make the inferences which are required to answer the natural language questions listed there.

Conclusion

In this paper, I have tried to describe some aspects of the lexical information an "intelligent" NLP system ought to be able to draw upon. Robust natural language processing has often been described as an AI-complete problem insofar as it presupposes the resolution of the most complex artificial intelligence problems. It should be clear that the lexical databases which can be used to meet the requirements described at the beginning of this paper should be seen as combinations of traditional dictionaries, thesauri, collocational and semantic networks. Much more research still needs to be done to find out how best to integrate the various (and other) approaches which have been alluded to in this paper and how to acquire all this knowledge, preferably by automatic means. Nobody knows when machines will be able to fully understand and manipulate natural language (whether spoken or written). However, if they ever manage to crack all these linguistic problems, the researchers and developers who try to meet these exciting challenges should be grateful to Sue Atkins for both initiating and developing the types of lexical resources they need and for

contributing to the advancement of our perception of what we should (or shouldn't) look for in dictionaries and corpora.

Acknowledgements

The original development of the Collins-Robert lexical-semantic database took place at the University of Liège (Belgium). Thanks are due to the publishers for granting us access to the tapes of the dictionary and for allowing us to go on using it for research purposes.

References

A. Dictionaries and Thesauri

Atkins, B. T. and Duval, A. 1978. *Robert-Collins Dictionnaire Français-Anglais, Anglais-Français.* (First edition; third edition edited by Sinclair, L. and Duval, A.) Paris: Le Robert and Glasgow: Collins. (CR)

Benson, M., Benson, E. & Ilson, R. 1986: *The BBI Combinatory Dictionary of English,* Amsterdam and Philadelphia, John Benjamins.

Fellbaum, C. (ed.) 1998 *WordNet: An Electronic Lexical Database.* Cambridge, Mass. and London: MIT Press.

Mel'čuk I. et al. 1984. *Dictionnaire Explicatif et Combinatoire du Français Contemporain.* Montréal: Presses de l'Université de Montréal.

Procter, P. (ed.) 1978: Longman Dictionary of Contemporary English, (2nd edition edited by D. Summers), Longman Group Ltd, Harlow.

Procter, P. (ed.) 1995. *Cambridge International Dictionary of English,* Cambridge University Press. (CIDE)

Sinclair, J. et al. (eds.) 1987. *Collins COBUILD English Language Dictionary.* (First edition.) Glasgow: HarperCollins. (Cobuild)

B. Other references

Amsler, R.A. 1980. The Structure of the Merriam-Webster Pocket Dictionary, Ph.D. Thesis, University of Texas at Austin, Austin.

Antelmi, D. & Roventini, A. 1992. 'Semantic relationships within a set of verbal entries in the Italian Lexical Database', in *Euralex'90 Proceedings,* Barcelona: Biblograf, pp.247-255.

Atkins, B.T., Kegl, J. & Levin, B. 1986. 'Explicit and implicit information in dictionaries', Lexicon Project Working Papers 12, Center for Cognitive Science, MIT, Cambridge, MA. Also available as Cognitive Science Laboratory Report 5, Cognitive Science Laboratory, Princeton University, Princeton, NJ.

Atkins, B.T. & Levin, B. 1991. 'Admitting Impediments', in U. Zernik (ed.) *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, Hillsdale, NJ, Lawrence Erlbaum Associates, pp.233-262.

Baker, C., Fillmore, C. and Lowe, J. B. 1998. 'The Berkeley FrameNet Project.' *Proceedings of ACL/COLING 1998.*

Boguraev, B. 1991. 'Building a Lexicon: The Contribution of Computers', *International Journal of Lexicography*, 4/3, pp.227-260.

Boguraev, B. & Briscoe, T. 1989. *Computational Lexicography for Natural Language Processing*, London and New York, Longman.

Calzolari, N. 1988. 'The dictionary and the thesaurus can be combined', in Evens (ed.) *Relational Models of the Lexicon*, Cambridge University Press, pp.75-96.

Church, K. and Hanks, P. 1990. 'Word Association Norms, Mutual Information and Lexicography.' *Computational Linguistics* 16.3: 22-29.

Cowie, A.P. (ed.). 1998. *Phraseology. Theory, Analysis, and Applications*. Oxford University Press.

Fillmore, C. J. 1982. 'Frame Semantics' in The Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm*, Seoul, Hanshin, 111-37.

Church, K., Gale, W., Hanks, P., Hindle, D. & Moon, R. 1994. 'Lexical Substitutability', in Atkins & Zampolli (eds) *Computational Approaches to the Lexicon*, Oxford University Press, pp.153-177.

Fillmore, C. J. and Atkins, B. T. S. 1992. 'Towards a Frame-Based Lexicon: the Case of RISK' in A. Lehrer and E. F. Kittay, (eds.), *Frames, Fields and Contrasts*. Hillsdale NJ: Lawrence Erlbaum Associates, 75-102.

Fillmore, C. J. and Atkins, B .T. S. 1994. 'Starting where the Dictionaries Stop: the Challenge for Computational Lexicography' in B. T. S. Atkins and A. Zampolli (eds.), *Computational Approaches to the Lexicon*. Oxford: Oxford University Press, 349-93.

Fillmore, C. J. and Atkins, B .T. S. 1998. 'FrameNet and Lexicographic Relevance.' *Proceedings of the Granada Conference on Linguistic Resources*, 417-23.

Fillmore, C. J., Wooters, C. and Baker, C. 2000. 'Buidling a Large Lexical Databank Which Provides Deep Semantics.' *Proceedings of the Pacific Asian Conference on Language, Information and Computation*. Hong Kong.

Fontenelle, T. 1997a. *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Tübingen: Max Niemeyer Verlag.

Fontenelle, T. 1997b. 'Using a Bilingual Dictionary to Create Semantic Networks.' *International Journal of Lexicography* 10.4: 275-303.

Fontenelle, T. 2000. 'A bilingual lexical database for frame semantics.' *International Journal of Lexicography* 13.4: 232-248.

Fontenelle, T. & Vanandroye, J. 1989. 'Retrieving ergative verbs from a lexical database', *Dictionaries: Journal of the Dictionary Society of North America*, Vol.11, pp.11-39.

Levin, B. 1993. *English Verb Classes and Alternations - A Preliminary Investigation*, Chicago and London, the University of Chicago Press.

Lowe, J. B., Baker, C. and Fillmore, C. 1997. 'A Frame-Semantic Approach to Semantic Annotation' in *Tagging Text with Lexical Semantics: Why, What, and How? Proceedings of the Workshop*. Special Interest Group on the Lexicon, Association for Computational Linguistics, 8-24.

Michiels, A. 1982. *Exploiting a Large Dictionary Database*. PhD Thesis, University of Liège, mimeographed.

Michiels, A. 2000. 'New Developments in the DEFI Matcher.' *International Journal of Lexicography* 13.3: 151-67.

Šikra, J. 1992. 'Dictionary Defining Language', in Tommola, Varantola, Salmi-Tolonen & Schopp (eds) *EURALEX'92 Proceedings I-II*, Fifth EURALEX International Congress, Studia Translatologica, Ser.A, Vol.1, University of Tampere, pp.295-300.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*, Oxford University Press.

Wilks, Y., Slator, B. & Guthrie, L. 1996. *Electric Words - Dictionaries, Computers and Meanings*, The MIT Press, Cambridge, MA and London.