Compounds: an intelligent tutoring system for learning to use compounds in english

Paul Boucher, Frédéric Danna, Pascale Sébillot

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Compounds: an Intelligent Tutoring System for Learning to Use Compounds in English

Paul Boucher, Frédéric Danna et Pascale Sébillot

N° 1974

Avril 1993

PROGRAMME 3

Intelligence artificielle,
systèmes cognitifs
et interaction homme-machine

ISSN 0249-6399

Compounds: an Intelligent Tutoring System for Learning to Use Compounds in English

Paul Boucher*, Frédéric Danna** et Pascale Sébillot**

Programme 3 — Intelligence artificielle, systèmes cognitifs
et interaction homme-machine
Projet Repco

Rapport de recherche n° 1974 — Avril 1993 — 21 pages

Abstract: In this paper, we present *Compounds*, an Intelligent Tutoring System which can teach French students of English how to produce and understand lexicalised and newly formed compounds. This ITS must be able to detect the difficulties of each student and to provide him or her with an appropriate set of exercises to improve him or her weak points. Here, we focus on the creation of the expert model. We study different linguistic theories of the English compounding process and we build a representation of the expert model together with a generic model of the errors that French students make. We describe the computer implementation of these two elements.

Key-words: ITS, second language learning, expert and student models, English compounds.

(Résumé : tsvp)

A French version is available by request to P. Sébillot at sebillot@irisa.fr

*Cerlico, université de Nantes, institut des langues, littératures et civilisations étrangères, section d'Anglais, chemin de la Cencive du Tertre, BP 1025, 44036 Nantes cédex 01
**{danna}{sebillot}@irisa.fr

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 99 84 71 00 – Télécopie : (33) 99 38 38 32

Compounds : tuteur intelligent pour apprendre le processus de composition anglais

Résumé : Dans cet article, nous présentons *Compounds*, un logiciel d'Enseignement Intelligemment Assisté par Ordinateur permettant à des étudiants francophones d'apprendre à maîtriser l'emploi et la compréhension de composés anglais nouveaux et existants. Ce logiciel doit s'adapter aux difficultés de chaque étudiant et lui proposer une suite d'exercices permettant de travailler ses points faibles. Ici, nous nous focalisons essentiellement sur la création d'un modèle de l'expert à partir d'une étude très précise de diverses théories linguistiques du phénomène de composition anglais. Nous présentons une représentation de ce modèle ainsi que celle d'un modèle générique des erreurs commises par des francophones. Nous décrivons une implantation informatique de ces éléments.

Mots-clé : EIAO, apprentissage d'une seconde langue, modèles de l'expert et de l'élève, composés anglais.

1 Introduction

Learning English involves, among other things, learning to form and use compounds. This is an important part of the learning process since compounding is an extremely productive source of word formation in English, especially in the technical and scientific fields.

A *compound in English* is a *word-level structure*, composed of at least two items, each of which belongs to one of the four lexical categories, N (noun), A (adjective), P (preposition) or V (verb). The compound itself belongs to one of the three categories N, A or V [Sel82].

Example:

$[[window]_N - [[wash]_V er]_N]_N$ $[[bird]_N\ [dog]_N]_N$ $[[high]_A\ [school]_N]_N$
$[[over]_P\ [dose]_N]_N$ $[[spoon]_N - [feed]_V]_V$ $[[rattle]_V\ [snake]_N]_N$

A French-speaking student who wants to learn English must be able to produce and understand lexicalised or newly formed compounds. Tests carried out on about 200 students at the University of Rennes 2 since 1989 [Bou92b] have shown the difficulties French-speaking students have in mastering the compounding process in English. In tests of comprehension and production, they make numerous errors of various sorts: morphological, syntactic and semantic errors. For instance, when asked to produce a compound corresponding to a definition in English (e.g.: a *house which dogs live in* or a *dog which hunts birds*), they may use more than the two terms required (*an orange made juice[1]), use illicit affixes or pluralize the left-hand term (*a flies net). They have trouble admitting that the semantic relationship between the terms of a compound must be partly implicit. Moreover, they often apply rules appropriate for French composition processes to English.

To summarize, the major difficulties French-speaking students have with the English system of compounding seem to derive from three different areas:

1. the nature of the compounding process is different in English from that of French (a *word-compounding* process versus a *stem-compounding* process);

2. the use of the tonic accent in compounds, which is essential in English, is poorly understood by French speakers;

   NB: The question of stress-marking will be left aside here (cf. [Blo33]).

3. The actual semantic relationships between the terms of a compound are quite different from those provided spontaneously by a French speaker.

In order to help French students of English, we decided to build an Intelligent Tutoring System (ITS) called *Compounds* which could teach them this

[1] In this article, incorrect examples will be marked with an asterisk.

aspect of learning English. This system must be able to adapt itself to the learning problems of each student, that is, it must be able to detect the correct and incorrect rules each student is applying and provide him or her with an appropriate set of exercises to improve his or her weak points.

An Intelligent Tutoring System is generally made up of four parts [NV88]: *the Expert Model*, which contains knowledge of the domain, *the Student Model*, which contains information about the student and his knowledge, *the Teaching Generator*, which generates teaching plans and *the User Interface*, which handles communication between the student and the ITS.

In order to diagnose student knowledge, we need to have a precise theoretic model of the knowledge area being taught. This model permits to represent expert knowledge of the domain, that is, the "correct" knowledge the student must acquire. Our first task was therefore to build a precise formal linguistic model of the compounding process in English. We have concentrated our attention initially on compounds of the form N-N, V-N, N-V, N-Ving, N-Ver, Ving-N and Ver-N, since these are the most frequently used types of compounds in English. Using this linguistic model, we then built an expert model which idealises the competence of an English speaker as the ability to produce the appropriate compound corresponding to a given definition and to generate the correct definition corresponding to a given compound.

This model allows us to generate two types of exercises in the current state of the project: converting a definition into a compound and converting a compound into a definition.

Aside from its use in teaching English, this ITS is of interest for several more general reasons:

- Relatively few ITS currently under development deal with the problem of second language learning. This is due in large part to the relatively unstructured nature of knowledge in this field, as compared with that of teaching mathematics or computer programming.

- The study of compounds is intrinsically interesting, since in fact they represent a sort of micro-world within the English system: the semantic relationships holding between the terms of a compound correspond to the same relationships holding between these terms when they are used as phrases in a sentence [Lie83].

- The problem of compounds has received relatively little attention in linguistic litterature.

- School books and grammars usually limit their remarks to vocabulary lists, translations or paraphrases. One is led to conclude either that the problem is too complicated to be taught, or that the only solution is to learn individual examples by heart. This gives the impression that there are no general rules for interpreting compounds.

In this article, we show how we built the Expert Model of our ITS from the theoretical linguistic model. We also present a generic model of student errors. In order to demonstrate the problems French students have in learning the compounding process in English, we refer first of all to the errors recorded in tests given at the University of Rennes 2. Then we describe the linguistic model we have built using various current theories. This model summarizes the correct knowledge needed for the production and interpretation of English compounds, but its rules are too abstract to be taught directly to French students. Using this theoretic model, we explain the way we represent knowledge in our ITS. We conclude by briefly describing how the ITS is programmed and the extensions currently under development.

2 Errors committed by French students

Building an ITS involves writing a formal model of expert knowledge as well as a model of the current state of student knowledge. In our case, the knowledge in question corresponds to the composition patterns used by an English-speaking native when he attempts to define a compound and vice versa.

The composition process can be codified using *logical representations*. A (correct or incorrect) composition pattern can be represented by a single initial logical formula:

$$<\text{predicate}>(<\text{argument}_1>^2, <\text{argument}_2>^3)$$

as well as by a set of correct or incorrect final formulae. For instance, the correct pattern for producing compounds of the form N-N is:

$$<\text{predicate}>(<\text{argument}_1>, <\text{argument}_2>) \rightarrow <\text{argument}_2>\text{-}<\text{argument}_1>$$

Using this pattern we obtain the compound *bird-dog* from the logical formula *hunt(dog, bird)*, which corresponds to the definition a *dog which hunts birds*.

This enables us to define the correct and incorrect patterns among the set of all possible ones which can be generated from the initial logical formula.

2.1 Correct Knowledge - The Expert Model

Using a semantic structure of the form <predicate>(<argument$_1$>,<argument$_2$>), an English-speaking native can produce the following types of correct compounds:

1. <argument$_2$>-<argument$_1$>: *bird dog, coffee mill, farm boy, cow pasture*;

[2] The grammatical subject (i.e. the agent or instrument of the action).
[3] The internal argument or the external argument (cf. section 3.1).

2. <predicate><affix>[4]-<argument$_1$>: *tracer bullet, worker bee*;
3. <predicate><affix>-<argument$_2$>: *swimming pool, cut throat, pick pocket, love nest*;
4. <argument$_2$>-<predicate><affix>: *data processing, window washer, day tripper*.

2.2 Student Knowledge

In the results of the tests given in Rennes, French students usually produced the following incorrect forms, working from a semantic structure of the form <predicate>(<argument$_1$>,<argument$_2$>):

1. <argument$_1$>-<argument$_2$>: **dog-bird* for a *dog which hunts birds*;
2. <argument$_1$>-<predicate><affix>: **bee-worker* for a *bee which works*;
3. <argument$_1$>-<argument$_2$>-<predicate><affix>: **dog-bird-hunting* for a *dog which hunts birds*;
4. <argument$_2$>-<argument$_1$>-<predicate><affix>: **bird-dog-hunting* for a *dog which hunts birds*;
5. <argument$_2$>-<predicate><affix>-<argument$_1$>: **window-washer-man* for a *man who washes windows* or **throat-cut-man* for a *man who cuts throats*;
6. <argument$_1$>-<predicate><affix>-<argument$_2$>: **dog-hunting-bird* for a *dog which hunts birds*;
7. <predicate><affix>-<argument$_1$>-<argument$_2$>: **hunting-dog-bird* for a *dog which hunts birds*;
8. <predicate><affix>-<argument$_2$>-<argument$_1$>: **hunting-bird-dog* for a *dog which hunts birds*.

Other errors found were due to lack of vocabulary or to the misapplication of correct composition rules. For example, students tend to add inappropriate suffixes (e.g.: **flies-net* instead of *fly-net*). A more detailed analysis of the results of these tests can be found in [Bou92b] and [Bou92a].

[4] In this article, an affix will be either *-er*, *-ing*, *-ed* or the empty suffix.

3 The theoretical linguistic model

In this part we present first of all the vocabulary needed to understand the following linguistic theories. Then we describe these theories, which we used to build our model. We comment on some of the results obtained with these theories and offer our own solution which is able to account for all the types of compounds we studied.

3.1 Vocabulary

- A *DO (or alternatively IDO)* is a direct (or indirect) object.
- An *Internal Argument*[5] is an obligatory argument (e.g.: the DO and the IDO for a verb like *give*).
- An *External* (or *semantic*) *Argument*[5] is an optional argument (e.g.: the DO for some verbs like *hunt*, *sing*, etc).
- The *Argument Structure* is a set of internal and external arguments.
- A *Predicative Term* is a term which has an argument structure (its lexical category is therefore V or P).
- The *Head* of a compound is the item which governs the compound semantically. The head is generally the right-hand term in English.
- A *Synthetic Compound* (Lieber [Lie83]) is an compound adjective or noun made up of any item and a second item (N or A) which is morphologically derived from a verb (i.e. N-(Ver), N-(Ving), (Ver)-N, etc).
- A *Primary Compound* (Lieber [Lie83]) includes all the other compound configurations, i.e. all those which do not contain a suffixed verbal base: N-V, V-N, etc.
- A *Verbal Compound* (Selkirk [Sel82]) is a synthetic compound with a verbally derived head term (i.e. V<affix>). Moreover, the head term must assign a thematic relation (i.e. agent, theme, goal, source, instrument, etc) to the second term. Some examples of verbal compounds are: *time-saver, handwoven, nice-sounding, surface adherence, water-repellent.*

  NB: Selkirk only discusses verbal compounds with right-hand heads, though she does point out the existence of a number of verbal compounds with left-hand heads, of the form V-P (e.g.: *grow-up, sit-in, step-out*).

[5] This terminology is borrowed from [Wil80], but we use the terms with slightly altered meaning.

3.2 Linguistic theories of English compounds

3.2.1 The work of Downing

A study devoted to compound nouns of the form N_1–N_2 was carried out by Downing. Her article [Dow77] is based on a survey made of a number of English-speaking subjects in order to determine the types of semantic relationships which are possible between the terms of a compound. Twelve types of semantic relationships were discovered: whole-part (e.g.: duck foot), half-half (e.g.: giraffe-cow), part-whole (e.g.: pendulum clock), composition (e.g.: stone furniture), etc.

The semantic relationship between the two terms depends largely on the semantic class of the head noun. Five main classes exist in Downing's opinion: human, animal, plant, natural object and synthetic object.

The following table gives the semantic relationships which may hold between two terms of a compound, by decreasing order of probability:

Class of the head-noun	Semantic relationship	Examples
Human	occupation, sexual and racial identity	police demonstrators, women officers, Negro woman
Animal	appearance, habitat	giraffe bird, Salt Creek coyotes
Plante	appearance, habitat	trumpet plant, Texas roadside flowers
Natural object	composition, origin, place	granite outcropping, cow hair, Montana beach
Synthetic object	goal	banana fork

Some compound nouns of the form N_1–N_2 were judged to be impossible (or improbable) by the subjects interviewed. The compounds were rejected for the following reasons:

- the set of referents of N_1 is equivalent to the set of referents of N_2 (e.g.: **lad-boy*);
- the set of referents of one term of the compound is a proper subset of that of the other term (e.g.: **book-novel*, **horse-animal*, **truck-vehicle*);
- the referents of N_2 are necessarily of the same type as those of N_1–N_2 (e.g.: **egg-bird*, **ground-flower*, **head-hat*, **hog-pork*, **mouth-whistle*, **time-hour*, **wind-flag*).

3.2.2 The work of Lieber

Lieber [Lie83] covers all the types of compounds by means of the *Feature Percolation Conventions* (FPC) and the *Argument Linking Principle* (ALP). The FPC and the ALP enable us to predict with a high degree of certainty what the underlying semantic relation of a given compound will be and whether a given pattern will be correct or not.

The FPC allow us to determine the syntactic and semantic features of a phrase after syntactic analysis and to apply the ALP. There are four conventions (cf. figure 1):

- **Convention I**: The features of a stem will percolate to the first non-branching node[6] that dominates it;
- **Convention II**: The features of an affix will percolate to the first branching node[7] that dominates it;
- **Convention III**: If a branching node receives no features by convention II, then the features of the next labeled node down the tree will automatically percolate to the unlabeled node;
- **Convention IV**: If two stems are sisters[8], then the features of the right-hand stem will percolate to the branching node dominating the stems.

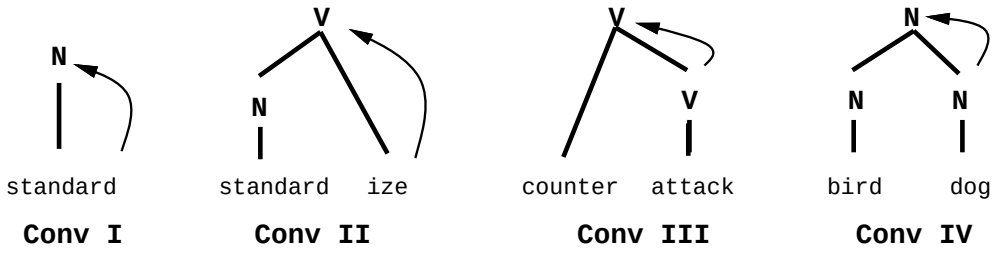


Figure 1: Feature Percolation Conventions

The ALP stipulates the conditions a compound must meet in order to be recognized as well-formed:

1. In the configuration $[\ ]_{\{V|P\}}\ [\ ]_{\alpha}$ or $[\ ]_{\alpha}\ [\ ]_{\{V|P\}}$ where α ranges over all categories, {V|P} must be able to link all internal arguments;
2. If α is *free*[9] in a compound which also contains an argument-taking stem, then α must be interpretable as a semantic argument (i.e. *instrumental, locative, manner, etc*) of the argument-taking stem.

The ALP therefore predicts that ***put-box* is incorrect, since the argument structure of *put* requires both a *theme* and a *location* argument and so the latter

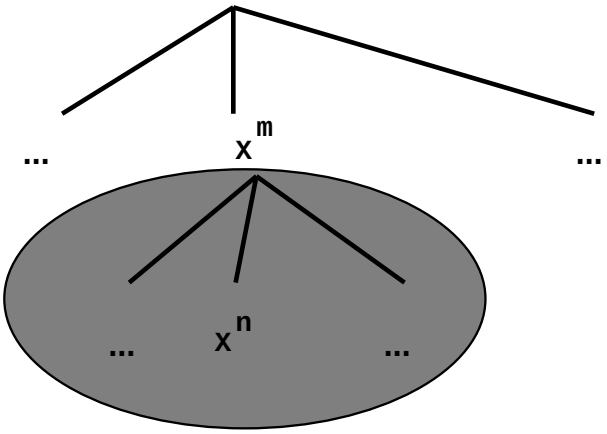
[6] a non-branching node is a node with only one daughter.
[7] a branching node is a node which has at least two daughters.
[8] i.e. they form a compound.
[9] i.e. that is, if it is not used as an internal argument.

internal argument will not be linked to any term. The ALP also allows us to predict the underlying semantic relations of a given compound [Bou92b, §II]. For example, *truck-driver* necessarily means *someone who drives trucks* (rather than *someone who drives (cars) in/under/near ... trucks*) since the argument structure of *drive* contains a *theme* internal argument which will necessarily be filled by *truck*. Finally, a compound verb like *spoon-feed* necessarily means *feed someone with a spoon* (rather than *feed a spoon to someone* or *feed a spoon*) since the FPC predict that the head-term *feed* will transmit its argument structure to the dominating node (and therefore link its theme argument outside the compound). This leaves the left-hand term free and so *spoon* will be read as an external argument of *feed*, in this case as *instrument*.

3.2.3 The work of Selkirk

Selkirk [Sel82] proposes a semantic analysis of compounds based on the *grammatical functions* of the terms. She bases her approach on the *Lexical Functional Grammars* developed in [BK81], which allows her to predict not only the semantic but also the syntactic relation holding between the terms of a compound.

A *grammatical function* (e.g.: subject, direct object, etc) corresponds to the syntactic role a term will be assigned in a sentence. This role is specified for each term in its lexical entry. In Selkirk's analysis, the *First Order Projection Condition* (FOPC) stipulates that all non-subject arguments of a lexical category X^i must be satisfied within the first order projection of X^i, i.e. in the grey area in figure 2.



The argument structure of X^n must be satisfied in the grey area.

Figure 2: Selkirk's FOPC

A compound containing more than one non-subject argument is therefore necessarily syntactically incorrect (e.g.: **baby toy handing* for "the handing of toys to babies").

Selkirk proposes the following context-free grammar for the syntactic analysis of compounds:

$$\begin{array}{lcllllll} N & \longrightarrow & (N & | & A & | & V & | & P) & N \\ A & \longrightarrow & (N & | & A & | & P) & A \\ V & \longrightarrow & (N^{10} & | & P) & V \end{array}$$

These simple rewrite rules cover all the various types of compounds (verbal and non-verbal, right-headed, left-headed, headless).

Specific semantic rules must be used for headless compounds (e.g.: *cutthroat, pickpocket, scarecrow, daredevil* for compounds of the form $[V-N]_N$; *redhead, longlegs, heavyweight* for compounds of the form $[A-V]_N$) and for left-headed compounds (e.g.: *sit-in, runaway, pushover* for compounds of the form $[V-P]_N$ and *worn out, laid off, tuned in* for compounds of the form $[V-P]_A$).

As for right-headed compounds, Selkirk treats the semantic analysis of verbal compounds separately from the semantic analysis of other types of compounds. As for non-verbal compounds, any relationship is possible between the two terms, although a certain number of subsets of relationships can be distinguished (cf. Downing). The semantic relation holding between the terms of a non-verbal compound therefore cannot be predicted with any precision. The semantic analysis of verbal compounds can be made using the grammatical functions specified in the lexical entries of the terms making up the compound. For instance, {*guerilla* | *child* | *Aztec*}*-constructed shelter* can be analysed as *shelter constructed by guerillas, children, Aztecs* by assigning the thematic role of *agent* to the term which is not in the head position. On the other hand, *often, home, factory* cannot be *agents* in {*often* | *home* | *factory*}*-constructed shelter*, but nonetheless these compounds are correct since no argument is specified in the lexical entry of *constructed*[11]; therefore the FOPC is satisfied.

Some compounds belong to both categories (*verbal* and *nonverbal*). For example, *tree-eater* can mean *eater of trees* where *tree* is the *theme* of *eater*, but it can also mean *eater in trees* where *tree* is a *locative* specifier of the verbal term. There may therefore be two distinct semantic analyses for a single syntactic analysis.

3.2.4 The working method chosen

We think it necessary to make the following choices:

- We choose to adopt Downing's analysis of head nouns for compounds of the form N-N;

[10] This category is not present in Selkirk's grammar. We have added it because the author gives an example of a compound of the form N-V later on in her text.

[11] Remember that the *subject* function is a necessary component of sentences but not of compounds.

- We use Selkirk's FOPC for the analysis of verbal compounds;
- The pair FPC-ALP is used to analyse compounds of the form N-V and V-N.

Remark: As we re-examine these choices, we realize that the compounds of the form N-V<suffix> where the relation between the terms is not a thematic one (e.g.: *party-drinker* where *party* is a *locative* term for *drink*, or *night-fishing* where night is *temporal*, etc) are not covered, nor are compounds of the form V<suffix>-N (e.g.: killer-shark).

Although some modifications of these theories are necessary, they still form the basis of the linguistic model used in our system.

3.3 Discussion of the linguistic theories dealing with compounding

3.3.1 Downing: N-N compounds

Downing's analysis only covers compounds of the form N-N. Her theory is based solely on the semantic features of the two terms to predict whether a compound is correct or not and its meaning.

Advantage:

- Downing is the only linguist (among those we examined) who deals in sufficient detail with the problem of analysing compounds of the form N_1-N_2. Most other studies propose only a general definition like: a N_2 *in some relation with a* N_1.

Limits:

- Given the absence of any explicit information as to a possible predicate in the semantic relation underlying compounds of the form N-N[12], the possible meanings of this type of compound can only be obtained with a certain degree of approximation, depending on the features of the head noun.
- This system does not always allow us to obtain the desired result. For instance, *bird-dog* may produce *a dog which looks like a bird* whereas an English speaker would produce *a dog which hunts birds*.

[12] The predicate is only implicitly suggested by the semantic features of the terms of the compound (e.g.: *bird-dog* means *a dog which hunts birds*. The predicate *hunt* is implicit in the meaning of the compound.).

3.3.2 Lieber: primary compounds

Lieber's theory is made up of two independent principles: the FPC and the ALP. These two principles allow her to determine whether a compound is correct or not and what its meaning will be.

Advantages:

- Lieber stresses the analogy existing between the lexicon and the base component of a grammar, that is between the semantic structure of compounds and the syntax of sentences. In fact, much of the originality of her study lies in her use of principles traditionally applied to the syntax of sentences (i.e. the argument structure of a verb must be satisfied in the sentence: **Sue hits* is incorrect since *hit* requires a DO; **The elephant disappeared the frog* is incorrect since *disappear* is intransitive and does not allow a DO) to analyse the semantic relations underlying compounds.

- This theory explains why compounds of the form N-N are so productive as compared with compounds in which one of the terms has an argument structure to satisfy. As Downing independently shows, N-N compounds are almost always possible (this configuration does not exist in the ALP, so it will never be rejected), whereas compounds like **put-box* which contain a term with an argument structure (*put*) may be rejected by the ALP since they cannot link a proper argument.

- These principles generate the different possible meanings of a synthetic compound. These meanings correspond to the two possible syntactic derivations: (N-V)er and N-(Ver). For example, *truck-driver* can be analysed as *(truck-drive)er*, in which case it will mean *a driver of trucks*. But it can also be recognized syntactically as *truck-(drive-er)*; the semantic analysis corresponding to this syntactic analysis is the same as the one proposed by Downing: *a driver in some relation with a truck*, for instance, *a driver owning a truck, a driver wearing a shirt with a truck on it, etc.*

- The main advantage of this theory is no doubt the fact that Lieber claims she can analyse all the different types of compounds (primary and synthetic) with these two principles.

Limits:

- Lieber doesn't really discuss compounds of the form N-N. She simply says that they are always possible and that their meaning depends too much on contextual factors to be calculated with any degree of precision. The author admits that there may be distinct subsets.

- Lieber proposes a semantic analysis of synthetic compounds based on a syntactic decomposition which does not respect the rules of English. For

instance, as mentioned above, *truck-driver* will be analysed syntactically in two different ways: [*truck-drive*]$_V$ *er* and *truck*-[*driver*]$_N$ thereby permitting two distinct semantic analyses. But the first syntactic analysis is incorrect. Some linguists claim that the compounds of the form N-V are syntactically correct only if the noun is not an internal argument of the verb. So *spoon-feed* is correct since *spoon* corresponds to an external *instrument* argument of the verb *feed*, whereas *[*truck-drive*]$_V$ is incorrect since *truck* is not allowed to be the *theme*.

- Some linguists claim that a compound like *truck-driver* has only one meaning: *someone who drives trucks*. The second meaning obtained with Lieber's principles would therefore be incorrect.
- The synthetic compounds of the form V<affix>-N like *killer-shark* are not covered by Lieber's theory.
- The ALP gives no precision about the level in which the argument structure of a term of a compound must be satisfied: must it be satisfied at the level of the node dominating the compound (i.e. in the grey area in the figure 2) or upper (i.e. at the level of one ancestor of x^m) ?

These limits explain why we have chosen to restrict the application of Lieber's principles to the analysis of primary compounds. Selkirk's FOPC is more precisely defined than the ALP for verbal compounds.

3.3.3 Selkirk: verbal compounds

Selkirk defines the FOPC to predict the correctness and the meaning of verbal compounds.

Advantage:

- Selkirk's theory is very precise for verbal compounds. The FOPC states that the argument structure must be satisfied under the mother of the node representing the derivation V<affix>.

Limit:

- Selkirk only considers the verbal compounds. She claims that they are the only compounds for which a precise meaning can be calculated. For the other compounds, she says that almost every relation is possible between the two components.

3.3.4 Unsolved problems

- Some linguists consider certain noun phrases composed of an adjective and a noun like compounds while others consider them as noun phrases (e.g.: *green-ladies, apprentice-welders*). The difference between both analyses can only be obtained using the tonic accent.

- There is no bijective relationship between the set of definitions and that of compounds. A definition of the form *someone who <predicate>s <object>s* can be generated by two different compounds: *<predicate>-<object>* or *<object>-<predicate>er*. For instance, *someone who washes windows* yields *window-washer*, while *someone who cuts throats* yields *cut-throat*. We have as yet found no way to determine which compound should be produced.

3.4 Solutions adopted

This section summarizes the different choices we have made and explains these choices. First, we present the problems we have not been able to solve and have had to abandon. Then we describe the different extensions we have added to the linguistic theories which the methodology is based on. We conclude this section by presenting the general algorithms developed from the augmented linguistic theories.

3.4.1 Unsolved problems

The analysis of left-headed compounds and headless compounds has been abandoned. This restriction is relatively unimportant because these compounds are not very productive as compared with the other types of compounds. However, this decision not to treat left-headed and headless compounds has led us to abandon compounds of the form V-N as well. As a matter of fact, most of these compounds are headless. Moreover, this is not a productive category of compounds and the meaning of its members is not always predictable (e.g.: *throat* is the theme of *cut* in the compound *cut-throat*, but *pocket* is a *locative* specifier for *pick* in *pick-pocket*).

3.4.2 Extensions

Two improvements have been made: one consists in generalising Lieber's FPC and the other concerns an extension of Selkirk's FOPC.

- The only affixes generally thought to possess a lexical category are suffixes. Prefixes do not change the lexical category of their stem. Therefore, only the features of the right-hand item can percolate up to the dominating node, and we can modify the FPC and transform them into FPC':

  FPC': all features of the right-hand term will percolate to the dominating branching node[13].

- Selkirk's FOPC only covers compounds of the form V<suffix>-N, whose head corresponds to a suffixed verb, if there is a thematic relation between

[13] This principle applies only to right-headed compounds. The specific rules for headless compounds or left-headed compounds must still be determined.

the verb and the noun. This mechanism could be applied to right-headed compounds of this type (e.g.: *killer-shark, worker-bee*) by adding the following principle:

In a right-headed compound of the form V<suffix>-N, the verb may not have an internal argument (since it could never be satisfied).

Now, the verb *kill* is sometimes used without a theme. It therefore seems clear that some verbs are *optional-transitives*. This theory is confirmed by most dictionaries, which have two entries for a given verb: one which indicates that the verb requires a theme (or DO) and the other which indicates that it does not require one.

The FOPC does not cover right-headed compounds of the form N-V<suffix> when the relation between the noun and the verb is non-thematic. However, these compounds could be covered by reformulating the FOPC as follows:

In the configuration N-V<suffix>, if the verb has an internal argument which must be satisfied, then the noun must correspond to this argument (e.g.: in *truck-driver*, *truck* must be the *theme* of *drive*)[14]. If the verb has no internal argument to link, then the noun must be interpreted as an external argument of the verb (e.g.: *party-drinker*: *party* is a *locative* specifier of *drink*).

The FOPC' (i.e. the FOPC with the above two extensions) allows us to analyse all right-headed compounds of the form N-V<suffix> and V<suffix>-N.

3.4.3 Algorithms

The aim of this section is to define the general algorithms allowing us to specify the correctness and the meaning of a compound. These algorithms are based on the FPC' and on Lieber's ALP for compounds of the form N-V, on the FOPC' for synthetic compounds and on Downing's ideas for N-N compounds.

The following table presents all the various syntactic configurations covered by the algorithms.

[14] Again, this is predictable from Selkirk's FOPC.

Configuration α – β					
α		β			
		unsuffixed		suffixed	
		exists	does not exist	exists	does not exist
unsuffixed	exists	$\emptyset$	ABANDONED	$\emptyset$	FAILURE
	does not exist	*spoon-feed*	*bird-dog*	*truck-driver*	*?truck-worker*
suffixed	exists	$\emptyset$	FAILURE	$\emptyset$	FAILURE
	does not exist	?	*swimming-pool*	*racing-driver*	*?swimming-worker*

Explanation of the table:
exists means that there is an argument structure for this term (i.e. it requires internal arguments (e.g.: a *theme* to link)).
suffixed means that a term must link its arguments at this stage in the derivation. In other words, the term derives from a rule of lexical redundancy: it is of the form <predicate><suffix>.

The examination of the various possible combinations leads us to write the following algorithms:

Algorithm 1 *Determining whether a compound is acceptable*

A compound is inacceptable if . . .
config=N-N →
the relation between the two terms is one of the unacceptable relations given by Downing.
config=N-V →
the noun is an internal argument of the verb.
config=N-V<suffix> →
the verb has an internal argument and the noun cannot satisfy this argument.
config=V<suffix>-N →
the verb has an internal argument (that is why some verbs must be marked optional-transitives).
EndAlgo

The above algorithm allows us to determine whether a compound is syntactically acceptable (according to Downing's principles, the FOPC' and the pair FPC'-ALP). We must now define the meaning of a compound:

Algorithm 2 *Determining the meaning of a syntactically correct compound*

The relation between the terms of a compound is:. . .
config=N-N →
a semantic relation defined by Downing.
config=N-V →

a semantic relation: manner, location, temporal, etc
config=N-V<suffix> →
either a thematic relation if the verb takes an internal argument,
or a semantic relation if no internal argument is specified.
config=V<suffix>-N →
agent/instrument thematic role
EndAlgo

4 Knowledge representation in the ITS

4.1 Expert Knowledge

Expertise corresponds to the expert's ability to generate correct combinations of terms and affixes corresponding to a given semantic relationship and, conversely, to analyse a given surface configuration in terms of the underlying semantic relation. This knowledge can be represented as follows:

Conversion of a compound into a definition		
example	compound	definition
bird-dog	<object>-<subject>	*a <subject> which is in some relation with a <object>*[15]
swimming-pool	<predicate>ing-<subject>	*a <subject> which is used for <predicate>ing*
worker-bee	<predicate>er-<subject>	*a <subject> who <predicate>s*
data-processing	<object>-<predicate>ing	*the act of <predicate>ing <object>s*
window-washer	<object>-<predicate>er	*someone who <predicate>s <object>s*

This initial table should be read as follows: thanks to the linguistic model chosen, if a compound like window-washer is analysed as an object (window) plus a predicate (wash) plus a suffix (er), then the corresponding definition (or semantic relation) will be *someone who <predicate>s <object>s*, that is someone who washes windows.

NB: These definitions are given in a rough form here. For instance, <object> will only take a plural suffix *s* if the <object> in question has the feature *countable*; the relative pronoun will be *who* or *which* depending on the value of the *human* feature of the <subject>; the determiner will be *a* or *an* depending on the morphology of the noun it precedes, and so on.

[15] This definition can be more specified using Downing's method. Several definitions with decreasing order of probability are obtained.

Conversion of a definition into a compound		
definition	compound	example
a <*subject*> {*which, etc*} *is used for* <*predicate*>*ing*	<predicate>ing-<subject>	*swimming-pool*
a <*subject*> {*which, etc*} <*predicate*>*s*	<predicate>er-<subject>	*worker-bee*
a <*subject*> {*which, etc*} <*predicate*>*s* <*object*>*s*	<object>-<subject>	*bird-dog*
someone {*which, etc*} <*predicate*>*s* <*object*>*s*	<object>-<predicate>er	*window-washer*
the act of <*predicate*>*ing* <*object*>*s*	<object>-<predicate>ing	*data-processing*

This table should be read as follows: thanks to the linguistic model chosen, if the definition is analysed, for instance, as *someone who* <*predicate*>*s* <*object*>*s* (someone who washes windows), then the correct compound corresponding to this definition will be <object>-<predicate>er, i.e. window-washer.

Some information is inevitably lost in the process of converting the definition into a compound, notably the predicate (cf. figure 3). This loss of information explains why compounds of the form N-N only implicitly express the relation holding between the subject and the object.

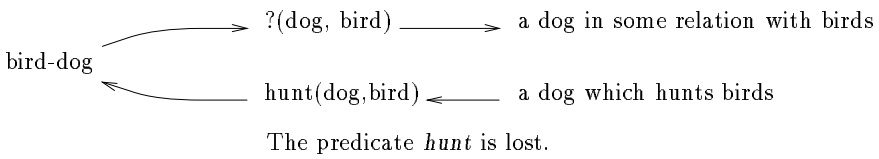


Figure 3: Example of loss of information

This representation allows us to determine what compound corresponds to a given definition as well as what definition corresponds to a given compound. The English speaker's (idealised) expertise is thus completely represented for the types of compounds chosen.

4.2 Student knowledge

The format used to represent expert knowledge can also be used to define the state of student knowledge at any given moment. The tests described in [Bou92b] and [Bou92a] indicate that some student errors are more frequent than others. These errors are presented in the following table[16]:

[16] These logical representations express the illicit patterns used by the students based on an initial logical representation of the form: <predicate>(<subject>,<object>).

Errors made in the conversion of a definition into a compound		
incorrect logical representation	example	
	incorrect compound	definition
<subject>-<object>	**dog-bird*	*a dog which hunts birds*
<subject>-<predicate><affix>	**bee-worker*	*a bee which works*
<subject>-<object>-<predicate><affix>	**dog-bird-hunter*	*a dog which hunts birds*
<object>-<subject>-<predicate><affix>	**bird-dog-hunter*	*a dog which hunts birds*
<object>-<predicate><affix>-<subject>	**window-washer-man*	*a man who washes windows*
<subject>-<predicate><affix>-<object>	**dog-hunting-bird*	*a dog which hunts birds*
<predicate><affix>-<subject>-<object>	**hunting-dog-bird*	*a dog which hunts birds*
<predicate><affix>-<object>-<subject>	**hunting-bird-dog*	*a dog which hunts birds*

This table represents a generic model of student errors, that is, the set of erroneous rules which students may apply. However, the given representation may be used to construct the model of a particular student. We see that the same mechanism can be used to interpret and classify the types of errors made by each student as were used to build the expert model. Moreover, we can relate each category of errors to a certain type of correction. This allows us to offer each student the most appropriate explanations and exercises.

5 Computer implementation

We have implemented our expert model of Compounds on a SUN4 station using the PrologII/Mali language. The syntactic and semantic analysis of the compounds and the definitions is made with a logical grammar. The entry for each word in the lexicon contains features specifying the morphological, syntactic and semantic information needed. These features, which are necessary for our methodology (see for instance the analysis of N-N compounds which uses the semantic features of the head noun), are represented as tree structures using feature descriptors (cf. [Seb89] and [Seb92]).

The syntactic analysis recognises the structure of the definitions and of the compounds and checks whether the items recognised are correct with the aid of filters which analyse the features specified for each term in the lexicon. The following example shows the application of a rule to analyse a definition.

a <subject> {*which, who*} *is used for* <predicate>*ing*:
ns np Det: *a*
Noun: <subject>
rel RelPro: {*which, who*}
Aux: *is*
VerbPaP: *used*
pp Prep: *for*
VerbPrP: <predicate>*ing*

{ human(np) ≡ human(rel) ∧ morpho(np) ≡ morpho(rel) ∧ ssCat(Verb) ≡ ssCat(pp) ∧ int(Verb) ≡ do(No) ido(No)[17] ∧ nb(Noun) ≡ Sing }

This rule stipulates that the definition *a pool which is used for swimming* can be analysed as a nominal syntagme (ns) made up of a noun phrase (np) and a relative (rel). The noun phrase includes a determiner (Det) and a noun. The relative is made up of a relative pronoun (RelPro), of the auxiliary verb *be* in the third person, present tense form, of the verb *use* in the past participle form and of a prepositional phrase (pp). The latter is composed of a preposition (Prep) followed by a present participle verb (VerbPrP). For this analysis to be accepted, there are several other conditions which must be met:

1. the value of the feature human of the noun phrase must unify with the value of the feature human of the relative pronoun,
2. the noun phrase must agree in number and person (morpho) with the relative pronoun,
3. the subcategorisation of the verb must correspond to that of the prepositional phrase.

The semantic analysis allows us to calculate the meaning of the definitions and of the compounds. This meaning is represented by a logical formula. The rules are of the following form:

<mother> → <list of daughters>
features ← <specification of mother's features>
logical representation ← <calculation of the logical representation of the mother>

This rule describes how we calculate the logical representation of the mother in terms of the logical representation of the daughters (based on Downing, the ALP and the FOPC') as well as feature percolation in the syntactic tree (following the FPC'). For instance, the rule below shows how we calculate the meaning of a compound of the form N-V.

Verb ⟶ Noun Verb
features ← features(Verb)
logical representation ← FPC' and ALP

This rule means that the features of a compound verb are obtained from the features of the verbal constituent and that the logical representation of this compound is based on the FPC' and the ALP.

The time of calculation needed to obtain a compound or a definition is between 0.2 and 0.6 second.

[17] In this case, the verb is necessarily intransitive.

6 Conclusions and future perspectives

The ITS for the English compounding process which we are developing is based on a solid linguistic theoretic model. We have used and adapted the work of Downing, Lieber and Selkirk in order to cover all the main types of compounds. The program currently treats the most common types of compounds in English but the linguistic methods used allow us to treat all of the various types of compounds (i.e. compound adjectives, compounds with non-head prepositions, etc). The answer times obtained for the syntactic and semantic analysis are satisfying. What's more, the same representation is used to codify the expert's knowledge and the (correct and incorrect) knowledge of the student. This similarity of treatment allows us to easily compare the student's compounding schemata with those of the expert.

The system contains several weaknesses in its current form. The system's ability to adapt to individual student errors is just beginning to be developed. Moreover, we have had to abandon compounds of the form V-N since they raise a problem we have not yet been able to solve. It should be noted however that these compounds are fairly rare. Finally, we do not consider the phonetic side of using compounds.

Our present and future work aims at developing a model of the student. Basing our work on the generic model which consists of a set of incorrect rules used by students (rules based on actual student errors in tests), we are trying to see what correct and incorrect rules are used by students for various types of exercises. Using statistics for a given student's results we are trying to find regular patterns in the problem solving strategies used in each context. For example, we are trying to determine whether the rules applied are systematically correct or not depending on certain features of the terms of the compounds. Finally, we are attempting to build models for the ability of individual students to learn and to forget.

References

[BK81] J. Bresnan and R. Kaplan. Lexical Functional Grammars: a formal system for grammatical representation. In *The mental representation of grammatical relations*, J. Bresnan Edt, MIT Press, Cambridge, Mass, 1981.

[Blo33] L. Bloomfield. *Language*. G. Allen and Unwin Ltd. London, 1933.

[Bou92a] P. Boucher. L'intelligence artificielle et l'apprentissage des langues : existe-t-il des tuteurs réellement intelligents ? *Cahiers de l'APLIUT*, 45(XI):9–23, 1992.

[Bou92b] P. Boucher. Teaching Compound Nouns in English: an application of the Charlie Brown principle. In *CIEREC/TRAVAUX LXXVI, Université de Saint-Étienne*, pages 33–50, 1992.

[Dow77] P. Downing. On the creation and use of English compound nouns. *Language*, 53(4):810–842, 1977.

[Lie83] R. Lieber. Argument Linking and Compounds in English. *Linguistic Inquiry*, 14(2):251–285, 1983.

[NV88] J.-F. Nicaud and M. Vivet. Les tuteurs intelligents : réalisations et tendances de recherches. *Technique et Science Informatiques*, 7(1):21–45, 1988.

[Seb89] P. Sébillot. *Modélisation de mécanismes de filtrage par arbres et par traits dans les grammaires logiques pour l'analyse automatique du langage naturel.* PhD thesis, INSA-Rennes, jun 1989.

[Seb92] P. Sébillot. A logical-based language for feature specification and transmission control. In *Logic Programming*, A. Voronkov Edt, Lecture Notes in Artificial Intelligence, N^0592, Springer-Verlag, 1992.

[Sel82] E. O. Selkirk. *The syntax of words.* Cambridge Mass., MIT Press, 1982.

[Wil80] E. Williams. Argument Structure and Morphology. *The Linguistic Review*, 1:81–114, 1980.

Unité de recherche INRIA Lorraine, Technôpole de Nancy-Brabois, Campus scientifique,
615 rue de Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105,
78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS
Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
(France)
ISSN 0249-6399