English-Arabic Machine Translation of Nominal Compounds

Zouhair MAALEJ
Regional Institute for Informatics and Telecommunications, Tunisia
Faculty of Letters Manouba, Tunis.

(1994)
(1994). "English-Arabic Machine Translation of Nominal Compounds." In: Pierrette Bouillon & Dominique Estival (eds.), *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*. Geneva: ISSCO, 135-146.

Introductory remarks

In pervading almost all text-types, English nominal compounds fill a great lexical gap in referring and naming, especially in fairly new disciplines such as computer science (e.g., word processor; word processing application), telecommunications (e.g., radio positioning; radio position finding), acoustics (e.g., sound power; sound pressure level), electronics (e.g., valence bond; valence bond method), mechanics (e.g., friction clutch; gear lever) to name only a few. Therefore, we believe that any translation project which does not make provisions to cope with such compounds is likely to be substantially impaired.

If it is true that our dealing with compounds has been motivated by their frequent occurrence in referring and naming, our major reason for considering them is their being a productive lexical composition process capable of generating from simple lexemes an infinite number of complex lexemes. In so doing, compounding widens the range of combinatorial possibilities for simple lexemes, creating new meanings not possible through derivation, which "involves an operation on a single lexeme"(Matthews ,1991: 84). Such productivity may be exemplified by the following combinations of lexemes:

bathroom
bathroom towel
bathroom towel rack
bathroom towel rack designer
bathroom towel rack designer training
bathroom towel rack designer training program (Selkirk, 1982: 15).

In this paper, it is not our intention to investigate machine translation in general or even compounding in English; this has been, and will be, the object of interest of many of you present here. Rather, we propose to describe some of the aspects of our modest experience with the translation of English nominal compounds into Arabic. For translation purposes, we take nominal compounds in the strict sense, i.e., we exclude from consideration compounds made up of adjective-noun and verb-noun sequences such as "public school" and "call girl" respectively. And even among the noun-noun sequences, we have selected only those compounds whose meaning is predictable through a compositional calculation of their constituents. The reason for this selective approach is rather technical: our translation system is designed to allow only for noun sequences to be processed within the structure of the NP automaton.

Neither is it part of our investigation to consider already institutionalized solid compounds such as "earthquake" and "scarecrow", since these do exist in the core dictionary as single lexical entries. According to Lyons (1977: 535), what we propose to study are "syntactic compounds" not "compound lexemes," which might suggest that our paper is only partly relevant to the topic of these study days. The reasons why we have been confined within these limitations are that (i) there is no way for computers to translate compounds whose meaning is not compositional; (ii) because we favour "intelligibility" and "fidelity" (Lewis, 1992: 104), we want our experimental project to yield translation which sounds natural and not, as Didaoui (1993: 157) calls it, "Araglish," i.e., a kind of Arabic which simply copies the English syntax . We hope it has become clearer that the causes are inherent in working with computers on natural language. Before getting into some of the details of our system of understanding of compounds, which is the core of this paper, we feel it quite useful to talk about the following items:

(i) the architecture of the translation module we are using;

(ii) the linguistic background to compounding.

I. The translation module

Computer-assisted translation (C. A. T., for short) has been launched at the Regional Institute for Informatics and Telecommunications (I. R. S. I. T.), Tunisia, in 1989 in collaboration with PC-Linguistics at Texas. The outcome of this cooperation has been TORJOMAN (literally, translator), an experimental project of aid to English-Arabic translation, which is currently run on a 386 PC under Arabic MS DOS. It is equipped with three electronic bilingual dictionaries

(Core - Specialized - Users). Initially, TORJOMAN has been entrusted with simple declarative sentences to translate into Arabic. Further developments involved the treatment of coordination at the level of all syntactic categories, and verbal idioms. Refinements are currently being added, involving the processing of compounds.

The translation module allows two options: translation without recourse to analysis and translation via analysis. Translation without analysis takes place so long as the sentence is proved to include an idiom or a proverb after the Specialized Dictionary has been consulted. In this case, a translation equivalent is provided, signalling the end of the journey for the sentence. In case, however, the sentence is not found to involve anything of the sort, analysis proper will start. Notice that the module, of course, analyses the English sentence and generates its Arabic counterpart.

A. The analysis of the SL sentence

This is taken care of by two types of analysers: a lexico-morphological analyser and a syntactic analyser. The lexico-morphological part determines the lexical properties of the lexical words making up the sentence. Such properties are grammatical category, number, gender, tense, etc. Then the whole thing is submitted to a lexical disambiguation unit, whose job is to make sure that each lexical item is assigned the right specification, scanning it in relation to its neighbours, and according to the grammar rules of English. Once disambiguated, the sentence is passed on to the syntactic analyser, which describes its internal structure, and assigns a function to each of its constituent parts according to its position. It is only then that the sentence is ready for transfer into the TL.

B. The generation of the TL sentence

This step takes full advantage of the morpho-syntactic information made available during analysis, and makes use of the bilingual core dictionary and the rules of Arabic to generate a TL sentence. Once generated, the sentence is submitted to an Arabic conjugator, which makes sure that verb inflections are in accordance with the number and gender of nouns, and that adjectives are in agreement with nouns with regard to number, gender, definiteness, etc. Lastly, a word sequencing device reorders the words in the Arabic sentence according to acceptable sequences, Arabic being basically a VSO language. Before getting

into the technicalities of the translation of compounds, a brief survey of compounding in linguistics is not out of order.

II. Linguistic background to compounding

It is beyond the scope of this paper to present a thoroughgoing descriptive survey of compounding in English. Instead, we will attempt to suggest a working division of the different trends within morphology into transformational and non-transformational. While both views hold it uncontroversial that compounds are made up of a head constituent (usually to the right) preceded by (a) nonhead constituent(s), they disagree, however, about the nature of compounding, i.e., how compounds are generated.

Roughly, the transformational model holds that compounds derive from a deep structure similar to that of a relative clause, where the head noun which is postmodified in deep structure is turned into a premodified head in surface structure, like with "a bird which is black" becoming "a blackbird" (in Matthews, 1991: 87). This model remained uncontested till the early 1970s when it started to be questioned, the reason for this refutation being that it failed to account for all compounds. For instance, "a suicide attempt" cannot be said to derive transformationally from "an attempt which is suicide" or "a suicide which is an attempt", simply because "the semantic relation obtaining between the head constituent and its sister nonhead constituent can vary considerably" (Selkirk, 1982: 22).

Proponents of the non-transformational model, however, claim that compounds are present in deep structure and that "the nonhead constituent [should be] interpreted as an argument of the head" (Selkirk, 1982: 18 - 23), i.e., it should be analysed according to the thematic relations of agent, theme, goal, source, instrument, etc. that it entertains with the head. Quirk et al (1985: 1570), following a lexical-functional approach, proposed "to adopt a mode of presentation which (where applicable) links compounds to sentential or clausal paraphrases." Their proposal, they claim, is supported by the evidence that, like sentences, compounds employ the same lexical categories (of nouns, verbs, adjectives, and adverbs), enjoy the same syntactic functions (of S, V, O, C, A), and involve head and nonhead constituents into a range of (more or less) complex relations.

We are fully aware that this overview of schools of thought about compounding is far from exhaustive and satisfactory for specialists in linguistic theory. Nevertheless, the enterprise we have embarked upon made us (i) simplify

linguistic data, (ii) collapse categories, and (iii) ignore barriers between theories and adopt a more eclectic approach. Accordingly, nominal compounds have been assigned a lexical-cum-syntactic classification. The criteria adopted in devising our system of understanding of English compounds have been observed as regularities in the human practice with translation.

III. Compounds: algorithms of analysis

In this section, two major classes of compounds will be isolated, namely, two-word compounds and multi-word compounds, in spite of Quirk et al's insistence that "compounds usually comprise two bases only, however internally complex each may be"(1985: 1567). The justification for postulating, for instance, that "question answering diagram" is not to be analysed as a nonhead compound (itself including a head) modifying the head ("diagram"), is economy- and efficiency-based, even though running counter to the recursive nature of compounds and at variance with well-established norms. If this compound were not analysed the way we indicated, the computer would have to do the job twice : once for the nonhead and another for the nonhead in relation to its head. Instead, we preferred to consider its constituents separately. This reminds us that "the grammatical formalism that most elegantly describes linguistic competence is not necessarily also best to process a language" (Patten, 1992: 30).

A. Two-word compounds

Maybe, these should have been named "two-noun" compounds, since in our scheme "automatic data processing" is still treated as an adjective premodifying a compound noun, the whole being a kind of complex NP (not to be confused with complex NP in Quirk et al's terminology, which includes a head noun and a relative clause postmodifying it). Quite obviously, this class of compounds consists of N1 and N2, where N1 is nonhead and N2 is head. Within this class, two major scenarios have been envisaged depending on whether or not N2 is a deverbal noun. Notice that we take the deverbativity of nouns to mean a noun deriving from a verb, and assuming any form:

(1) Scenario 1: N2 is not a deverbal noun

In case N2 is not a deverbal noun and if the determiner is "a," "an," or zero, then N1 does not take a definite determiner in Arabic, and the compound is rendered as two nouns annexed (or postfixed), with N1,

which comes second in Arabic, receiving the diacritic feature related to annexation (which consists in the sound [in]) as in:

a dinner party

Haflu 3aè?è?in (1)

(party dinner of)

However, if the determiner is "the," N1 takes the definite determiner "al" (the), changing the feature of annexation into the sound [i] as in:

the dinner party

Haflul 3aè?è?i

(party the dinner)

Where the determiner, on the other hand, is a possessive, N1 takes the possessive marker corresponding to the speaker's pronoun, addressee's pronoun or obviative pronoun, and both N1 and N2 remain in the indefinite like in:

my car key

miftèHu sajja:rati:

(key car my)

Notice that beside their comparative multiplicity, Arabic pronouns are marked at more than a level. Apart from speaker pronouns, which are marked for number only, the others are marked for number (singular -- dual -- plural) and gender. This makes the diacritic feature related to annexation dependent on the type of person used.

Lastly, in case the determiner is a demonstrative, a demonstrative pronoun, agreeing in number and gender with N1, is inserted to the left of N2 in Arabic, and N1 takes the definite article as in:

this car key

miftèHu hèδihis sajja:rati

(key this (feminine) the car)

It should be noted at this level that in case it is not a deverbal noun, N2 is never made definite in the Arabic translation, and that what seems to be governing this subclass of two-word compounds is the type of determiner used, with N1 remaining annexed to N2 independently of the kind of determiner adopted.

(2) Scenario 2: N2 is a deverbal noun

In case N2 is a deverbal noun, the criterion that seems to be governing the translation of this subclass is whether or not N1 is

adjectivalizable(2). Needless to say that in the core dictionary nouns are tagged for deverbativity in Arabic. Once N2 is proved to be a deverbal noun, the automaton checks whether N1 is adjectivalizable and if it is masculine or feminine in Arabic. In case N2 is masculine, and the determiner is "a," "an," then N2 takes no definite article and ends with the sound [jun] as in:

a bank discount
xaSmun bankijjun
(discount bank of (masculine))

But if N2 is feminine and abstract, both N1 and N2 are made definite and N1 takes a femininization ending marker as in:

machine translation
aTTarZamal?èlijja
(the translation the mechanical)

On the other hand, if N1 is not adjectivalizable and not abstract, N1 is pluralized, taking a definite article, and N2 translates as a verbal noun ("maSdar" in Arabic) as in:

letter writing
kitèbaturrasè?ili
(writing (of) the letters)

Pluralization takes place regardless of the form N2 may assume (e.g., nouns in -ing, -er, -ion). This homogeneity in the Arabic translation in the face of a multiplicity of derivational suffixes in English, could be accounted for by the fact that (i) most of these compounds in English talk of occupations, (ii) in Arabic only one form (the verbal noun or "maSdar") is available. Such derivational suffixes may be exemplified by "house building," "bell founder," "business administration." In order for pluralization to obtain, the computer uses an algorithm for the conversion of singular nouns into the plural even though the plural of nouns exists in the dictionary, which is a further complication.

B. Multi-word compounds

Before dealing with this class of compounds, a few words about compounds recognition are not out of place. The parser being essentially a non-deterministic Augmented Transition Network(ATN), the structure of the NP can cope, at least theoretically, with an infinite number of nouns in both nominative

and accusative positions. Recognition proper takes place when the parser moves to the structure of the VP.

We consider a stretch of language a multi-word compound if it satisfies one requirement: it should include at least three constituents, of which only the medial one may be a deverbal noun or an adjective (or any equivalent prenominal modifier). This class of compounds has been divided into two major categories depending on whether their medial constituent is a deverbal noun or an adjectival passive participle. For the sake of ease of reference, the former are called "deverbal compounds" and the latter "passive compounds."

1. Deverbal compounds

These compounds are named after their deverbal component for the following reasons:

(i) the deverbal noun facilitates recognition within the structure of the NP, in that it may appear in a finite set of forms determined by suffixes such as -ing, -ion, -er, and -ment. With a bit of backtracking, the compound is thus localized and processed;

(ii) because of its being a deverbal noun, N2 follows most of the steps implemented in the translation of two-word compounds, with a few modifications;

(iii) since the deverbal noun originates in a verb, the computer has to track its verbal counterpart to see whether it is a self-transitive verb or an intransitive (prepositional) one. In case it is of the latter kind, which preposition should follow the deverbal noun has to be determined by appeal to the core dictionary.

In dealing with this category of compounds, we realized how much Quirk et al's (1985: 1570) proposal that compounds could be linked to "sentential or clausal paraphrases" is attractive. Their suggestion is all the more invaluable that the structure of Arabic lends itself to this kind of analysis, in that it tends, so to speak, to imitate our thought processes, leaving nothing implicit or ambiguous among the constituents of the compound. This kind of structural elaboration (or what the French call "étoffement") is at the root of the translation of this class of compounds. This method of translation, which Newmark (1981: 41) calls "cognitive," consists in "explicat[ing] the relation of all multiple compounds (e. g., 'data acquisition control system': system to control the acquiring of data')." That is exactly what happens in Arabic.

Accordingly, instructions to the computer consist in the following: Given the sequence N1 + N2 + N3, where N2 is a deverbal noun in -ing, -ion, -ment, or -er:

1. If N1 is uncountable in Arabic, translate N3 as an indefinite noun; turn N2 into a verbal noun (or "maSdar"), prefixing it with [li] of purpose; and keep N1 as a singular definite noun after selecting the preposition that fits N2 and inserting it to the right of N1, if applicable:

an oil prospection firm
è*arikatun littanqi:bi 3aninnafTi*
(firm for prospection about oil)

2. If N2 is countable in Arabic, translate N3 in the indefinite; turn N2 into a verbal noun, prefixing it with [li] of purpose; and pluralize N1, making it definite, looking for the right preposition, if need be, that works with N2, and insert it to the right of N1 as in:

tumour killing cells
xalèjè lilqaDa?i 3alal awra:mi
(cells for killing on the tumours)

Notice that the structure of compounds may be optionally made more complex by the addition of premodifiers as in:

automatic picture (N1) transmission (N2) system (N3)
niDa:m utumatiki li?irsèlisSuwar
(system automatic for transmitting the pictures)

automatic data (N1) processing (N2) auxiliary equipment (N3)
aZhiza musè3ida lilmu3a:la□al utumatikijja lilma3lumèt
(tools helping for processing the automatic for the data)

However, it should be noticed that while the first occurrence of "automatic" is a prenominal modifier to N3, the second relates to N2, which occasions serious semantic and even pragmatic difficulties the computer is not equipped to cope with.

2. Passive compounds

These compounds are quite recognizable since they include N1 + Adj (or an equivalent passive participle) + N2, where the adjective derives from a verb, and where, semantically, N1 is instrumental agent, and N2 is affected by the action. Passive compounds are of two kinds:

(i) If N1 is uncountable and if the determiner is "a," "an," or zero, the computer is instructed to translate N2 as an indefinite noun; to look for the verb corresponding to the adjective and conjugate it in the passive participle form, making it agree with N2 in number, gender and definiteness; then to translate N1 as a definite noun, prefixing it with instrumental [bi] ("with ") as in:

a gas-cooled reactor

mufè3il mubarrad bil Ra:z

(reactor cooled by the gas)

If, however, the determiner is definite and N2 is of an irrational type, the computer will be instructed to make N2 and the passive participle in the definite; to conjugate the passive participle simply appending the suitable femininization marker to it; and add the definite [al] to N1, prefixing it with instrumental [bi] as in:

rule-governed transformations

taHwi:lèt munaDDama bilqawè3id

(transformations governed by the rules)

(ii) In case N1 is adjectivalizable, given the passive adjective is in -able, and if the determiner is "a," "an," or zero, the computer is made to translate N2 as an indefinite noun; to look for the verb form corresponding to the passive adjective in -able and conjugate it in the passive voice in Arabic, providing necessary agreements with N2; and then to translate N1 either as a specificative in [yan] or as a noun preceded by instrumental [bi] as with:

machine-readable data

ma3lumèt tuqra?u bil ?èlati / ?èlijjan

(data readable by the machine or mechanically)

If, on the other hand, the determiner is definite, the compound has to be transposed into a relative clause, with N2 in the definite and left insertion of the relative pronoun that fits the gender and number of N2 as in:

the machine-readable data.

alma3lumèt allati tuqra?u bil?èlati / ?èlijjan

(the data that [are] read by the machine / mechanically)

Concluding remarks

Working in the field of automatic translation is challenging but also time- and energy-consuming and unrewarding in the short term as results never always match all these investments. Furthermore, working in automatic translation with compounds has enabled to emerge to the surface the often contested assumption, however anachronistic it may be thought of, that meaning may be calculated from the constituents making up a sentence. Well, it seems that this compositional dimension of meaning still has some relevance to computer translation, otherwise there is no way for a machine to have access to semantic information.

The factors governing the translation of compounds into Arabic are so common to linguistic theory that no-one imagined them to be so decisive in such a translation project. Such factors are the notion of transitivity of verbs (too much cherished by Halliday), and the notions of countability, definiteness, abstractness, adjectivalizability, and deverbativity of nouns. It goes without saying that since compounds are essentially noun-based, the factors relevant to their translation have been those related to the semantics of nouns in language.

To finish with this presentation, three concluding remarks relating to compounding and the dictionary will be made. In connexion with the translation of multi-word compounds, it should be mentioned that algorithms have not yet been implemented, for we have been pressed for time putting a marketable product on its feet. With regard to the translation of both classes of compounds, the methods of translation that have been undeliberately followed are the "semantic" method for two-word compounds and the "cognitive" method (a variant on the "semantic") with multi-word compounds. The dictionary, on the other hand, shows two major deficiencies, namely, unsystematic tagging and inexistent tags. With prepositional verbs in the Arabic translation, for instance, not all verbs have been furnished with their prepositions, which makes the treatment of many compounds using such verbs not immediately feasible. Inexistent tags such as those for transitive/intransitive verb distinctions, countable/ uncountable nouns are now more urgent to add.

ENDNOTES

(1) The following is a phonological description of Arabic consonants and vowels:

δ : voiced inter-dental fricative
D: voiced pharyngalized inter-dental fricative
θ: voiceless inter-dental fricative
T: voiceless pharyngalized alveolar stop
S: voiceless pharyngalized alveolar fricative
è: voiceless alveo-palatal fricative
Z: voiced alveo-palatal fricative
?: alveolar trill or flap
q: voiceless uvular stop
x: voiceless uvular fricative
3: voiced pharyngeal fricative
R: voiced uvular fricative
H: voiceless pharyngeal fricative
j: palatal glide
è: mid-low front vowel
a: low vowel
u: high back vowel
i: high front vowel

(2) I owe a great debt to Professor Salem Ghazali, head of the C.A.T. project at I.R.S.I.T.; Miss Lamia Abed, a computer engineer, and our late colleague Chedly Belfalah, for their having been at the origin of the adjectivalizability criterion.

REFERENCES

Butler, C.S. (ed.) (1992). *Computers and Written Texts.* Oxford: B. Blackwell.

Cachia, Pierre (1973). *The Monitor. A Dictionary of Arabic Grammatical Terms.* London and Beirut: Librairie du Liban and Longman.

Didaoui, M. (1993). "Scientific Translation in Arabic." Proceedings of the XIII FIT World Congress, 147 - 162.

Lewis, Derek (1992). "Computers and Translation." In C.S. Butler, ed. *Computers and Written Texts*, 75 -113.

Lyons, John (1977). *Semantics* (vol. II). Cambridge: C.U.P.

Matthews, P.H. (1991). *Morphology*. Cambridge: C.U.P.

Newmark, Peter (1981). *Approaches to Translation.* London: Prentice Hall.

Patten, Terry (1992). "Computers and Natural Language Parsing." In C.S. Butler, ed., *Computers and Written Texts*, 29 - 52.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985). A *Comprehensive Grammar of the English Language.* London and New York:Longman.

Selkirk, Elisabeth O. (1982). *The Syntax of Words*. Cambridge/Mass.: The MIT Press.

** This paper has been published in P. Bouillon & D. Estival (1994) (eds.), Proceedings of the Workshop on *Compound Nouns: Multilingual Aspects of Nominal Composition*. ISSCO: University of Geneva.