Hierarchical NN compounds in a cross-linguistic perspective

Giorgio F. Arcodia, Nicola Grandi & Fabio Montermini

The identification of consistent classes of compounds has been an issue since the research of early Indian grammarians and more recently it has received new attention in the linguistic literature. Starting from Bisetto & Scalise's proposal (2005), namely that compounds may be divided into three classes, each of which may contain both endocentric and exocentric complex words, we shall show that these classes are not discrete, but rather that they constitute the points of a *continuum*. We shall then test the behaviour of compounds belonging to these three classes in fusional languages from the *Standard Average European* area and in languages from the East and South-East Asian region, namely Chinese (isolating) and Japanese (agglutinating), to provide an example from each major morphological type. Our findings are that Bisetto & Scalise's attributive / appositive (henceforth ATAP) compounds and subordinate (henceforth SUB) compounds apparently behave similarly in different languages, but having a phrasal constituent is possibly a unique property of subordinate compounds. As far as coordinate (henceforth CO) compounds are concerned, we shall argue that two subclasses of coordinating compounds should be distinguished, namely "hyperonymic" and "hyponymic" compounds, as they behave in a rather different way[*].

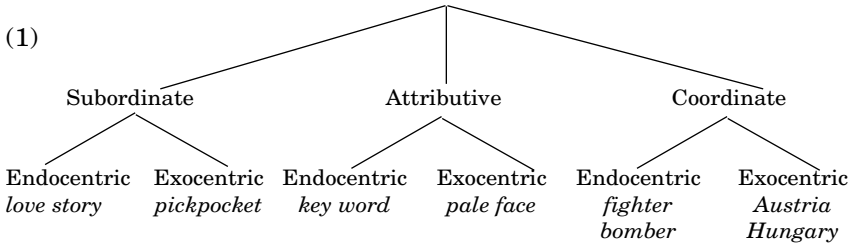
1. Theoretical premises: the classification of compounds

The classification of compound words in the world's languages has received new attention in the past few years. Traditionally, compound classifications were mainly based on the distinction of Sanskrit compounds in (at least) three categories: *dvandva*, *tatpuruṣa* and *bahuvrihi* (cf. e.g. Bloomfield 1933; Benveniste 1974 (1967), among others). Roughly, for dvandva compounds the meaning of the whole is simply the sum of the meanings of the two parts (Eng. *Alsace-Lorraine*); in tatpuruṣa compounds the two members are in a dependency or modification relation (Eng. *love story*); finally, bahuvrihi compounds refer to an entity which is not designated by any of the members (Eng. *redskin*), and correspond to what we would call, in modern terms, exocentric compounds. As already pointed out by Bisetto & Scalise (2005), an unsatisfactory aspect of this classification, and of those which are based on it, is that it does not clearly distinguish semantic and grammatical criteria, such as the relation between the two members from headedness, that is the endocentric-

Rivista di Linguistica 21.1 (2009), pp. 11-33 *(ricevuto nel marzo 2009)*

ity / exocentricity of the whole compound (i.e. the presence / absence of an head). Bisetto & Scalise (2005)[1] instead propose a tripartite distinction of compounds, in which the endocentricity / exocentricity criterion is orthogonal to the semantic classification, and split each of the semantic classes in two (examples are from Bisetto & Scalise 2005):[2]

(1)



(Bisetto & Scalise 2005:326)

The goal of this paper is to reassess Bisetto & Scalise's classification by taking into account both semantic and formal criteria. In view of the linguistic data, we will claim that the tripartition in (1) should be viewed as a continuum scale and not as a universally valid framework in which each compound can be unambiguously placed. As we will see, ATAP compounds are more similar to SUB compounds, in particular if we take into account formal criteria. However, from a purely semantic point of view, the distinction between ATAP and CO compounds is not so clear. Bisetto & Scalise (2005:327) characterize CO compounds as "those formations whose constituents are tied by the conjunction 'and'", and ATAP compounds as formations "where the non-head very often is used somehow metaphorically, expressing an attribute of the head". It would be interesting, consequently, to evaluate if this distinction has a correlate at the level of the structure or of the grammatical properties of compound structure, since the – purely semantico-pragmatic – notion of metaphor can hardly be taken as a solid criterion to establish a linguistic class, should it be the only one. In what follows, we shall take into account only N+N compounds. Let us, for instance, take some examples of CO and ATAP compounds from Bisetto & Scalise (2005:327-328):

(2)	CO COMPOUNDS		ATAP COMPOUNDS	
	Eng.	fighter-bomber	Eng.	ape man
		girlfriend		ghost writer
		actor author		key word
	Sp.	poeta pintor		snail mail
		'poet-painter'		

A first remark is that ATAP compounds fall under the definition of "formations whose constituents are tied by the conjunction 'and'": though in a metaphoric sense, a *ghost writer* is at the same time a writer 'and' a ghost. More precisely, the word denotes a man sharing some peculiar characteristics with a ghost (e.g., in this case, invisibility, though with a figurative meaning). What we suspect is that the characterization of an 'and' relationship between the two members of a compound as 'literal' or 'metaphorical' (and consequently of a compound as CO or ATAP) is more a matter of pragmatics than of grammar. Actually, the possibility for two nouns of being coordinated in a literal sense is limited to nouns having a large number of common semantic features, and probably limited to some very specific semantic classes of nouns. If we consider the CO compounds in (2) (and, generally speaking, the examples of typical CO compounds proposed in the literature), three of them are exclusively composed of [+human] nouns, and two of these (*actor author* and *poeta pintor*) are nouns designating a profession, made up themselves of two nouns designating a profession. The first compound in column 1 in (2) (*fighter bomber*) is made up of two nouns denoting an instrument.[3] Of course, a precise characterization of the semantic types more likely to be found in CO compounding would deserve a specific study, and we can only sketch the picture here. We want to observe, however, that the CO / ATAP compound distinction cannot be based solely on the notion of metaphor, and that only the observation of systematic differences in their construction and in their behaviour would justify such a distinction. This holds, however, only for hyperonymic CO compounds; we shall introduce the distinction between hyperonymic and hyponymic CO compounds in section 2.

In order to assess the pertinence of the three classes identified by Bisetto & Scalise (2005), we will take into consideration three criteria: i) the possibility of observing a recursive structure within a compound; ii) the possibility of a compound containing an inflected member or a full syntactic unit; iii) the possibility of an anaphoric element being coreferent with one of the members of a compound (or, in other words, the possibility for a compound to violate anaphoric islandhood,

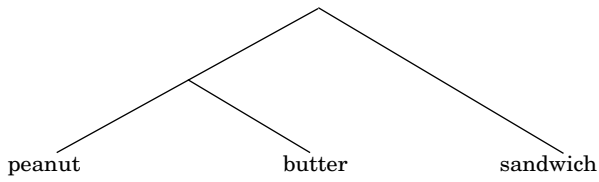
cf. Postal 1969). Clearly, the three criteria given above are crucial for identifying the nature of the elements which constitute a compound. In a strictly lexeme-based morphology, a compound should be considered as a lexeme made up of two (or more) lexemes. As we will see, such a simple definition cannot be mechanically applied to all types of compounds. Rather, the nature of its members can also be a criterion for the characterization of a compound as CO, ATAP or SUB. Eventually, we will be led to discuss how the above points should be considered in relationship to the morphology / syntax distinction, and the place of compounding within this distinction.

What we call recursivity is the possibility of a compound containing another compound. The compounds in (3) illustrate this type of constructions:

(3) Eng. peanut butter sandwich
It. centro assistenza pneumatici
'tyre assistance center'

Recursivity should be distinguished from the inclusion of syntactic structures within compounds (see below). Even if the non-head of the compounds in (3) is a complex structure, it cannot be considered as being constructed by syntax, but rather as a compound itself:

(4)



As observed by some scholars (cf. Bisetto 2004:42), apparently only subordinate compounds can have another compound in non-head position.[4]

As far as the second parameter is concerned (inclusion of an inflected constituent and / or of a syntactic structure), we should first observe that, in those languages in which inflection is overtly marked, the head element of a compound is generally the one which bears inflectional markers:

(5) It. gli squali$_{PL}$ balena ?gli squali$_{PL}$ balene$_{PL}$
'whale sharks'
le donne$_{PL}$ oggetto ?le donne$_{PL}$ oggetti$_{PL}$
'object women'

The examples in (5) should probably be classified as ATAP compounds: the referent of the whole is an X (the head noun) which shares some of the characteristics of Y (the non-head), not an X which is also a Y in the literal sense. With CO compounds, on the other hand, the two elements are more likely to covary:[5]

(6) It. attore-regista attori$_{PL}$-registi$_{PL}$
'actor-director'
Rus. ženščina vrač ženščiny$_{GEN}$ vrača$_{GEN}$
'woman doctor'

SUB compounds too only bear inflectional markers on the head. The non-head constituent may also appear as inflected on the surface. However, in this case the scope of the inflection is not the whole compound, but only the noun which bears it. In (7) we give two examples of Italian SUB compounds:

(7) It. capo$_{SG}$gruppo capi$_{PL}$gruppo
leader+group
'group leader'
It. ufficio$_{SG}$ informazioni uffici$_{PL}$ informazioni
office+information
'information desk'

The compound *capogruppo* can be inflected only on its head (*capo*). On the other hand, in *ufficio informazioni* the non-head (*informazioni*) bears a plural marker, independently of the number of the head. In this case, the scope of the plural inflection is only the non-head member of the compound. This property can be connected to another property of some compounds, namely the possibility of having a syntactic construction in non-head position.[6] It is often claimed that this possibility is limited to lexicalized phrases. Actually, this does not seem to be the case, as the examples in (8) suggest (cf. also Lieber & Scalise 2006:10-12 for a discussion):

(8) It. scaldachiodi di sommergibile
'submarine nail heater'
[*la Repubblica*, June 5, 2008]
It. articoli […] ammazzaindagini sulla mafia
'mafia investigation killing articles'

The last criterion we will take into account is the possibility for an anaphoric element to be coreferent with one of the elements of a compound. Morphologically complex words are generally considered to be 'anaphoric islands'; several works on anaphoric islandhood,[7] however, have pointed out that, among complex words, compounds are the most likely to violate it:

(9) Eng. Although *cocaine*$_{i}$ use is down, the number of people using *it*$_{i}$ routinely has increased
(Ward *et al.* 1991:454)
It. Era disponibile a diventare capo*gruppo*$_{i}$ di *quello*$_{i}$ da noi appena costituito
'He was available for being the leader of the group we just constituted'
(Montermini 2006:139)
Jp. *kenkyuu*$_{i}$-shitara, *sore*$_{i}$ ga hyooka sareta
'after I had research$_{i}$ed it$_{i}$ received appreciation'
(Lombardi Vallauri 2005:324)

The three criteria considered strongly suggest that, at least in some cases, one of the elements of a compound is 'more' than a lexeme form, since it preserves, at least partially, some of its syntactic characteristics, as well as its referential capacity. This property of compounds, clearly, poses a challenge to a strictly modular view of grammar, in which every morphological rule applies before every syntactic rule.

In the next two sections of this work we will test these criteria on NN compounds from languages belonging to two different areas (Standard Average European, henceforth SAE, and East and South-East Asia), in order to understand to what extent they can be unambiguously assigned to the three classes indicated in (1) and, moreover, to evaluate how clear-cut the boundaries among them are.

2. ATAP, CO, and SUB compounds in SAE languages

As shown in (1), the most recent classification of compounds (Bisetto & Scalise 2005) identifies three classes of compounds, mainly on the basis of the grammatical relation between the constituents.

Nevertheless, such a classification does not capture the fact that the endo- / exocentricity criterion can assume different values with regard to the three classes identified in (1). On the one side, as far as SUB and ATAP compounds are concerned, it is legitimate to discriminate the compounds which have a head from the compounds which do not. Moreover, in endocentric compounds the formal and semantic heads usually coincide. As far as coordinate compounds are concerned, on the other hand, the notions of exocentricity and endocentricity are sometimes inadequate, when we move from the formal to the semantic level. As a premise, it is necessary to emphasize that cross-linguistically many types of coordinate compounds are attested. However, two of them are the most widely diffused and, moreover, their diffusion seems clearly delimited from the areal point of view. In Bisetto & Scalise's (2005) classification these types are labelled as endocentric and exocentric compounds and can be exemplified by It. *studente lavoratore* ('student worker') and Thai *phɔ̂ɔmêɛ* 'parents' (*phɔ̂ɔ* 'father' + *mêɛ* 'mother'), respectively.

If we consider a form like *sword fish*, a typical ATAP compound, the relation of modification is unidirectional: *sword* modifies *fish*, but not *vice versa*. In other words, a *sword fish* is a kind of fish, but not a kind of sword. On the contrary, as to It. *studente lavoratore*, an endocentric coordinative compound (according to the traditional terminology), the relation of modification is bidirectional: a *studente lavoratore* is both a kind of student and a kind of worker. Thus, the so-called coordinative endocentric compounds are better labelled as two-headed compounds or, most properly, as hyponymic compounds: the whole compound is a hyponym of both constituents.

Also the label 'exocentric coordinative compounds' is unsuitable to describe the data, since in this case there is not any modification relation: the two (ore more) members of a compound equally contribute to the whole meaning. The crucial point to emphasize is that the compound is a hyperonym of its constituents. In a compound like Thai *phɔ̂ɔmêɛ* 'parents' both members, *phɔ̂ɔ* 'father' and *mêɛ* 'mother', indicate a kind of parent.

Hyponym coordinate compounds are attested in SAE languages and often express accidental coordination, that is coordination between items which are in a semantically loose relation. Hyperonymic compounds are frequently attested in East and South-East Asian languages and typically express natural coordination, that is coordination between items which have a close semantic relationship.[8] Consequently, in this section the label CO compounds will be used to mean hyponymic coordinate compounds of the *singer actor* type.

If we go back to the criteria identified in the previous paragraph, as far as SAE languages are concerned, some interesting and promising similarities between ATAP and SUB compounds seem to emerge.

At first, as already seen in (8), both ATAP and SUB compounds can violate the well known 'No phrase constraint': the complement / modifier slot can be filled by a linguistic item larger than a word, i.e. by a syntactic construction:

(10) Dutch *lach of ik schiet humor*
'laugh or I shoot humour'
(Bisetto & Scalise 1999:35)
Eng. *green-and-red truck driver*
(Google search May 07)
God is dead theology
(Lieber & Scalise 2006:10)
It. *raccolta rifiuti tossici e ingombranti*
'toxic and bulky waste collection'
ragazza casa e chiesa
'well-educated girl' (lit: 'girl+house and church')

Furthermore, these compounds can contain an inflected constituent in non-head position:

(11) It. *lavapiatti*
'dishwasher'
Icel. *hafnar-garður*
'harbor wall, dock'

In the Italian form, the second member is a plural noun, independently of the number of the whole compound (cf. *la lavapiatti* / *le lavapiatti* 'dishwasher$_{SG\ /\ PL}$'); in the Icelandic form, the first constituent is inflected for the genitive. According to Indriðason (1999), such a form belongs to a productive class of Icelandic compounds labelled as 'genitive compounds', which display formal and semantic relations to noun phrases with genitive complements. So, they are usually described as derived in syntax.[9]

Finally, as shown by the data in (3), these compounds can be recursively expanded.[10]

Possibly the only criterion that does not apply to ATAP compounds of SAE languages is the coreference between a member of a compound and an external anaphoric element.

After a preliminary rough survey the following picture can be sketched:

(12)

	CO	ATAP	SUB
Violation of the 'No phrase constraint'		✓	✓
Inflected member		✓	✓
Recursivity		✓	✓
Violation of anaphoric islandhood			✓

Nevertheless, a deeper investigation shows some crucial differences between ATAP and SUB compounds. Consider the data in (8) and (10), which illustrate the presence of a possible syntactic construction within a compound: while the syntactic nature of the non head constituent of SUB compounds such as *raccolta rifiuti tossici e ingombranti* can hardly be disputed, some doubts emerge if we consider ATAP compounds. In this case, non head constituents larger than a word usually correspond to binomials (It. *casa e chiesa* in *ragazza casa e chiesa*) or to 'frozen phrases' (En. *floor-of-a-birdcage* and *one-hat-per-student* in *floor-of-a-birdcage taste* and *one-hat-per-student stipulation* respectively), that is to linguistic items that were created in syntax, but underwent a process of lexicalization afterwards. So, in a synchronic perspective, these constructions cannot be considered as typically syntactic. This assumption is further supported by the fact that in these compounds the non head constituent can hardly be expanded: a compound as It. *ragazza casa e chiesa* can be expanded only by adding such items as adverbs (*molto*) or adjectival modifiers (*tutta*):

(13) [[ragazza]$_{N}$ [casa e chiesa]$_{NP?}$]
[[ragazza]$_{N}$ tutta / molto [casa e chiesa]$_{NP?}$]

However, the modifiers are external to the second constituent, as demonstrated by the fact that *tutta* agrees with *ragazza*, not with *casa* and / or *chiesa*. In fact, if the head noun is masculine, the same modifier needs to agree with it for gender: *ragazzo tutto casa e chiesa*. This fact shows that it is not the second constituent which is expanded, unlike the case of a SUB compound as in (14):

(14) [[*porta*]$_{V}$ [*anelli*]$_{NP}$] > [[*porta*]$_{V}$ [*anelli, orecchini o piccoli monili*]$_{NP}$]
'ring box' (lit: carry+ring$_{PL}$) 'box for rings, earrings or small jewels'
(Ricca 2005:479)

This demonstrates that the relation between the two constituents of an ATAP compound is looser than that between the two members

of a SUB compound: in many respects, *ragazza casa e chiesa* behaves more like a NP containing an adjectival modifier than as a compound. As a consequence, the tick in the ATAP cell in the first line of Table in (12) should probably be followed by a question mark.

Therefore, according to the data discussed so far and summarized in the Table in (12), whereas CO compounds can be classified as pure morphological objects, built up from lexemes (abstract forms) and constituting a lexeme (though often a non-prototypical one), SUB compounds can be considered as the outcome of word formation processes that systematically involve both syntax and morphology. In other words, they seem to be composed by a morphological constituent (usually a stem; it is always the head in endocentric compounds) and a syntactic construction (usually a NP). The consequence of this claim is to admit that syntax can 'feed' morphology not exceptionally or occasionally (i.e. via lexicalization and 'fossilization' of widely used phrases), as it has often been argued, but systematically, as some recent theoretical models of word formation already acknowledge (cf. Lieber & Scalise 2006 for an overview). In this respect, in a formal perspective ATAP compounds seem more similar to SUB compounds than to CO compounds. However, in spite of the clear picture that emerges from the criteria listed in the first paragraph and summarized in table (12), quite paradoxically the highest number of ambiguous cases with respect to the tripartite classification in (1) concerns the distinction between ATAP and CO compounds. In other words, there are many more forms that can be interpreted both as ATAP and CO compounds than forms which can be interpreted as ATAP and SUB compounds. As observed above in connection with data in (3), the distinction between ATAP and CO compounds is usually based on a metaphoric interpretation of the non-head constituent, which means that the distinction is often a matter of pragmatics. This is undoubtedly a rather weak criterion, all the more so in SAE languages, where, as stated above, CO compounds always belong to the hyponymic type (i.e. En. *secretary treasurer*) and should be placed in an intermediate position between the most typical CO compounds, those expressing natural coordination, and 'hierarchical' compounds.[11] So, it can be assumed that, for European languages, the difference between hyponymic CO compounds and ATAP compounds is the most difficult to capture. In this case, it is necessary to find formal correlates to the semantic – and pragmatic – variables that influence their interpretations.

In order to represent the meaning of compounds and their members, we will adopt Lieber's (2004) framework and terminology. In the theoretical picture drawn by Lieber, the semantics of a lexeme is subdivided in two components, the skeleton and the body. The skeleton is hierarchically organized and includes all information relevant to syntax. The body contains encyclopaedic information connected to a lexeme: it cannot be deconstructed and formally represented by features. While the skeleton is relatively fixed and hardly modifiable across time, the variation of the information stored in the body is higher and highly conditioned by use.

The difference between ATAP compounds and hyponymic CO compounds is mainly a matter of body, so we will neglect the skeleton. In a general perspective, if there is a perfect match between the bodies of the two members of a compound (that is, if the bodies are identical with respect to the quantity and the nature of the information they express) except for one feature, then the compound is coordinate in most cases:

(15)	*studente*	*lavoratore*	>	*studente lavoratore*
	'student'	'worker'		
	body			
	<natural>	<natural>		<natural>
	<human>	<human>		<human>
	<male>	<male>		<male>
	<he studies>	<he works>		<he studies and works>

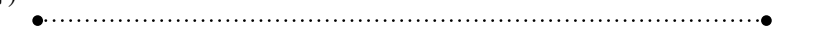
On the contrary, if the match between the bodies is limited to just one feature, irrespective of the quantity and the nature of the information, then the compound tends to belong to the ATAP class:

(16)	*snail*	*mail*	>	*snail mail*
	body			
	<gastropod>	<institution>		<institution>
	<secretes slime>	<means of communication>		<means of communication>
	<very slow>	<takes time>		<very slow>

(cf. Bisetto & Scalise 2005)

We can formulate the hypothesis that ATAP and hyponymic CO compounds are placed on a *continuum*, with the examples in (15) and (16) at the extreme poles:

(17)



CO — ATAP
studente lavoratore — *snail mail*

There are, however, many complex forms in between for which an unambiguous judgment can hardly be formulated, as in the cases of *ghost writer* or *woman doctor* previously discussed. In these situations, it is even more pressing to resort to formal criteria in order to resolve ambiguity. For this purpose, the most forceful criterion seems to be the reversibility of the constituents' order, suggested by Renner (forthcoming) to identify CO compounds: "the compound X.Y could be named Y.X". This test allows us to unambiguously classify a compound as *ghost writer* as ATAP: if we invert the order of constituents, we obtain a form with a different referential meaning (a *ghost writer* is an 'invisible' writer; a *writer ghost* is a ghost that uses to write). On the contrary, if the position of *studente* and *lavoratore* are inverted, we get a compound with the same referential meaning (but probably with different pragmatic nuances). We can conclude, then, that a compound can be classified as coordinate if its members' bodies match except for one feature and if the order of its members can be inverted.[12]

Therefore, we can modify the table in (12) as follows:

(18)

	CO	**ATAP**	**SUB**
Violation of the 'No phrase constraint'		✓?	✓
Inflected member		✓	✓
Recursivity		✓	✓
Violation of anaphoric islandhood			✓
Constituents' reversibility	✓		

3. Features of compounding in isolating languages: Chinese

The hypotheses we put forward above about the class features of SUB, ATAP and CO compounds have been developed through the observation of the behaviour of such complex words in the rather 'familiar' Indo-European languages of Europe, i.e. typical fusional languages (with the possible exception of English). Let us now turn to a language which is different from the point of view of morphological typology such as Chinese. Despite its fundamentally isolating character, Chinese in

fact possesses a number of productive word-formation processes which construct complex words, similar in principle to what we call compounds in Indo-European languages, i.e. words which can intuitively be defined as being made up of two or more words and which are endowed with "lexical autonomy" according to speakers' judgments (Grandi 2006). First of all, we shall make a rather obvious observation. Chinese is an isolating language with no inflectional markers for e.g. gender and case on nouns and adjectives, and only possesses a small set of aspectual markers for verbs (which apparently are almost never obligatory; cf. Li & Thompson 1981:193, 206). Thus, the criterion concerning the inclusion of an inflected constituent plays no role in this language. On the other hand, one can find interesting data on SUB and ATAP compounds which seemingly contain a syntactic constituent:[13]

(19) 追索侵占物诉讼
zhūisuǒ *qǐnzhànwù* *sùsòng*
pursue seize-thing lawsuit
'trover'

The syntactic structure underlying this compound may be represented as in (20):

(20) [[[*zhūisuǒ*]V [*qǐnzhànwù*]N]VP [*sùsòng*]N]N
pursue seize-thing lawsuit

The semantic relationship between the head *sùsòng* and the non-head may be defined as 'aimed at' (a trover being a legal action aimed at recovering goods wrongfully taken) and thus we may consider (19) as a SUB compound. Such cases are far from being exceptions in the language: we can observe productive word formation patterns which involve a phrasal non-head and a lexical head (often, a bound lexical morpheme). See, for instance, some words which have 机 *jī* 'machine' as the head:

(21) a. 自动取款机
zìdòng *qǔkuǎnjī*
self-move withdraw-sum-machine
'ATM'

b. 吹风机
chuīfēngjī
blow-wind-machine
'hair dryer'

c. 打洞机
dǎdòngjī
punch-hole-machine
'perforator'

It may be argued that the non-head constituents of the words in (21a-b) have a word status, since these elements are listed in dictionaries as separable "verb-object compounds" (see Li & Thompson 1981:73-81) and have an ambiguous state between 'true compounds' and phrases. They can be separated by certain morphological elements (or even undergo topicalization of the object constituent), but share some properties of compounds, such as non-transparency of meaning (though these characteristics are different for each verb-object compound). The verb-object construction in (21c), *dǎdòng* 'make a hole', however, is clearly phrasal,[14] and the structure of the whole compound may be represented as:

(22) [[[*dǎ*]V [*dòng*]N]VP [*jī*]N]N
punch hole machine

According to He (2004), only constructions such as those in (19), i.e. with a disyllabic verb and a disyllabic object as the non-head constituent may be regarded as truly syntactic. He's arguments for this analysis are the impossibility of taking the human plural / collective marker *-men*, which should be possible with all human nouns, and the possibility of having an adjunct modifying either the verb or the object. In (23a-b) we give He's examples (cf. He 2004:3; glosses adapted).

(23) a. 盗竊國寶犯
dàoqiè guóbǎo fàn
steal country-treasure criminal
'thief of state treasures'

b. *盗竊國寶犯們
**dàoqiè guóbǎo fànmen*
steal country-treasure criminal:PL
'thieves of state treasures'

According to He (2004), this happens because these are not "canonical lexical structures", but are rather built on syntactic principles.

Despite the apparently hybrid nature of these compounds, which involve a lexical and a syntactic constituent just as the SUB compounds in the languages of Europe we dealt with above (see section 2.), we

could not find any clear-cut instances of violation of anaphoric island-hood for these compounds in Chinese: this could be either a matter of chance (even the Italian examples in (9) are uncommon and difficult to record), or it might be explained by other features of the language (as e.g. the way anaphora works in Chinese; see Li & Thompson 1981:657).

Just as in English (compare the often-quoted example *bathroom towel rack designer training course notes*; Selkirk 1982:15), in Chinese ATAP compounds may be recursively expanded, as in examples (24a-c):

(24) a. 信息服务
xìnxī fúwù
information service
'information service'

b. 管理信息服务
guǎnlǐ xìnxī fúwù
manage information service
'management information service'

c. 公共管理信息服务
gōnggòng guǎnlǐ xìnxī fúwù
public manage information service
'Common Management Information Service (CMIS)'

This possibility is not attested with the same productivity in all languages. In (3) we gave an example of a recursive compound in Italian. However, as Scalise (1994:141) points out, many English recursive compounds correspond to 'true' syntactic structures in Italian, containing determiners, prepositions, etc. This is probably due to the morphological type to which a language belongs: recursive compounds are apparently more frequent in more isolating languages.

Hyperonymic compounds, which are very common in the Chinese lexicon (see e.g. Wälchli 2005), normally constitute fixed, lexicalized structures. However, we have a semantic class of coordinating compounds, which Wälchli calls "non-pairing additive co-compounds", which "express collection complexes which are exclusively listed by the parts" (2005:139), such as 刀叉 *dāochā* knife-fork 'knife and fork', which may be expanded by the addition of other coordinands:

(25) a. 港澳
Gǎng-Ào
Hong Kong-Macao
'Hong Kong and Macao'

b. 港澳台
Gǎng-Ào-Tái
Hong Kong-Macao-Taiwan
'Hong Kong, Macao and Taiwan'

One should not however exaggerate the importance of this phenomenon, which seems to be restricted both by phonological-prosodic constraints (it is not common for Chinese to have words containing more than three syllables) and by pragmatic constraints (the coordinands must evoke some conceptual unit, in Wälchli's words, and not be a mere juxtaposition of meaningful units). Generally speaking, the order of constituents in a Chinese CO compound cannot be inverted. This may be motivated by pragmatic and cultural factors (e.g. in a patriarchal society, a compound such as 'mother-father' for 'parents' cannot be reversed) or simply because of lexicalization. Feng (1998:223) gives several examples of bimorphemic coordinate constructions attested with both possible orders of constituents (AB and BA) from Ancient and Middle Chinese; later, when these constructions lexicalize and become "proper" CO compounds, their order becomes fixed.

In summary, the three classes of compounds (SUB, ATAP, CO) in Chinese seem to have a behaviour similar to that of compounds in European languages: SUB compounds in Chinese may contain a phrasal non-head constituent (and, possibly, allow word-internal anaphora, but this depends on what instances one regards as compounds); ATAP compounds may be recursively expanded and seem to forbid word internal anaphorical reference; CO compounds may undergo expansion, but seemingly not in a productive and regular way. Let us now turn to the examination of an agglutinating language, Japanese.

4. Features of compounding in agglutinating languages: Japanese

Compounding in Japanese is pervasive "in both lexical and syntactic domains, as opposed to the rather modest affixes and inflections" (Kageyama 2009:512). The distinction between compounds and phrases in Japanese, according to Kageyama's (2009:512) analysis, is in principle consistent with general morphological theory: "compounds are diagnosed by the absence of case markers, inflections and other functional categories from the non-head position". In Japanese, there are three (non-homogeneous!) classes of compounds which contain a syntactic element (Kageyama 2009:518); these are (examples and glosses adapted from the source; characters added):

(26) "Phrasal compounds"
きれいな町造り
kirei na machi-zukuri → [[*kirei na machi*]NP *zukuri*]N
clean:INFL town-making
'construction of a clean town'

(27) "Possessive compounds"
日の出
hi no de
sun:GEN rise
'sunrise'

(28) "Word plus-level compounds"
貿易会社社長
bōeki-gaisha shachō
trading-company president
'president of a trading company'
(with a short pause between *bōeki-gaisha* and *shachō*)

The non-head constituent in phrasal compounds may for instance contain an adjectival modifier (note the presence of internal inflectional markers), such as in (26), or a coordinated NP; possessive compounds such as (27) are lexicalized NPs with an internal overt genitive marker *no*. As Kageyama (2009:519) puts it, "[d]espite their syntactic flavor, those two types of phrase-like compounds exhibit all the traits of lexical words in terms of compound accent, limited productivity, and lexical conditioning."

"Word plus" (henceforth, W+) compounds may neither contain a phrasal constituent nor have overt internal markers of functional / grammatical categories, but nevertheless they are pronounced as two prosodically independent words and "are immune from lexical conditioning" (Kageyama 2009:519). What is more interesting from our perspective is their visibility for word-internal anaphora (examples adapted from Kageyama 2009:520):

(29) a. 大統領は明日友好条約に調印する予定だ。
daitōryō wa asu yūkō-jōyaku ni chōin suru yotei da
president TOP tomorrow amity-treaty DAT sign do schedule COP

b. 同条約最終案によると…
dō-jōyaku saishūan ni yoru to
same-treaty final version according to
The President is going to sign the amity treaty. According to the final version of that treaty...'

We will not discuss the issue of W+ here any further due to lack of space: for the theoretical correlates of such a notion, see Kageyama (2001 and 2009).

In short, the classes – or rather subclasses – of SUB and ATAP compounds in Japanese seem to allow the presence of a syntactic non-head constituent, displaying internal grammatical markers (cf. (8)-(9)). At least for the W+ level, SUB compounds also allow word-internal anaphora. Another difference between SUB and ATAP compounds in Japanese is that apparently only the latter is 'immune' from word-internal grammatical markers.

It is worth noticing that, in Japanese, for CO compounds at the W+ level, the coordinands may be separately visible for anaphora (adapted from Kageyama 2009:515):

(30) 夫婦は互いを励ました
fūfu wa tagai o hagemashita
husband-wife TOP each other OBJ cheer up:PAST
'the husband and wife cheered each other up'

A sentence like (30) seems also possible in Chinese, a language for which a W+ analysis has never been proposed, to our knowledge. Moreover, in Chinese these constructions are normal phonological words. This aspect of the behaviour of CO compounds is possibly the only relevant difference in the overall configuration of the compounding domain between fusional European languages on one side and Chinese and Japanese on the other.

5. Concluding remarks

The examination of data from languages of Europe (SAE) and from the East and South-East Asian area, belonging to the three major morphological types, namely fusional, isolating and agglutinating, allows us to formulate some tentative generalizations about the behaviour of the three classes of compounds identified by Bisetto & Scalise (2005). Our findings may be summarized in the table in (31), which completes the picture we gave in (18):

(31)

	CO		ATAP	SUB
Violation of the 'No phrase constraint'	HYPONYMIC	HYPERONYMIC	✓?	✓
Inflected member			✓	✓
Recursivity			✓	✓
Violation of anaphoric islandhood		✓		✓
Constituents' reversibility	✓			

As far as the distinction between ATAP and SUB compounds is concerned, 'Asian' data seem to support our conclusions about SAE languages, namely that SUB and ATAP compounds, generally speaking, behave similarly in many respects, but the latter tend to be 'tighter' in term of anaphoric islandhood. Also, whereas we have plenty of examples of SUB compounds containing a non-head syntactic constituent, it is less clear whether elements formed productively in syntax may be part of an ATAP compound (see exx 8-10).

However, we have found some significant differences in the behaviour of CO compounds in SAE and in East Asian languages. This, however, has nothing to do with idiolinguistic or areal features; differences are motivated by the kind of compounds that we analyzed for the two 'groups': hyponymic compounds for SAE languages and hyperonymic compounds for Chinese and Japanese. What the two subclasses of CO compounds have in common is that they show more integrity, from the morphological point of view: we have found no violations of the 'No phrase constraint' in CO compounds. Recursivity seems to be rare in the domain of CO compounds, possibly for pragmatic reasons (it would be odd to have long sequences of coordinated elements).

The violation of anaphoric islandhood is rather common for hyperonymic CO compounds in Chinese and Japanese, while apparently impossible in the 'European' hyponymic ones: since such compounds have a single referent, it would be logically unacceptable to make reference only to one of its constituent parts. Reversibility of constituents, as said before, may be constrained by pragmatic and cultural factors in East Asian languages; what plays the most important role, however, seems to be lexicalization, which prevents reversal of constituents.

To conclude, our data showed that the behaviour of ATAP compounds is quite similar cross-linguistically, as well as that of SUB compounds; as far as CO compounds are concerned, we see 'mirror' features for the two subclasses (hyponymic and hyperonymic), having in common the apparent impossibility of having a syntactic element as a constituent. It would be interesting to test the consistency of such findings on a broader, typologically balanced sample.

Addresses of the Authors:

Giorgio Francesco Arcodia, Facoltà di Scienze della Formazione, Università degli Studi di Milano-Bicocca, ed. U6, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy <giorgio.arcodia@unimib.it>
Nicola Grandi, Dipartimento di Studi Linguistici e Orientali, Università degli Studi di Bologna, via Zamboni 33, 40126 Bologna, Italy <nicola.grandi@unibo.it>
Fabio Montermini, UMR 5263 CLLE – ERSS, Maison de la Recherche, Université de Toulouse le Mirail, 5, allées Antonio Machado, 31058 Toulouse Cedex 9, France <fabio.montermini@univ-tlse2.fr>

Notes

* Although this work is the outcome of joint research, F. Montermini is responsible for section 1, N. Grandi for section 2, and G. F. Arcodia for sections 3, 4 and 5. We are grateful to Jesse Tseng and to an anonymous reviewer for their remarks on a previous version of this text.

The abbreviations used in the text are:

COP	copula
DAT	dative
GEN	genitive
INFL	inflectional marker
OBJ	object marker
PAST	past tense
PL	plural
TOP	topic marker

1 Cf. also Scalise *et al*. (2005).

2 Henceforth, we will refer to the three classes as SUB (subordinate), ATAP (attributive-appositive) and CO (coordinate).

3 Note that the compounds in (2) all indicate a hyponym of the two elements: a *fighter bomber* is, at the same time, a particular sort of fighter and a particular sort of bomber (plane). In other languages, however, a CO compound may indicate a hyperonym of the two elements, as we will see in the next section.

4 Unlike Bisetto (2004), we do not regard Italian compounds such as *centro assistenza pneumatici* as belonging to a particular register, outside of regular, productive Italian morphology (see also Baroni *et al*. 2007, who state that these

constructions "belong to a particular syntactic register and are not part of morphology").

[5] In some particular cases, compounds may have inflectional markers only on one of the members, independently of its role in the compound. This is true in particular for lexicalized compounds (cf. It.: *pescecane* / *pescecani* 'shark', lit. 'fish+dog').

[6] Gaeta & Ricca (2009) define subordinate compounds displaying a syntactic construction in non head position as morphological units which are not lexical units.

[7] Cf. Postal (1969) for the notion; Dressler (1987), Ward *et al*. (1991) and Montermini (2006), among others, for a discussion.

[8] For a framework for coordination in a cross-linguistic perspective, cf. Haspelmath (2004); for the difference between natural and accidental coordination, cf. among others Wälchli (2005).

[9] Indriðason (1999) states that a compound as *vél-ar-hljóð* 'machine sound' is formed from the phrase *hljóð vél-ar* 'sound of a machine' with an inversion of the constituents, and a subsequent merger of them.

[10] On recursivity in compounding, cf. Haider (2001) who draws an interesting correlation between this property and the final position of the head.

[11] In his monograph on co-compounds, Wälchli (2005) does not include compounds as It. *studente lavoratore* ('student worker') in the number of coordinate compounds, considering them closer to ATAP compounds.

[12] Inversion of constituents' order within a coordinate compound has often pragmatic reasons, as has been well argued by Malzahn (2000) on Sanskrit compounds.

[13] The existence of word formation elements such as 者 *zhě* 'human suffix' which act as heads for complex constructions having as a non-head a (supposed) syntactic constituent, such as 不符合条件者 *bùfúhétiáojiànzhě* 'not qualified' (lit. 'non-conforming-to-conditions-zhe') has led Dong (2004:85 ff.) to propose the category of "semi-free morphemes" for Chinese, which may act as word affixes or as clitics; here we shall not deal with this issue, since it does not concern the examples which will be analysed here.

[14] Here we adopt Packard's (2000:115-125) position: a verb-object construction which cannot have another object (valency being saturated by the object inside the construction) and has not undergone idiomaticization of meaning is (or acts as) a phrase. Lexicalized verb-object constructions (i.e. compounds) should either be able to be followed by an object (like 投资 throw-money 'to invest') or have an idiomatized, non-compositional meaning (such as 担心 *dānxīn* carry-heart 'to worry').

Bibliographical References

Baroni Marco, Emiliano Guevara & Roberto Zamparelli 2007. Italian deverbal compounds: Words, phrases or either?. Communication presented at IGG 33, Bologna, March 3, 2007.

Benveniste Emile 1974 (1967). Fondements syntaxiques de la composition nominale. In Id. *Problèmes de linguistique générale*. 2. Paris: Gallimard. 145-162.

Bisetto Antonietta 2004. Composizione con elementi italiani. In Grossmann Maria & Franz Rainer (eds.). *La formazione delle parole in italiano*. Tübingen: Niemeyer. 33-51.

Bisetto Antonietta & Sergio Scalise 1999. Compounding: morphology and/or syntax? In Mereu Lunella (ed.). *The Boundaries of Morphology and Syntax*. Benjamins: Amsterdam. 31-48.

Bisetto Antonietta & Sergio Scalise 2005. The classification of compounds. *Lingue e linguaggio*. IV.2. 319-332.

Bloomfield Leonard 1933. *Language*. New York: Holt.

Dong Xiufang 2004. 汉语的词库与词法 (Chinese lexicon and morphology). Beijing: Beijing Daxue Chubanshe.

Dressler Wolfgang U. 1987. Morphological islands: Constraint or preference? In Steele Ross & Terry Threadgold (eds.). *Language Topics. Essays in Honour of Michael Halliday*. Amsterdam / Philadelphia: Benjamins. 71-79.

Feng Shengli 1998. Prosodic structure and compound words in Classical Chinese. In Packard Jerome (ed.). *New Approaches to Chinese Word Formation*. Berlin / New York: Mouton de Gruyter. 197-260.

Gaeta Livio & Davide Ricca 2009. *Composita solvantur*: Compounds as lexical units or morphological objects? This volume.

Grandi Nicola 2006. Considerazioni sulla definizione e la classificazione dei composti. *Annali dell'Università di Ferrara* 1. 31-52. http://eprints.unife.it/annali/lettere/2006vol1/grandi.pdf

Haider Hubert (2001). *Riesengratulationkompositum - *Kompositumgratulationriesen* or: why are there no complex head-initial compounds?*. In Schaner-Wolles Chris, John R. Rennison & Friedrich Neubarth (eds.). *Naturally! Linguistic studies in honour of Wolfgang Ulrich Dressler presented on the occasion of his 60th birthday*. Torino: Rosenberg & Sellier. 165-174.

Haspelmath Martin 2004, Coordinating Constructions: An overview. In Id. (ed.), *Coordinating Constructions*. Amsterdam: Benjamins. 3-39.

He Yuanjian 2004. The words-and-rules theory: evidence from Chinese morphology. *Taiwan Journal of Linguistics* II.2. 1-26.

Indriðason Þorsteinn G. 1999. Um eignarfallssamsetningar og aðrar samsetningar í íslensku [On genitive compounds and other compounds in Icelandic]. *Íslenskt mál* 21. 107-150.

Kageyama Tarō 2001. Word Plus. In van der Weijer Jeroen Maarten & Tetsuo Nishihara (eds.). *Issues in Japanese Phonology and Morphology*. Berlin: Mouton de Gruyter. 245-276.

Kageyama Tarō 2009. Isolate: Japanese. In Lieber Rochelle & Pavol Štekauer (eds.). *The Oxford Handbook of Compounding*. Oxford: Oxford University Press. 512-526.

Li Charles N. & Sandra N. Thompson 1981. *Mandarin Chinese: a Functional Reference Grammar*. Berkeley / Los Angeles: University of California Press.

Lieber Rochelle 2004. *Morphology and Lexical Semantics*. Cambridge: Cambridge University Press.

Lieber Rochelle & Sergio Scalise 2006. The Lexical Integrity Hypothesis in a new theoretical universe. *Lingue e Linguaggio* V.1. 7-32.

Lombardi Vallauri Edoardo 2005. When are phrases "compounds"? The case of Japanese. In Grossmann Maria & Anna M. Thornton (eds.). *La for-*

mazione delle parole. Atti del XXXVII Congresso Internazionale di Studi della Società di Linguistica Italiana. Roma: Bulzoni. 309-334.

Malzahn Melanie 2000. Die Genese des indogermanischen Duals. In Ofitsch Michaela & Christian Zinko (eds.). *125 Jahre Indogermanistik in Graz*. Graz: Leykam. 291-315.

Montermini Fabio 2006. A new look on word-internal anaphora on the basis of Italian data. *Lingue e linguaggio* V.1. 127-148.

Packard Jerome 2000. *The Morphology of Chinese. A Linguistic and Cognitive Approach*. Cambridge: Cambridge University Press.

Postal Paul M. 1969. Anaphoric islands. In Binnik Robert I., A. Davison, G.M. Green & J.L. Morgan (eds.). *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*. Chicago: University of Chicago. 209-239.

Renner Vincent forthcoming. On the semantics of english coordinate compounds. *English Studies. A Journal of English Language and Literature*.

Ricca Davide 2005. Al limite tra sintassi e morfologia: i composti aggettivali V-N nell'italiano contemporaneo. In Grossmann Maria & Anna M. Thornton 2005 (eds.). *La formazione delle parole. Atti del XXXVII Congresso Internazionale di Studi della Società di Linguistica Italiana*. Roma: Bulzoni. 465-486.

Scalise Sergio 1994. *Morfologia*. Bologna: Il Mulino.

Scalise Sergio, Antonietta Bisetto & Emiliano Guevara 2005. Selection in compounding and derivation. In Dressler Wolfgang U., Dieter Kastovsky, Oskar E. Pfeiffer & Franz Rainer (eds.). *Morphology and Its Demarcations*. Amsterdam / Philadelphia: Benjamins. 133-150.

Selkirk Elisabeth O. 1982. *The Syntax of Words*. Cambridge MA: MIT Press.

Wälchli Bernard 2005. *Co-Compounds and Natural Coordination*. Oxford: Oxford University Press.

Ward Gregory, Richard Sproat & Gail McKoon 1991. A pragmatic analysis of so-called anaphoric islands. *Language* 67.3. 439-474.