Morphosyntax and Semantics in the NooJ Italian Dictionary of Simple Words

Mario Monteleone
Dipartimento di Scienze Politiche e della Comunicazione
Università degli Studi di Salerno – Italy
mmonteleone@unisa.it

Abstract

The main topic of this paper is to describe how to transform effectively the "lexical matter" of a language (not only Italian) into a formal and taxonomic morphosyntactic classification of the simple words stored and tagged inside NooJ electronic dictionaries. The ultimate goal is to build electronic dictionaries for NooJ that can be used effectively in Natural Language Processing (NLP) and Automatic Textual Analysis (ATA). To achieve this task, we will start from the following assumption: nouns inflection codes may prove useful not only to describe morphological behaviours and features, but also to "predict" syntactic combinations inside sentences. This means that inside NooJ finite-state automata (FSA) and transducers (FST), nouns inflection codes can account for Lexicon-Grammar (LG) co-occurrence and selection restriction rules, thus also providing for several aspects of formal semantics (FS).
The method we intend to outline here can therefore become a descriptive and applicative standard, reusable for all those languages in which word inflection has a value not only morphological (as for gender, number and possibly case) but also syntactic. As is known, such a formal descriptive approach is absent in paper dictionaries, in which there is a tendency to "flatten" the morphosyntactic description of words in favour of a list of words which uses rely heavily on the linguistic competence of readers and speakers.

Keywords: NooJ, NooJ Grammars, Italian morphosyntax, Formal Semantics, Lexicon-Grammar.

Morfosintaxis y Semántica en el Diccionario NooJ Italiano de Palabras Simples

Resumen

Este artículo tiene como tema central describir cómo transformar efectivamente la "materia léxica" de una lengua (no solo el italiano) en una clasificación morfosintáctica formal y taxonómica de las palabras simples almacenadas y etiquetadas en los diccionarios electrónicos de NooJ. El objetivo final es construir diccionarios electrónicos en NooJ que se puedan usar de manera efectiva en el procesamiento de lenguaje natural (PLN) y el análisis textual automático (ATA). Para llevar a cabo esta tarea, partiremos del supuesto de que los códigos de flexión nominal pueden resultar útiles no solo para describir comportamientos y características morfológicas, sino también para "predecir" combinaciones sintácticas dentro de las oraciones. Esto significa que, dentro de los autómatas de estado finito (FSA) y los transductores de estado finito (FST) de NooJ, los códigos de flexión nominal pueden dar cuenta de las reglas de restricción, selección y coocurrencia de Léxico-Gramática (LG), proporcionando así varios aspectos de la semántica formal (FS).
Creemos que el método que pretendemos esbozar aquí puede convertirse en un estándar descriptivo y aplicativo, reutilizable para todas aquellas lenguas en las que la flexión de las palabras tiene un valor no sólo morfológico (para género, número y posiblemente caso) sino también sintáctico. Este enfoque descriptivo formal generalmente está ausente en los diccionarios en papel, en los que existe una tendencia a "aplanar" la descripción morfosintáctica de las palabras y presentar una lista de palabras basada, en gran medida, en la competencia lingüística de lectores y hablantes.

Palabras clave: NooJ, gramática de NooJ, morfosintaxis del italiano, semántica formal, Léxico-Gramática

1. Introduction

The main topic of this paper is to describe how to transform effectively the "lexical matter" of a language (not only Italian) into a formal and taxonomic morphosyntactic classification of the simple words stored and tagged inside NooJ electronic dictionaries. The ultimate goal is to build electronic dictionaries for NooJ

that can be used effectively in Natural Language Processing (NLP) and Automatic Textual Analysis (ATA). To achieve this task, we will start from the following assumption: nouns inflection codes may prove useful not only to describe morphological behaviours and features, but also to "predict" syntactic combinations inside sentences. This means that as inside NooJ finite-state automata (FSA) and transducers (FST), nouns inflection codes can account for Lexicon-Grammar (LG) co-occurrence and selection restriction rules[1], thus also providing for several aspects of formal semantics (FS).[2]
The method we intend to outline here can therefore become a descriptive and applicative standard, reusable for all those languages in which word inflection has a value not only a morphological (as for gender, number and possibly cases) but also syntactic. As is known, such a formal descriptive approach is absent in paper dictionaries, in which there is a tendency to "flatten" the morphosyntactic description of words in favour of a list of words which uses rely heavily on the linguistic competence of readers.

2. Main Features of NooJ Electronic Dictionaries

The specificities of a NooJ electronic dictionary are of two types: formal and content-related.

From a formal point of view, the structure of an electronic dictionary must be that of a lexical database,[3] i.e. unambiguously readable and applicable by a specific software.

As for its content, an electronic dictionary must:

- Reach a comprehensive and taxonomic coverage of the lexical matter of a language, not taking into account, for example, the fact that one or more words have a low level of use, therefore not relying on statistics to identify the words to be listed;
- Be rigorously separating synchrony from diachrony. This means, for example, that the electronic dictionary of the language of Molière must not lemmatize words only used in 21st century French, and vice versa.

However, as in the case of Occam's Razor, *entia non sunt multiplicanda praeter necessitatem*. At the same time, recalling Parmenides, two are the only roads of enquiry there are to think of: one, that it is and that it is not possible for it not to be; the other, that it is not and that it must not be. Therefore, the whole description of the lexical material of a language, including all formal and content-related aspects, must above all be pragmatic, non-redundant, and based on concrete and unambiguous data, both morphosyntactic and (sentence-related) semantic. To approach these topics correctly, in the following article we will deal with the interactions between Morphology and Syntax in the creation of semantically acceptable and grammatical sentences.

2.1. Morphology

In linguistics, the term *Morphology* (cf. the Greek words morphé "form" + lógos "study") traditionally denotes the branch of grammar that studies the form of words, as opposed to Syntax, which deals with the function and combination of words in sentences and discourses.[4] In other words, morphology studies the paradigms of words and the organization of grammatical categories, while syntax deals with word sequences, and the syntagmatic relations required by acceptable and grammatical contexts. From the point of view of the peculiarities of a language, a distinction is made between descriptive (synchronic) morphology, which deals with the morphological structure of a language, at a given moment, and historical (diachronic) morphology, which studies the evolution of the morphological structure of the language and its development prospects.

---

[1] See 2.4.
[2] See https://en.wikipedia.org/wiki/Formal_semantics_(natural_language).
[3] See Monteleone, M., *Lexicographie et dictionnaires électroniques. Des usages linguistiques aux bases de données lexicales*, Université de Marne-la-Vallée, Thèse pour obtenir le grade de Docteur de l'Université de Marne-la-Vallée, présentée et soutenue publiquement par M. Mario Monteleone, 8 décembre 2003. Available for download at https://tel.archives-ouvertes.fr/tel-00627599/document.
[4] The difference between sentence and discourse is here to be understood as established by Zellig S. Harris: a sentence is also called nuclear sentence, since it contains one and only one predicate; a discourse is instead defined as a concatenation of nuclear phrases produced through parataxis and hypotaxis.

2.2. Syntax

As already stated, *Syntax* is the branch of linguistics which studies the way in which words combine to form sentences or discourses in a given language. With regard to these specific themes, our approach is openly structuralist, and above all in the theoretical-practical perspective provided by Maurice Gross's Lexicon-Grammar.[5] In turn, Lexicon-Grammar itself has its roots in the structural and transformational linguistics of Zellig S. Harris[6] and in the structuralist linguistics of Lucien Tesnière[7].
As for Lexicon-Grammar specific approach to syntax, in 2.4 we give the definition of co-occurrence and selection-restriction rules, which in fact allow to connect the distributional and transformational aspects of syntax to the meanings of sentences and speeches, thus laying the foundations for the birth and development of FS. From a purely linguistic point of view, syntax studies:

- Word order, in terms of distribution and allowed/possible transformations;
- Grammatical categories or parts of speech, mainly in terms of word agreements;
- Rection phenomena, i.e. the syntactic/semantic dependencies among words;
- Grammatical functions, that is mainly how morpholexical elements are used inside sentences and discourses, according to syntax rules, as for instance grammatical agreement, or those previewed by the syntactic and semantic profiles of predicates.

2.3. Semantics

Overall, we will here consider *Semantics* only as that branch of linguistics that studies the meaning of natural language communication, for instance, what we want to convey by statements. In this sense, a more detailed consideration becomes here necessary. The semantics we intend to deal with does not concern making evaluations on the truth / falsity of sentences, nor does it depend on such evaluations. The semantics treated here is therefore not as it is understood in Semiotics that is, a tool that allows interpreting communication signs in relation to the surrounding world. On the contrary, we will deal here only with Semantics of Sentences (SoSs), in the strictest sense established by Ferdinand De Saussure, for whom any utterance in a given language must be not only grammatically well built, but also and above all understandable and (re)usable by the community of speakers of that language. Therefore, of the sentence:

1) *Today it snows*

Pronounced in the middle of summer on a sunny beach of the Mediterranean Sea, we will not evaluate whether it describes an event really in progress, but we will evaluate the correct morphosyntactic, distributional and semantic relationship between the elements of the sentence itself. Hence, we will not be dealing with the truth conditions of a statement, critical discourse analysis, and pragmatics (as a branch of Semiotics).
There is a close relationship between SoSs and Syntax, similar to the one that exists between the shape of a container and its content. Essentially, we can say that Syntax is a support for SoSs, the former coping with statements forms, the latter with their content. Besides, as we shall see, each change in the shape of the container can only correspond to an identical change in the shape of its content – or to rephrase, any specific content requires its appropriate containers.
The main objects of study we therefore attribute to SoSs are:

- The meaning of words, simple or compound, with reference to specific sentence contexts;
- The relations of meaning between words, that is, at the lexicon level, homonymy, synonymy, antonymy, polysemy, hyperonymy, hyponymy, and so on;
- The distribution of semantic roles (agents) within sentences ad discourses, on the basis of the syntactic behaviour of predicates.

In the following paragraph, we will see how to use these three last specificities of SoSs inside NooJ electronic dictionaries, in order to address efficiently ATA.

---

[5] See https://en.wikipedia.org/wiki/Lexicon-grammar.
[6] See www.zelligharris.org.
[7] See https://en.wikipedia.org/wiki/Lucien_Tesnière.

2.3.1. Knowledge Domains, Semantic Fields and Terminology

As stated before, one of SoSs topics of study is the meaning of words, simple or compound, with reference to specific sentence contexts. In this sense, the division between simple and compound words becomes of crucial importance: the former, as is well known, are very often polysemic (i.e. present lexical ambiguity). This polysemy is reduced when simple words are contextualized within sentences and discourses. The latter, on the other hand, consisting of more than one simple word, are rarely ambiguous. In fact, the semantic expansion[8] to which they are subjected gives them a unique and specific overall meaning. This peculiarity is evident if we consider, for instance, the Italian words, *martello* (hammer), *martello pneumatico* (pneumatic drill) and *pesce martello* (smooth hammerhead): although all three have a simple word in common, their overall meanings are extremely different, Besides, the two compound words are non-ambiguous, even outside specific contexts.

As for word use contexts, especially with reference to polysemy zeroing and compound words specific meanings, it is also necessary to introduce the notions of Knowledge Domain (KD), Semantic Field (SF), and Terminology (TR).

In order to define KD, we must necessarily refer to how human knowledge can be subdivided into many and various sectors and sub-sectors, which can be interconnected to each other, to varying degrees. For instance, if we consider the KD of Medicine (MED) and its very specific knowledge, we note that it shares part of such knowledge with other domains and subdomains, to it connected more or less closely: for example, Anatomy (ANAT), Physiology (PYSIOL), Virology (VIROL), Radiology (RAD), or Nuclear Medicine (NUCLMED), and so on. From a linguistic and communicative point of view, each domain such as those just mentioned must necessarily use specific terms, due to the type of knowledge to describe and share. In a very large part, these terms are compound words,[9] which as already said, thanks to semantic expansion, as signifiers have a unique and unambiguous relationship with their meanings, a feature extremely important as regards all KDs. In this sense, we can therefore state that the SF of MED, for example, will use a specific set of unambiguous words, and that this set will be completely or partially different from that of other KDs: MED will therefore have many terms in common with ANAT and few if any with metallurgy (METALL).

Therefore, as for Lexicology, we can state that a semantic field is represented by a set of terms, simple or compound, bearing certain specifiable meaning relations to one another and used to describe the (conceptual) knowledge of a specific domain. All the sets of terms pertaining to KDs are the object of study of TR, a discipline that analyses:

- The names of objects or concepts used by different KDs;
- The functioning in the language of terminological units, as well as the problems of translation, classification and documentation that arise about them.

The ISO 1087 standard defines TR as "*the scientific study of concepts and terms used in specialized languages*". As well, TR also analyses all sets of terms, rigorously defined, which are specific to a science, a technique, or a particular field of human activity.

---

[8] Zellig S. Harris was the first to define the concept of semantic expansion. More precisely, in his paper *From Morpheme to Utterance*, he dealt specifically with nominal group semantic expansion: *"We now consider sequences of these thirty-odd morpheme classes, to see what sequences of morphemes can be substituted for single morphemes. Sequences of morpheme classes which are found to be substitutable in virtually all environments for some single morpheme class, will be equated to that morpheme class: AN=N means that* ***good boy****, for example, can be substituted for* ***man*** *anywhere. If we write DA=A (****quite old*** *for* ***old****), then DA can be substituted for A wherever A appears, e.g. in AN =N (****old fellow*** *for* ***man****, where we can substitute* ***quite old*** *for* ***old****, and obtain* ***quite old fellow*** *DAN=AN=N). There is nothing to prevent us from substituting DA for A, even in the equation DA=A. We would then obtain DDA=A:* ***really quite old*** *for* ***old****.*" (Zellig S. Harris, From Morpheme to Utterance, Language 22, 1946, pp. 161-83. Republished in *Papers in Structural and Transformational Linguistics*, D. Reidel Publishing Company, Dordrecht, Holland, 1970, pp. 108-109).

[9] Here, we give some examples of MED compound words, some of which can be used also in other SFs:

- Aarskog-Scott syndrome (MED)
- Goll kernel (MED)
- Potassium ascorbate (MED, but also Biology (BIOL) and Chemistry (CHEM))
- Morbid fear of open spaces (MED, but also Psychology / Psychoanalysis (PSIC))

Consequently, each scientific domain has and uses a specific terminological set of words. This means that we will have a distinct dictionary for each KD. Currently, 181 terminological dictionaries have been identified for Italian, corresponding to an equal number of KDs.[10] For each of these, electronic dictionaries have been created to embed in NooJ. The entries of each electronic dictionary are tagged with all the pertaining SF terminological labels.

2.3.2. Semantic Fields of Simple Words

It is possible to identify SFs also for the simple words of any language, even if the methods of identification are different from those described for compound words. Compared to that of compound ones, the number of simple terminological words is much lower. It follows that the identifiable SFs for simple words will mostly refer to relative and variable linguistic uses or semantic values, and not absolute from the point of view of meaning.
As already stated, simple words have a polysemic index higher than compound ones: that is, they are more ambiguous, because they are not subjected to semantic expansion. We achieve simple words disambiguation only by contextualising them inside sentences and discourses. This necessary contextualization, however, is extremely important for the definition of simple words SFs. For example, if we consider the two sentences:

2) *Paolo entra in ufficio* (Paul enters the office)
3) *L'ufficio ride alla battuta di Paolo* (The office laughs at Paolo's joke)

We may state that the word *ufficio* in 2) is a locative noun, as it is selected by the motion verb "*entra*"; while in 3), it is a human noun, as it is selected by the verb *ride*, which describes a specifically human activity. Aa a consequence of this, inside the Electronic Dictionary of Italian Simple Words (DELAS) these two words will be entered as follows:

Paolo,N+NPR+HUM+m+s
ufficio,N12+LOC+m+s

Here, NPR stands for "proper name", "m+s" for "masculine, singular" while N12 is the inflection code[11] for *ufficio* and all the other Italian nouns that drop the final vowel *o* to obtain their plural forms, in this case *uffici.*
2) and 3) demonstrate, once again, that the already mentioned co-occurrence and selection-restriction rules help identify the correct semantic value use of simple words, and consequently label them with SFs-specific tags. Hence, for the word *ufficio*, we will have two different tags/SFs: HUM for human, and LOC for locative. Therefore, we can rewrite formally 2) and 3) into:

4) N0Hum entra (in+nella) N1Loc
5) N0Hum ride (al+alla) N1EComOr[12] di N2Hum

Actually, these formal rewrites stand for very large sets of sentences expressing the same type of semantic relationship between their words, as for instance in:

6) *Paolo entra nella (casa+foresteria+stanza*) (Paul enters the (house+guesthouse+room))

---

[10] See Monteleone M., Elia A., De Bueriis G., Di Maio F., "Le polirematiche dell'italiano", in De Bueriis, G. (a cura di) Lessici elettronici e descrizioni semantiche, sintattiche e morfologiche. Risultati del Progetto PRIN 2005 Atlanti Tematici Informatici - ALTI, Collana "Lessici & Combinatorie", n. 2, Dipartimento di Scienze della Comunicazione dell'Università degli Studi di Salerno. Salerno, Plectica.
[11] In NooJ electronic dictionaries, an inflection code is a set of formal instructions that, starting from a lexical morpheme, allows producing all its inflected forms. For instance, form the lexical morpheme of a verb, we will use a specific inflection code to produce all the necessary attested inflected forms, in their modes and tenses. As well, from the lexical morpheme of a noun, an adjective, or other similar inflectional categories, we will produce all the necessary attested inflected forms, in their genders and numbers. See Silberztein, M. (2003) for complete instructions on how to create inflection codes with NooJ.
[12] See Table 1.

7) *Il commissariato ride al (racconto+resoconto+riassunto) di Paolo* (The police station laughs at Paul's (story+report+summary))

Therefore, if applied to the verbs, non-verbal predicates and nouns of a given language, the method of analysis sketched here allows to:

- Tag semantically and catalogue formally all simple words;
- Factorize the rules of co-occurrence and selection restriction through the association of labels referring to sets words with homogeneous characteristics, be morphosyntactic, distributional and transformational for verbs and non-verbal predicates, and semantic for nouns.

Applied to Italian, the method described here has helped to identify the list of simple words SFs shown in Table 1:

TAG	SEMANTIC FIELD DEFINITION
Abb	noun of a clothing article
AlimE	noun of an edible substance
AlimP	noun of a potable substance
ALoc	locative adverb
AMod	modal adverb
Anim	animal noun or animate non-human being noun
AQuant	quantity adverb
~~**Ast**~~	~~abstract noun~~
Assmbl	noun of object which can be manually/mechanically assembled
Atmo	noun of an atmospheric event
ATmp	time adverb
Coll	collective human noun
~~**Conc**~~	~~concrete noun~~
DDef	defined (pre)determiner
Des	obsolete term
DIndef	indefinite (pre)determiner
DNum	numerical (pre)determiner
EComOr	oral element of communication
EComScr	written element of communication
Farm	noun of a drugs or medication
Fig	figurative noun (also as a part of idiomatic sentences)
Gramm	noun of a grammatical, morphological, syntactic element
HDisc	noun referring to investigative/contemplative humanistic disciplines
Hum	human noun
Lin	noun of a tongue, a dialect, a jargon
Loc	locative noun, noun of a place
Lud	game/sport noun
Mal	illness/disease noun
Mass	mass noun
Mis	unit of measure noun
Mon	currency noun
Mus	musical instrument noun
Num	numeric noun
PC	body-part noun
Psic	noun of sentimental, psychic or psychological state
Qual	noun expressing a quality
QuantD	Conventionally-defined quantifier noun
QuantI	undefined quantifier noun
SDisc	noun referring to investigative scientific disciplines

SostG	noun of a gaseous state substance
SostL	noun of a liquid-state substance
SostS	noun of solid-state substance
Strum	mechanical tool noun
Tmp	noun of a period of time/event with defined or undefined duration
Veg	noun of plants, vegetables, flowers

Table **1.** Tag list for Italian simple-word knowledge domains

We note that in this list the classic opposition between the concepts "concrete" and "abstract" has been annulled, in the sense that it has been "diluted" into more specific conceptual subcategories. This annulation is useful in different ways. First, from a logical-deductive point of view, it bypasses the difficulty of defining precisely when and by means of which semantic characteristics a noun is "concrete" or "abstract". Second, by cancelling this opposition, it is possible to bring Logic, with its specific features, slightly closer to NLP.
As already stated, using these labels within NooJ electronic dictionaries is very useful for obtaining an even more precise and effective ATA. However, before moving on to the practical demonstration of such advantages, it is necessary to do a quick assessment of the prevalence of morphosyntax in the building of acceptable and grammatical sentences and discourses.

3. Morphology, Syntax and Semantics in Sentence Formation

As is well known, in all languages, be they isolating, agglutinating, inflectional or polysynthetic, a clear-cut separation between morphology, syntax and sentence semantics is possible only theoretically. In fact, in both spoken and written language, the creation of semantically and grammatically acceptable sentences and discourses takes place through the concatenation of morpholexical elements. Such a concatenation is governed by the rules of co-occurrence and selection-restriction. Co-occurrence and selection restriction rules are distributional and transformational rules used during sentence/discourse formation and based on the syntactic-semantic properties of predicates. More precisely, they define how each lexical element, when used in sentences and discourses, brings as a contribution some of its morphological, grammatical and semantic characteristics, i.e. those selected by the predicate syntactic profile and at the same time required by the meaning to be produced.
As for this theme, Z. S. Harris[13] was the first to admit the necessity of representing lexical restrictions:

"We now have sets of morphemic components (and residues), so set up that as nearly as possible all sequences and combinations of them occur (...) It may not be convenient to represent by means of components such limitations of occurrence among morphemes as do not intersect with other limitations involving the same morphemes, or as do not lead to the division of a class into sub-classes clearly differentiated on that basis (...) This is frequently the case for morphemes classes which are grouped together into a general class on the basis of major similarities, but which have small and unpatterned differences in distribution." (pp. 309-312).

Some years later, this approach was formalized and applied to many languages thanks to Maurice Gross's Lexicon-Grammar.
The importance of co-occurrence and selection-restriction rules is more evident in sentences such as the following:

8) The dog chases John and Paul

Versus

9) John and Paul chase the dog

[13] See Harris, Z. S. (1951). *Structural linguistics*. Chicago, University of Chicago Press.

We can see clearly how sentence semantics always comes from both the distribution of words and the morphosyntactic relationships between them. As for the subject and/or complement slots, in the two previous sentences the verb *to chase* can select either a human noun (the group *John and Paul*) or an animated noun (*the dog*). This allows for the mirroring of the two complements and for reversal of the semantic roles of *John* and *the dog*. The result are two sentences that are semantically and grammatically acceptable, syntactically very similar, but have opposed meanings. Moreover, these two sentences allow us to determine that every change of form (syntax) corresponds to a change of content (semantics / meaning of the sentence), be it slight or considerable.
Therefore, we can state that the final syntactic form of acceptable and grammatical sentences is always the result of the interaction between morphology (in terms of morpholexical elements) and syntax. Hence the use of the word "morphosyntax" to describe such interactions.

3.1. Morphosyntax Prevalence in Sentence Formation

The use of English collective nouns[14] is a classic example of morphosyntax prevalence in sentence formation. An English collective noun is a noun that represents a collection of individuals, usually people (more than two persons), such as a team, a family, a crew, and so on. Besides, English collective nouns are formally used without or do not have plural mark, but may behave as plural nouns. If we consider the following English sentence:

10) *Police **are** all over the place*

We may notice that a morphologically singular word (*police*) has actually a plural value, hence it must co-occur with a verb in the plural form (*are*). At the same time, we note the following translation from Italian to English:

11) *Police **forces** of all countries **have** been alerted =: Le **polizie** di tutte le nazioni sono state allertate*

In which the English plural form police forces corresponds to the Italian plural word *polizie*.
Other examples of collective nouns are the *words committee, jury, senate, company, audience, police, army*, and so on. Many collective nouns are common nouns, but they can also be proper nouns when they are the name of a company or other organisation with more than one person, for example *Microsoft, IBM, Apple, The Financial Times, The United Nations, Cambridge University, Manchester United*, and so on. A collective noun can be singular or plural, depending on our view of the individuals in the group. If we consider the individuals as acting together, then it is possible to treat the collective noun as singular (with singular verbs and singular pronouns), as in:

12) *The **jury has** delivered **its** conclusion to the judge*

On the contrary, if we consider the individuals as acting individually, then it is possible to treat the collective noun as plural (with plural verbs and plural pronouns), for example:

13) *The **jury have** not reached a conclusion because **they are** still arguing among **themselves***

British English tends to treat collective nouns as plural, while American English tends to treat them as singular. Therefore, in the example above, American English speakers might use a singular verb with *jury* and rephrase the rest of the sentence to avoid a logical incongruity:

14) *The **jury has** not reached a conclusion because **its members are** still arguing among **themselves**.*

However, even in American English, it is acceptable to use a plural verb to emphasize the individuality of the collective noun members:

15) *The San Francisco **crowd were their** usual individualistic selves*

In American English, it is also possible to use a plural pronoun with a singular verb, as in:

[14] See https://www.englishclub.com/grammar/nouns-collective.php.

16) *The family next door **is** very quiet. We never hear **them**.*

In all varieties of English, the collective noun *police* is always treated as plural:

17) *The **police are** coming.*
18) *The **police were** the first on the scene*
19) *The **police have** issued **their** report*

In most cases, a collective noun can itself be plural, i.e. it is possible to have more than one collective noun (two teams of a football game, many families in a street there are many families). In such cases, a plural verb is used automatically, as in these examples:

20) *The many London football **teams were** playing a friendly tournament*
21) *The two **families have** been quarreling for over a week*

In the following, we give some more examples of collective noun treated as both singular and plural depending on sentence semantics:

22) *The club **was founded** in 2003 / The club **are** currently displaying their best photos*
23) ***Does** Sony make mobile phones? / Do Sony plan to make cars?*
24) *The board of directors **uses** this room for its meetings / The board of directors **are** eating sandwiches for their lunch*
25) *The family next door **is** very quiet. We never hear **them** / My family **are** always arguing. The neighbours often hear **us***
26) *The school **reopens** in September / The school **are** preparing for their winter marathon*
27) *CNN **does** like to blow its own trumpet / CNN **do** like to blow their own trumpet*

From a morphosyntactic point of view, as for collective nouns, Italian has less particularity than English. For example, all logically collective nouns like "*polizia*" (police officers), "*giornale*" (newspaper workers) or "*famiglia*" (family members), always co-occur with singular verb forms, while in the plural they simply have the meaning of "more than one" and always co-occur with plural verb forms.
Moreover, the classic English opposition between countable / uncountable nouns is also less marked in Italian, even if it presents some interesting peculiarities, which can be accounted for inside NooJ electronic dictionaries, by means noun inflection codes. For instance, in the following two sentences:

28) *Paolo mangia (E+del) vitello* (Paul is eating veal)

Vitello is intended as a mass noun, in the sense of "calf meat", contrary to what happens instead in the sentence:

29) *Paolo possiede un vitello* (Paul owns a calf)

In which *vitello* is an animate countable noun. We add that in Italian, all the nouns that indicate edible animal meat,[15] being mass nouns, to preserve this meaning must be used only in the singular. Often, they are not introduced by any determinant, or are preceded by a generic predeterminant.
Therefore, in the case of veal and of all the nouns that simultaneously indicate animals and food animal meats, we will have the following entries in an electronic dictionary:

vitello,N+N7+ANIM[16]*+m+s*
vitello,N+N608+MASS[17]*+m+s*

While the inflectional code N7 accounts also for the plural form *vitelli*, N608 provides only the singular noun *vitello*.
Other examples of strict selection restrictions of some verbs on specific Italian nouns are listed below.

[15] In this sense, very particular is the case of the noun *maiale*, which can simultaneously indicate a type of edible meat (pork, in English), an animal (pig, in English), a very dirty individual, which arouses repugnance (pig or slob, in English), or an individual who eats with animal greed (glutton, in English).
[16] See Table 1.
[17] See Table 1.

- *Paolo respira SOSTG*[18] *(i.e. ossigeno+nitrogeno+butano+etano+metano…)* (Paolo breaths (oxygen + nitrogen + butane + ethane + methane…))
- *Paolo beve SOSTL (i.e. vino+acqua+birra+aranciata+benzina…)* (Paolo drinks (wine+water+orange juice+fuel…))
- *Paolo paga in MON+p (i.e. dollari+corone+dracme…)* (Paul pays in (dollars + crowns + drachmas…)

In the previous list, the semantic relationships between verbs and nouns are immutable. In fact, if we try to replace the SFs tags, we produce unacceptable/doubtful sentences:

- **Paolo beve SOSTG (i.e. ossigeno+nitrogeno+butano+etano+metano…)* (*Paolo drinks (oxygen + nitrogen + butane + ethane + methane…))
- **?Paolo respira SOSTL (i.e. vino+acqua+birra+aranciata+benzina…)* (Paolo breaths (wine+water+orange juice+fuel…))
- **Paolo paga in MON+s (i.e. dollaro+corona+dracma…)* (*Paul pays in (dollar+crown+drachma…)

3.2. Some Examles of NooJ Grammars

We give here are some examples of NooJ syntactic grammars in which, using the tags of Table 1, we prepare the automatic recognition and annotation of groups of sentences the verbs and nouns of which are in the same type of semantic relationship:

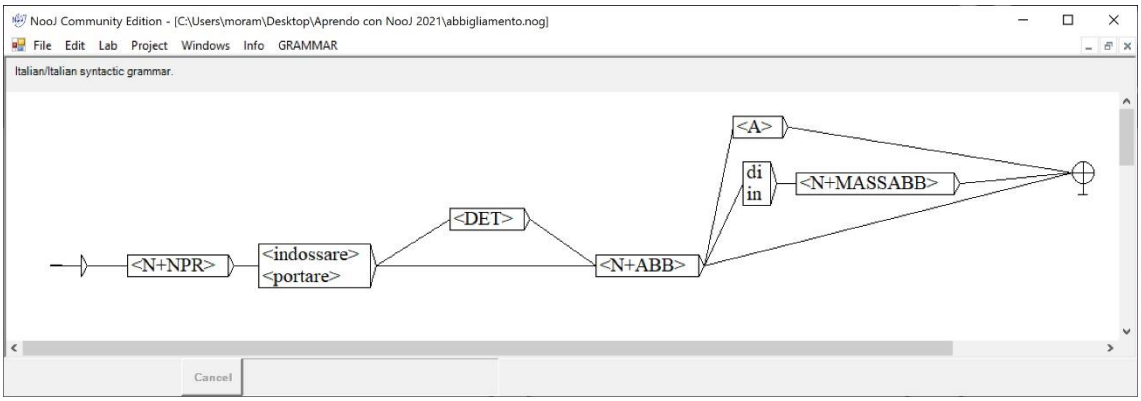

Figure **1**.: NooJ syntactic grammar for ABB SF

The grammar in Figure 1. processes sentences of the type:

- *Paolo porta delle scarpe (di+in) cuoio* (Paolo wears leather shoes)
- *Paolo indossa una camicia blu* (Paolo wears a blue shirt)
- *Paolo porta un maglione di lana* (Paolo is wearing a wool sweater)

[18] See Table 1.

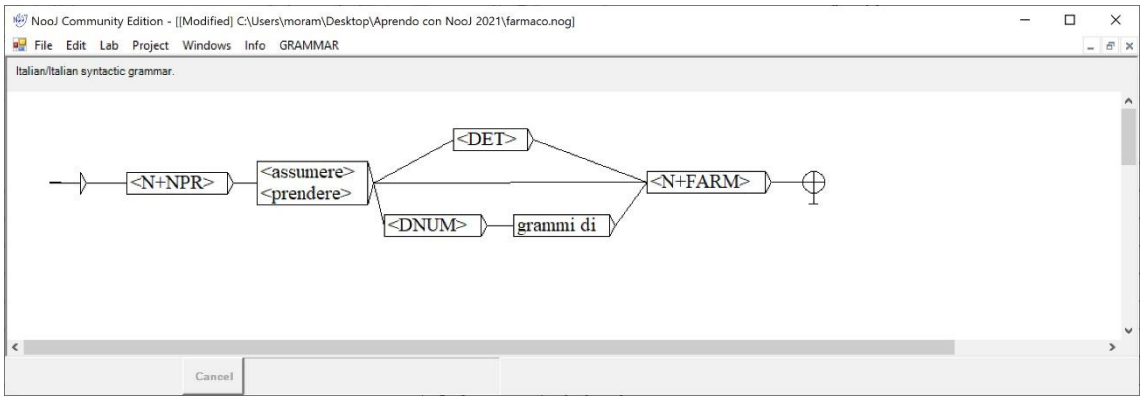

Figure **2**.: NooJ syntactic grammar for FARM SF

The grammar in Figure 2. recognises sentences of the type:

- *Paolo prende Ibuprofene* (Paul takes Ibuprofen)
- *Paolo assume*[19] *dieci grammi di cortisone* (Paolo takes ten grams of cortisone)
- *Paolo assume un antipiretico* (Paul takes an antipyretic)

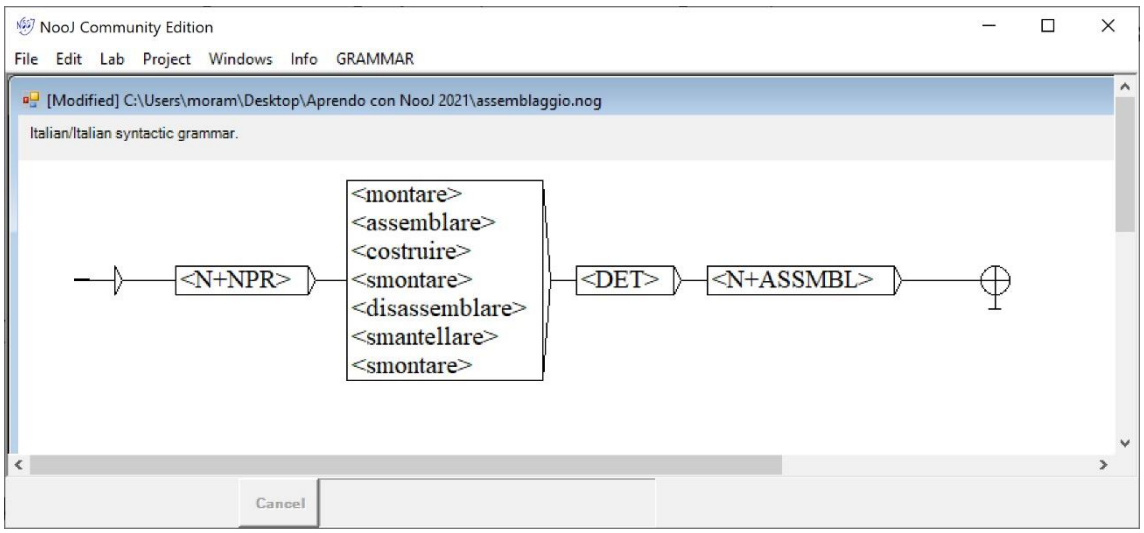

Figure **3**.: NooJ syntactic grammar for ASSMBL SF

The grammar in Figure 3. processes sentences of the type:

- *Paolo (assembla+disassembla) il treno* (Paolo (assembles+disassembles) the train)

This grammar has two peculiarities:

- It provides for the use of verbs with opposite but contiguous meanings, such as for example *installare* (to install) and *disinstallare* (to uninstall);
- It contributes to the disambiguation of the Italian verb *montare* in sentences such as:
    - *Paolo monta **il** treno* (Paolo mounts the train)
    - *Paolo monta **sul** treno* (Paolo jumps on the train)

[19] In this case, we note that thanks to the relationship with the N+FARM, the Italian verb *assumere* takes the meaning of "to ingest" and not of "to employ", which is by far more commonly used in Italian.

In which *treno* may be either an object to mount or a place.

4. Conclusions

The three previous grammars can account for the recognition and annotation of a high number of sentences. This is clearly the most important result of the whole procedure outlined here, because it allows us to effectively process large amounts of data with minimal effort, namely that of labelling simple words in an electronic NooJ dictionary.
In addition to this, it is worth remembering that with NooJ, creating both specific transformation rules and the Grammar-> Generate option, in any language it is possible to produce a very large quantity of acceptable and grammatical, sentences starting from grammars as those of Figures 1., 2., and 3. This possibility makes the method outlined in these pages even more effective in terms of linguistic data coverage.

5. References

https://en.wikipedia.org/wiki/Formal_semantics_(natural_language)

Monteleone, M. (2003). *Lexicographie et dictionnaires électroniques. Des usages linguistiques aux bases de données lexicales*. Université de Marne-la-Vallée. Thèse pour obtenir le grade de Docteur de l'Université de Marne-la-Vallée, présentée et soutenue publiquement par M. Mario Monteleone, 8 décembre 2003. Available for download at https://tel.archives-ouvertes.fr/tel-00627599/document.

Harris, Z. S. (1946). "From Morpheme to Utterance", Language 22, pp. 161-83.
Harris, Z. S. (1951). *Structural linguistics*. Chicago, University of Chicago Press.

Harris, Z. S. (1970). *Papers in Structural and Transformational Linguistics*, D. Reidel Publishing Company, Dordrecht, Holland.

Monteleone M., Elia A., De Bueriis G., Di Maio F. (2005). "Le polirematiche dell'italiano". In De Bueriis, G. (a cura di) *Lessici elettronici e descrizioni semantiche, sintattiche e morfologiche. Risultati del Progetto PRIN 2005*. Atlanti Tematici Informatici - ALTI, Collana "Lessici & Combinatorie", n. 2. Dipartimento di Scienze della Comunicazione dell'Università degli Studi di Salerno. Salerno, Plectica.

Wikipedia Encyclopedia, entry *Lexicon-Grammar*, available at the page https://en.wikipedia.org/wiki/Lexicon-grammar.

Z. S. Harris Web Site Homepage www.zelligharris.org.

Wikipedia Encyclopedia, entry *Lucien Tesnière*, available at the page https://en.wikipedia.org/wiki/Lucien_Tesnière.

Silberztein, M. (2003 -). *The NooJ Manual*. Available for download at: http://www.nooj-association.org

Silberztein, M. (2007). *An Alternative Approach to Tagging*. Proceedings of NLDB 2007. LNCS Series, Springer-Verlag Eds, 1-11.

Silberztein, M. (2007). *Les unités linguistiques et leur annotation automatique*. Modèles Linguistiques n. 55.

Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. ISTE Ltd and John Wiley & Sons Inc, London. EAN: 9781848219021

Harris, Z. S. (1951). *Structural linguistics*. Chicago, University of Chicago Press.

English Club Web Site, page https://www.englishclub.com/grammar/nouns-collective.php.