THE TERMINOLOGICAL TAGGING OF THE NOOJ ITALIAN COMPOUND WORD DICTIONARY

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

Abstract

In this paper, and in relation to the construction of electronic dictionaries for NooJ, we will deal with the tagging of Italian compound words, and with how it differs from that of simple words as for methods, functions, and purposes. We will especially focus our attention on the tagging of technical-scientific compound words, demonstrating how this operation, in NooJ, represents a crucial tool for both information extraction and knowledge automatic management and representation. Furthermore, with the intention to producing a complete analysis, we will provide the definitions of simple word and compound word, from both a formal and a linguistic point of view. As for the linguistic examination, we will adopt two different approaches. For the first one, we will use the analytic methods of Zellig S. Harris, who first set out, in structuralist terms and in relation to English, the study of the composition of different morphemes in more complex linguistic units, hence also of word groups or phrases. For the second one, we will make extensive reference to the methodological framework of language formalization described by Maurice Gross' Lexicon-Grammar, as to its subsequent adaptations to the Italian language. Finally, as we will see, it will be of fundamental importance for us to differentiate the definitions that we will give here of compound words from the more generic and less precise one of multiword expressions (MWE). To justify this differentiation, we will provide not only formal indications, but also lexical, morphosyntactic and semantic ones.

Keywords: NooJ, Lexicon-Grammar, NooJ Electronic Dictionaries, Compound Words, Terminology

EL ETIQUETADO TERMINOLÓGICO DEL DICCIONARIO ITALIANO DE PALABRAS COMPUESTAS DE NOOJ

Resumen

En este artículo, y en relación con la construcción de diccionarios electrónicos para NooJ, trataremos el etiquetado de palabras compuestas en italiano y cómo se diferencia del de palabras simples en cuanto a métodos, funciones y propósitos. Centraremos especialmente nuestra atención en el etiquetado de palabras compuestas técnico-científicas, demostrando cómo esta operación, en NooJ, representa una herramienta crucial tanto para la extracción de información como para la gestión y representación automática del conocimiento. Además, con la intención de producir un análisis completo, daremos las definiciones de palabra simple y palabra compuesta, tanto desde el punto de vista formal como lingüístico. En cuanto al examen lingüístico, adoptaremos dos enfoques diferentes. Para el primero, utilizaremos los métodos analíticos de Zellig S. Harris, quien por primera vez planteó, en términos estructuralistas y en relación con el inglés, el estudio de la composición de diferentes morfemas en unidades lingüísticas más complejas, por lo tanto, también de grupos de palabras. o frases. Para el segundo, se hará amplia referencia al marco metodológico de formalización de la lengua descrito por el Léxico-Gramática de Maurice Gross, así como a sus posteriores adaptaciones a la lengua italiana. Finalmente, como veremos, será de fundamental importancia para nosotros diferenciar las definiciones que daremos aquí de palabras compuestas de la más genérica y menos precisa de expresiones multipalabras (MWE). Para justificar esta diferenciación aportaremos no sólo indicaciones formales, sino también léxicas, morfosintácticas y semánticas.

Palabras clave: NooJ, Léxico-Gramática, diccionarios electrónicos NooJ, palabras compuestas, terminología

1. Definition of Simple Word, Compound Word and Multiword Expression

1.1. Simple Words

As for written language, a simple word is any meaningful sequence of letters whose segmentation into immediate constituents almost never produces other words, but morphemes, mainly lexical and grammatical, and also derivational. Similar segmentation results are very frequent in Romance languages such as, for instance, Italian and Spanish, and in contrast to English, where lexical morphemes are very rarely also autonomous words, as shown by the following quick example:

It. ragazz-o = Sp. muchach-o = Eng. boy
It. fatal-mente[1] *= Sp. fatal-mente = Eng. fatal-ly*

Thus, a simple word is a linguistically certified, inalterable and semantically autonomous concatenation of morphemes, inside which no other lexical material can be added. This same definition is therefore also valid for orthographic or formal words (sequences of characters endowed with a meaning or a grammatical function and enclosed between two blank spaces), and morphological words (words composed by more than one morpheme and enclosed between two blank spaces). It is also valid for monorhematic words (dictionary entries consisting of a single word), also called Simple Atomic Linguistic Units (SALUs) or simple lexemes (units of the lexicon, in general a simple word coinciding with the lemma of a dictionary, which can also be a radical thematic basis that cannot be analyzed further). Finally, in all inflectional languages, simple word forms are obtained through the inflection of monorhematic words/SALUs/simple lexemes, with which they complete the set of simple words of a specific language.

1.2 Compound Words and Multiword Expressions

The first attestation relating to groups of words is due to Z. S. Harris,[2] who in 1946 mentioned, for English, the possibility that one or more morphemes did combine in sequences, or phrases, to create unities of meaning. More specifically, in "From Morpheme to Utterances", Harris first defines the concept of free sequences of simple words, stating that he wants to present *a formalized procedure for describing utterances directly in terms of sequences of morphemes rather than of single morphemes. It thus covers an important part of what is usually included under syntax. When applied in a particular language, the procedure yields a compact statement of what sequences of morphemes occur in the language, i.e. a formula for each utterance (sentence) structure in the language. At present, morpheme classes are formed by placing in one class all morphemes, which are substitutable for each other in utterances, as "man" replaces "child" in "The child disappeared". The procedure outlined below consists, essentially, in extending the technique of substitution from single morphemes (e.g. "man") to sequences of morphemes (e.g. "intense young man"). In so far as it deals with sequences, it parallels the type of analysis frequently used in syntax, so that the chief usefulness of this procedure is probably its explicitness rather than any novel of method or result.*[3]

Harris also defines a possible typology of word groups. We may have:

1 Among all possible exception in Italian and Spanish, we find the adverbs of manner obtained through the suffix *-mente*. In fact, the segmentation of these adverbs highlights the suffix itself together with a word form, namely a feminine singular adjective, as in *It. corretta-mente = Sp. correcta-mente*.
2 See Harris Z.S. (1970), especially the 1946 article titled "From Morphemes to Utterances", followed by "Componential Analysis of a Paradigm" and "Immediate-Constituent Formulation of English Syntax".
3 *Ibid.* p. 100. To preserve the clarity of the quotation, we emphasize here once again that in English there is very often perfect coincidence between lexical morpheme, simple word and lexeme.

- Nominal groups (*white dog, defence dog*);
- Verbal groups (*to back off, to kick the bucket*);
- Adjectival groups (*bright red, India red*);
- Adverbial groups (*on the stairs, in spare time*); prepositional groups (*in the middle of, in the direction of*);
- Conjunctions groups (*and again, even if*);
- Exclamatory groups (*ooh-la-la, for God's sake*);
- Determiner group (*some kind of, a lot of*);
- Pronominal groups (*his own, you both*).

Subsequently,[4] while taking its cue from Harris' definition, Maurice Gross' Lexicon-Grammar adds an important differentiation between compound words and free groups of simple words. Indeed, a formal differentiation would not suffice to separate the formers from the latters. For instance, the two noun groups *small house* and *public house* are structurally identical (composed by an adjective followed by a noun), but while the first group refers to a house that has small dimensions – hence has a compositional meaning, the second refers to an establishment providing alcoholic beverages to be consumed on the premises – hence, has a non-compositional meaning. To stress such difference, Gaston Gross (1986) proposes to adopt a series of syntactical criteria useful to distinguish compound nouns[5] from free noun groups of the same structure. Here, we recall the general principle of distinction:

A sequence of simple words is fixed (or compound) if at least one of its syntactic, distributional or semantic properties cannot be deduced from the properties of its constituents.

Therefore, considering the non-compositional compound word *piece of cake* (used with reference to something that is very easy to manage), we notice that it can be used as a noun ("The English test was a piece of cake"), or and as an adjective ("This is a piece-of-cake situation"). Therefore, we can say that *piece of cake* is a compound word because we cannot deduce its distinctive morphosyntactic properties and uses from the simple words that form it. The same may be said for *Green Beret* or *Red Army*, which despite their adjective-noun composition are human nouns that can be used as subjects inside declarative sentences, as *The Green Berets/The Red Army attacked the enemy*. Hence, we will classify them as compound words because we cannot deduce these distributional properties from the words *green, berets, red* or *army*, neither can we paraphrase the two sequences in *berets that are green* or *army that is red*. On the other hand, very different and somewhat less fine-grained is the definition generally given of MWEs, since they are supposed to include all "expressions which are made up of at least 2 words and which can be syntactically and/or semantically idiosyncratic in nature. Moreover, they act as a single unit at some level of linguistic analysis."[6] Hence, MWEs include:

- Fixed expressions, i.e. fully lexicalized expressions that can neither be morphosyntactically varied nor modified internally (i.e. *in short, by and large, every which way*);
- Semi-fixed expressions, in which word order and composition are strictly invariable, inflection is possible (i.e. *kick the bucket* can be inflected as *he kicks the bucket*) while some transformations are not allowed (as in * *the bucket was kicked by Paul*, which does not have the same meaning);
- Compound nominal semi-fixed expressions, as *car park* or *peanut butter*, which are syntactically unalterable but can inflect for number (i.e. *two car parks*);
- Proper names, which are also semi-fixed expressions, as can occur in different forms (i.e. the name of the U.S. sports team the *San Francisco 49ers* can occur as the *49ers* or as a modifier in the compound noun a *49ers player*);

4 See Gross, G. (1986) and Silberztein, M. (1989).

5 More specifically, those of the formal kind Noun of Noun, as *kettle of fish*, or *house of cards* from *months of the year*, or *wheel of a car*.

6 See https://aclweb.org/aclwiki/Multiword_Expressions.

- Syntactically-flexible expressions, which have a wider range of syntactic variability than semi-fixed expressions. They occur in the form of decomposable idioms, verb-particle constructions and light verbs:
    - Decomposable idioms are likely to be syntactically flexible to some degree (i.e. *let the cat out of the bag* and *sweep under the rug*). Yet, it is hard to predict which kind of syntactic variation a given idiom can undergo.
    - Verb-particle constructions (i.e. *write up* and *look up*, which are made up of a verb and one or more particles). Either they are semantically idiosyncratic as *brush up on* or compositional as *break up* in *the meteorite broke up in the earth's atmosphere*.
    - Light verb constructions, i.e. *to make a mistake, to give a demo*. Though they are highly idiosyncratic they have to be distinguished from idioms;
- Institutionalized or conventionalized phrases, such as *salt and pepper, traffic light, to kindle excitement*, which are semantically and syntactically compositional, but statistically idiosyncratic.

Since it copes with all grammatical categories, MWE definition is actually much broader than the one provided by Lexicon Grammar for compound words. In our opinion, this definition creates confusion and does not help define a coherent, cohesive and effective approach to the same MWEs. First, contrarily to what previously seen with Lexicon-Grammar as for distributional and transformational syntax, MEWs miss specific tests that could help accurately separate them from non-MWEs. Second, the line of demarcation between free word groups and MWEs is not well defined. Hence, it is invalidated that part of the explanation in which a MEW is defined as a "lexeme-like unit made up of a sequence of two or more lexemes that has properties that are not predictable from the properties of the individual lexemes or their normal mode of combination". Finally, different linguistic levels of analysis are improperly associated, while they would require specific analytical tools. In fact, when we speak, for example, of light verbs, which correspond to Lexicon-Grammar support verbs,[7] or of idiomatic sentences, we enter more closely into a pure syntactic sphere, while the phenomenon of MWEs occurs, fundamentally, as a syntagmatic lexical manifestation, therefore impossible to deal with in the same way as sentences. Therefore, in this paper, we will stick to Lexicon-Grammar definitions and use the term "compound word" instead of MWE, as this will help us producing a more detailed and clear analytic study.

2. Compound Words and Terminology

In all languages, there is a close relationship of necessity between compound word and terminology.[8] In fact, terminology needs compound words, and this is confirmed by the presence in specialty lexicons[9] of a very large number of compounds, in some cases more than 90% of the entire lexical group listed. However, it is worth remembering that the use of compound words is widely attested even in registers that are not terminologically marked, although simple and monorhematic compound words are more widespread in them.

7 See Gross M. (1975). Méthodes en syntaxe, Paris, Cantilène.

8 As is known, terminology is composed by those sets of terms that refer to concepts and tools belonging to a particular sector of knowledge or human activity. Terminology consists of common words to which a specific meaning is attributed, of borrowings, calques, and more rarely of real neologisms. The set of terms that form a terminological domain must be unambiguous, i.e. they must have a single precise meaning for all specialists in that sector. Actually, this is not always possible, particularly in human sciences. Periodically, the associations that bring together the experts of each sector review its terminology to update it, eliminate eventual problems of polysemy, and draw up specialty dictionaries, which carry out a uniforming action.

9 Specialty lexicons are all those homogeneous lexical subsets that contain terms used specifically and in a semantically univocal way within the various domains of knowledge. In this sense, the domain of knowledge - or semantic field - of the economy will have its own specialized lexicon, and the same will be true for physics, biochemistry, geodesy, and so on. Furthermore, these lexical subsets are generally cataloged and described within specific paper works, also called specialty dictionaries.

As highlighted by Lexicon-Grammar, terminological compound words,[10] contrary to simple words of generic use, are not polysemic, as they can be tagged univocally – that is, even though they belong to different semantic fields, in each of them they will have on one and only one meaning. This characteristic is of great value for terminological language, which needs to be as precise as possible in the combination of their signifieds and signifiers. The main purposes of terminology in fact concern the unambiguous classification of objects and concepts, and therefore in the second analysis the achievement of a non-dysfunctional technical-scientific communication. Terminological language, by definition, cannot be ambiguous, and therefore finds in compound words the most adequate and suitable form of linguistic formulation. It is so worth noting that thanks to compounds syntagmatic structure, in a given terminological field, it is possible to build meaning open series, useful to define conceptual subsets and establish logical-inclusive relationships, creating terminological cognitive networks of which the different nodes are compound words having specific lexical elements in common. This is the case of an Italian and English open series of compounds, belonging to the specialized lexicon of Psychology (PSIC), made up of fifty-nine entries and that share, as head element, the nominal group ***paura morbosa*** (morbid fear):

ITA	ENG
paura morbosa degli spazi aperti	morbid fear of open spaces
paura morbosa dei bambini	morbid fear of children
paura morbosa dei cani	morbid fear of dogs
paura morbosa dei colori	morbid fear of colors
paura morbosa dei fiori	morbid fear of flowers
paura morbosa dei gatti	morbid fear of cats
paura morbosa dei pesci	morbid fear of fish
paura morbosa dei precipizi	morbid fear of precipices
paura morbosa dei pulcini	morbid fear of chicks
paura morbosa dei ragni	morbid fear of spiders
paura morbosa dei serpenti	morbid fear of snakes
paura morbosa dei suoni	morbid fear of sounds
paura morbosa dei vermi	morbid fear of worms
paura morbosa del buio	morbid fear of the dark
paura morbosa del calore	morbid fear of heat
paura morbosa del colore	morbid fear of color
paura morbosa del denaro	morbid fear of money
paura morbosa del disordine	morbid fear of disorder
paura morbosa del dolore	morbid fear of pain
paura morbosa del freddo	morbid fear of cold
paura morbosa del fuoco	morbid fear of fire
paura morbosa del mare	morbid fear of the sea
paura morbosa del matrimonio	morbid fear of marriage
paura morbosa del peccato	morbid fear of sin
paura morbosa del piacere	morbid fear of pleasure
paura morbosa del ridicolo	morbid fear of ridicule
paura morbosa del sesso	morbid fear of sex
paura morbosa del sonno	morbid fear of sleep
paura morbosa del tuono	morbid fear of thunder
paura morbosa del veleno	morbid fear of poison
paura morbosa del vento	morbid fear of the wind
paura morbosa del vetro	morbid fear of glass
paura morbosa dell'amore	morbid fear of love
paura morbosa dell'errore	morbid fear of making a mistake
paura morbosa dell'idrofobia	morbid fear of hydrophobia
paura morbosa dell'infinito	morbid fear of infinity

10 For an assessment of this type of simple words and their polysemy, see Gross (1989).

paura morbosa della crescita	morbid fear of growing up
paura morbosa della divinità	morbid fear of divinity
paura morbosa della fatica	morbid fear of fatigue
paura morbosa della felicità	morbid fear of happiness
paura morbosa della folla	morbid fear of crowds
paura morbosa della gente	morbid fear of people
paura morbosa della gravità	morbid fear of gravity
paura morbosa della lebbra	morbid fear of leprosy
paura morbosa della luce	morbid fear of light
paura morbosa della nebbia	morbid fear of fog
paura morbosa della neve	morbid fear of snow
paura morbosa della pioggia	morbid fear of rain
paura morbosa della polvere	morbid fear of dust
paura morbosa della profondità	morbid fear of depth
paura morbosa della responsabilità	morbid fear of responsibility
paura morbosa delle api	morbid fear of bees
paura morbosa delle deformità	morbid fear of deformities
paura morbosa delle feci	morbid fear of feces
paura morbosa delle foreste	morbid fear of forests
paura morbosa delle infezioni	morbid fear of infections
paura morbosa delle malattie	morbid fear of disease
paura morbosa delle scale	morbid fear of stairs
paura morbosa di tutto	morbid fear of everything

3. The NooJ Italian Electronic Dictionary of Compound Words

The Italian Electronic Dictionary of Inflected[11] Compound Words (DELACF) for NooJ is essentially terminological, therefore in it entries may have also more than one terminological tag, based on the sector(s) of knowledge in which they have has been attested. Currently, the DELACF includes 283.641 entries, subdivided into 173 different sectors of knowledge, as shown in the following table:

Tag	Italian Knowledge Domain	English Knowledge Domain
ABB	Abbigliamento	Clothing
ACC	Accessori	Accessories
ACUS	Acustica	Acoustics
AGR	Agricoltura	Agriculture
ALIM	Alimentazione	Diet
ANAT	Anatomia	Anatomy

11 It is worth noting that the inflectional typology of compound words is very wide-ranging, as it does not always require the pluralization of all the constituent elements, and is often governed by the categorical properties of the constituents themselves. This is confirmed, for example, by the compound *casa di cura* (nursing home), which is of the type NDN (name-preposition DI-name). It has a plural form in *case di cura* (nursing homes), while the form **case di cur**__**e**__ is not acceptable. Again, the aforementioned compound *paura morbosa degli spazi aperti* (morbid fear of open spaces), which is of the type NAPADINA (noun-adjective-name-articulated preposition DI+gli-name-adjective type). It has no plural form, as **paur*__*e*__ *morbos*__*e*__ *degli spazi aperti* is not acceptable, while **paura morbosa dell*__*o*__ *spazi*__*o*__ *apert*__*o*__ is not attested. At the same time, instead, compounds of the noun-adjective type, such as *berretto verde* (green cap), pluralize both constituent lexical units, as in *berretti verdi*. We also underline that to automatically inflect DELACF entries, we use a morpho-grammatical and inflectional description based on a binary matrix, in which the letter *m* indicates a masculine form, the *f* a feminine form, the *s* a singular one and the *p* a plural one. To these letters, we add the use of the signs + and -, which provide further information on the possible inflected forms. Thus, a tag like *fs*-+ will indicate that the compound word is feminine singular (*fs*), has no corresponding masculine form (as indicated by the - sign), and has instead a feminine plural form (as indicated by the + sign). Similarly, a tag like *mp*-- will indicate that the word is masculine plural, and that it has no other inflection form, either masculine or feminine.

ANTROP	Antropologia	Anthropology
ARALD	Araldica	Heraldry
ARCH	Architettura	Architecture
ARCH NAV	Architettura navale	Naval architecture
ARCHEOL	Archeologia	Archeology
ASTROFIS	Astrofisica	Astrophysics
ASTROL	Astrologia	Astrology
ASTRON	Astronomia	Astronomy
AUT	Automatismi	Automation
AUT IND	Automatismi industriali	Industrial automations
AUT INT	Automatismi intelligenti	Intelligent automations
AUT UFF	Autmatismi per ufficio	Office autmatisms
BASI DATI	Basi di dati	Databases
BIOCHIM	Biochimica	Biochemistry
BIOFIS	Biofisica	Biophysics
BIOL	Biologia	Biology
BIOL MOL	Biologia molecolare	Molecular biology
BOT	Botanica	Botany
CARTOGR	Cartografia	Cartography
CHIM	Chimica	Chemistry
CHIM ANAL	Chimica analitica	Analytical chemistry
CHIM FIS	Chimica fisica	Physical chemistry
CHIM INORG	Chimica inorganica	Inorganic chemistry
CHIM ORG	Chimica organica	Organic chemistry
CITOL	Citologia	Cytology
CLIMATOL	Climatologia	Climatology
COMUN	Comunicazioni	Communications
COSM	Cosmologia	Cosmology
COSTR AER	Costruzioni aeronautiche	Aeronautical construction
COSTR NAV	Costruzioni navali	Shipbuilding
CRIOGEN	Criogenia	Cryogenics
CRISTAL	Cristallografia	Crystallography
CRONOM	Cronometria	Timekeeping
DANZA	Danza	Dance
DIGE	Dizionario generico	General dictionary
DIR	Diritto	Right
DISP	Dispositivi	Devices
DISP ELAB	Dispositivi elaborazione dati	Data processing devices
ECOL	Ecologia	Ecology
ECON	Economia	Economy
EDIL	Edilizia	Building
EDIT	Editoria	Publishing
ELAB	Elaborazione dati	Data processing
ELAB DISTR	Elaborazione dati distribuita	Distributed data processing
ELETTR	Elettricità	Electricity
ELETTROMAG	Elettromagnetica	Electromagnetics
ELETTRON	Elettornica	Electronics
EMBRIOL	Embriologia	Embryology
ENOL	Enologia	Enology
EVOL	Evoluzionismo	Evolutionism
FANT	Fantastico	Fantastic
FARM	Farmacologia	Pharmacology
FERR	Ferrovia	Railroad
FIG	Figurato	Figured
FIS	Fisica	Physics
FIS ATOM	Fisica atomica	Atomic physics
FIS NUCL	Fisica nucleare	Nuclear physics
FIS PLASMA	Fisica del plasma	Plasma physics

FIS SOL	Fisica dei solidi	Physics of solids
FIS SUBNUCL	Fisica subnucleare	Subnuclear physics
FISC	Fisco	Tax
FISIOL	Fisiologia	Physiology
GASTR	Gastrologia	Gastrology
GEMMOL	Gemmologia	Gemmology
GEN	Generazione dati	Data generation
GENET	Genetica	Genetics
GEOCHIM	Geochimica	Geochemistry
GEOD	Geodinamica	Geodynamics
GEOFIS	Geofisica	Geophysics
GEOGR	Geografia	Geography
GEOL	Geologia	Geology
GIOCO	Gioco	Game
GRAF	Grafica	Graphics
IDROL	Idrologia	Hydrology
INF	Informatica	Informatics
ING	Ingegneria	Engineering
ING ACUS	Ingegneria acustica	Acoustic engineering
ING AER	Ingegneria aeronautica e aerospaziale	Aeronautical and aerospace engineering
ING CHIM	Ingegneria chimica	Chemical engineering
ING CIV	Ingegneria civile	Civil engineering
ING MECC	Ingegneria meccanica	Mechanical engineering
ING MIN	Ingegneria mineraria	Mining engineering
ING NAV	Ingegneria navale	Naval engineering
ING NUCL	Ingegneria nucleare	Nuclear engineering
ING PETROL	Ingengeria petrolifera	Petroleum engineering
ING SIS	Ingegneria dei sistemi	Systems engineering
INT ART	Intelligenza artificiale	Artificial intelligence
INTERAZ	Interazione	Interaction
ISTOL	Istologia	Histology
LETTER	Letteratura	Literature
LING	Linguistica	Linguistics
MAR	Dizionario marittimo	Maritime dictionary
MAT	Matematica	Mathematics
MATER	Materia	Matter
MECC	Meccanica	Mechanics
MECC FL	Meccanica dei fluidi	Fluid mechanics
MECC QUANT	Meccanica quantistica	Quantum mechanics
MECC STAT	Meccanica statica	Static mechanics
MED	Medicina	Medicine
MEMOR	Dispositivi di memorie	Memory devices
MEST	Nomi di mestiere	Job names
METALL	Metallurgia	Metallurgy
METEOR	Meteorologia	Meteorology
MICOL	Micologia	Mycology
MICROBIOL	Microbiologia	Microbiology
MIL	Militare	Military
MINERAL	Mineralogia	Mineralogy
MUS	Musica	Music
NAVIG	Navigazioni	Navigations
NOTAZ	Notazioni	Notations
NUMER	Numerazioni	Numberations
OCEANOGR	Oceanografia	Oceanography
ORG	Organizzazione	Organization
ORG AZ	Organizzazione aziendale	Business organization
ORG DATI	Organizzazione dati	Data organization
ORG IND	Organizzazione industriale	Industrial organization

OTTICA	Ottica	Optics
PALEOBOT	Paleobotanica	Palaeobotany
PALEONT	Paleontologia	Paleontology
PATOL	Patologia	Pathology
PATOL VEG	Patologia vegetale	Plant pathology
PELL	Pellame	Leather
PERIF	Periferiche	Peripherals
PETROGR	Petrografia	Petrography
PITT	Pittura	Painting
POL	Politica	Politics
PROG MECC	Programmazione meccanica	Mechanical programming
PROGR	Programmazione	Programming
PSIC	Psicologia, psicanalisi e psichiatria	Psychology, psychoanalysis and psychiatry
PT	Poste e telecomunicazioni	Post and telecommunications
RAP ART	Rappresentazioni artistiche	Artistic performances
RELAT	Relatività	Relativity
RELIG	Religioni	Religions
RETI	Reti	Networks
SART	Sartoria	Tailoring
SCI TEC	Scienza e tecnica	Science and technique
SCOL	Scuola	School
SCULT	Sculture	Sculptures
SICUR	Sicurezza	Safety
SILVIC	Silvicoltura	Forestry
SIS CONTR	Sistemi di controllo	Control systems
SIST	Sistemi	Systems
SIST ELAB	Sistemi di elaborazione dati	Data processing systems
SPETTR	Spettrografia	Spectrography
SPORT	Sport	Sport
ST	Storia	History
STAT	Statistica	Statistics
SVIL	Sviluppo	Development
SVIL SIST	Sviluppo sistemi	Systems development
TASSON	Tassonomia	Taxonomy
TEAT	Teatro	Theater
TECN	Tecniche	Techniques
TECN ELAB	Tecniche di elaborazione dati	Data processing techniques
TECNOL	Tecnologia	Technology
TELECOM	Telecomunicazioni	Telecommunications
TEOR	Teorie	Theories
TERMOD	Termodinamica	Thermodynamics
TESS	Termini tessili	Textile terms
TRATT	Trattamento	Treatment
TRATT DATI	Trattamento dati	Data processing
TRATT TESTI	Trattamento testi	Text processing
TUR	Turismo	Tourism
VETER	Veterinaria	Veterinary
VIROL	Virologia	Virology
ZOOL	Zoologia	Zoology
ZOOL INVERT	Zoologia degli invertebrati	Invertebrate zoology
ZOOL VERT	Zoologia dei vertebrati	Vertebrate zoology

The terminological domains with more entries are that of Medicine (MED label, with approximately 63,000 inflected forms), Economics (ECON label, with approximately 58,000 inflected forms), Information Technology (INF label, with approximately 38,000 inflected entries), Law (DIR label, with approximately 14,000 inflected forms), and Engineering (ING label, with approximately 5,000 inflected forms).

3.1. The use of NooJ DELACF in automatic textual analysis

As known, in NooJ we can use tagged entries to retrieve information automatically from texts and represent by means of charts the knowledge they include. To demonstrate this, we have analyzed an Italian essay on computer science and the Internet[12] applying the Italian NooJ DELACF, of which we give a quick example in the figure below:

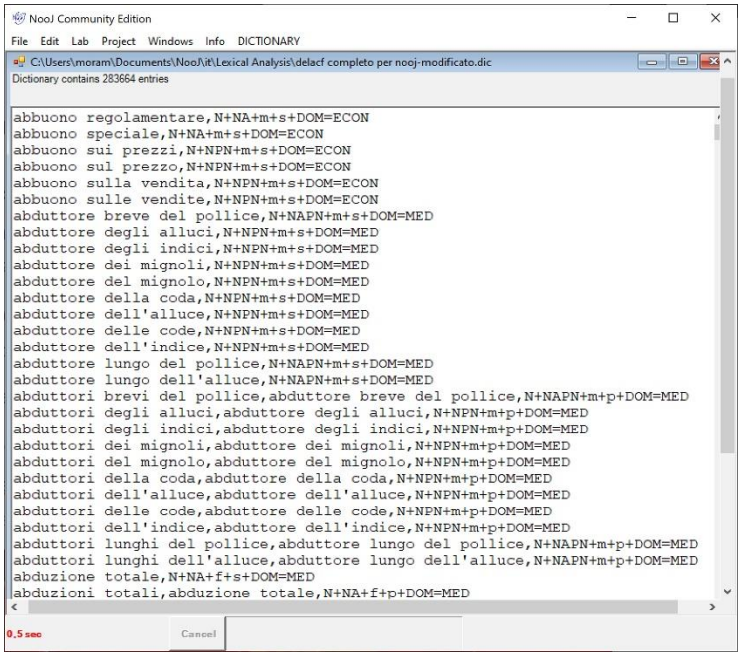
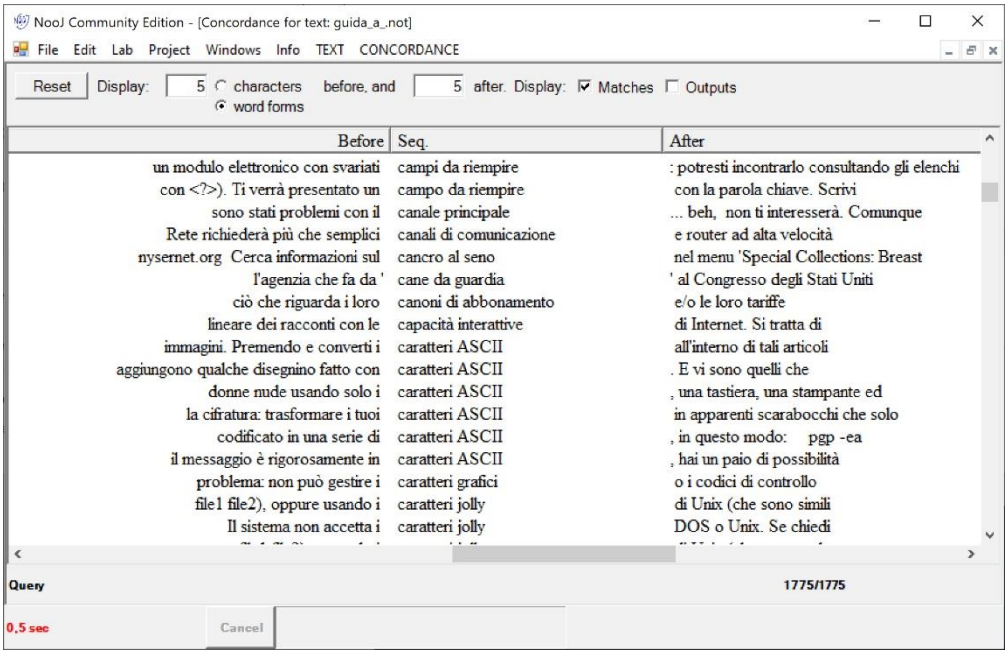


Fig. 1. An excerpt of Italian NooJ DELACF

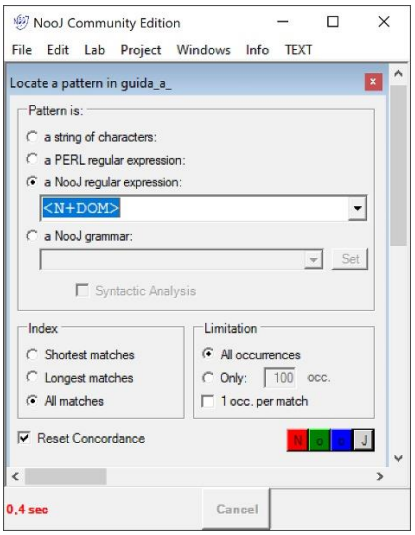
Once achieved the linguistic analysis with NooJ, and by accessing the Locate Pattern panel from a given text, it will be possible to extract all the terminological compound words, which as for our text are 1775:

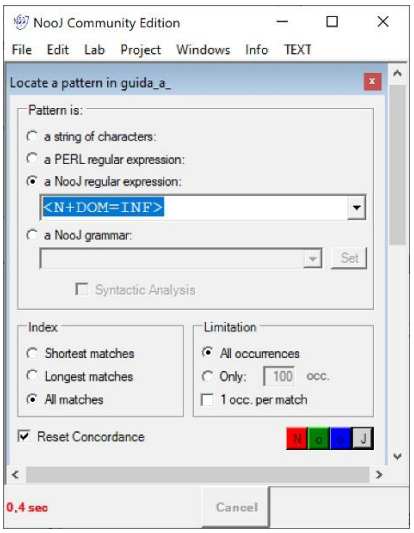


12 See https://www.liberliber.it/online/opere/download/?op=2345413&type=opera_url_txt.

Fig. 2. NooJ Concordance.

It will also be possible to search and extract both all compound words and those belonging to a specific terminological domain, such as INF (Information Science), using the regular expressions shown in the following figures:





Figs. 3 and 4. NooJ Locate Pattern/regular expression for compound words extraction.

In this case, our text contains 890 INF compound words. The results of this extraction can be displayed in the form of a chart, using the option Statistical Analysis -> Standard Score. For each part of a given text, the chart displays the points and relative values of greatest and least occurrence of the compound words set searched for:

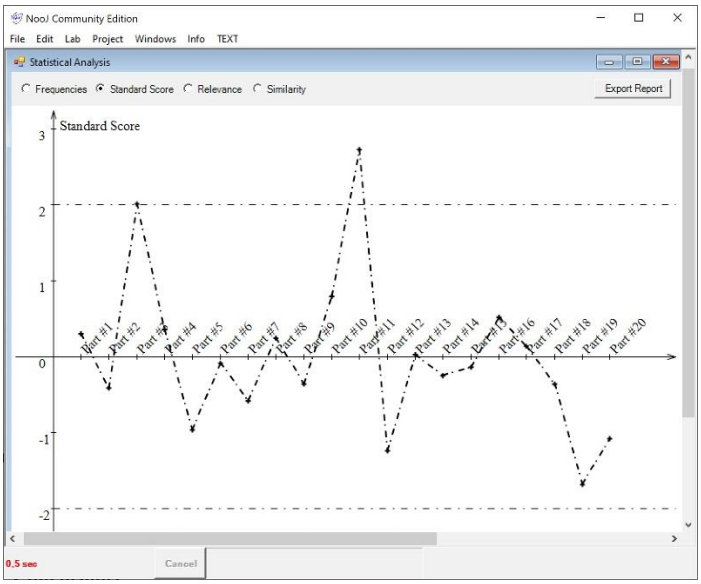


Fig. 5. A sample of Standard Score chart.

4. Conclusions

As is known, NooJ allows the creation and analysis of even very large corpora, which can be parsed as previously shown, both separately and comparatively. In the case of terminological texts, this allows not only to achieve a lexical and semantic analysis, but also to represent the knowledge contained in such texts and corpora. This is due to the already mentioned characteristic

of terminological compound words, i.e., their necessity of being as less ambiguous as possible. Therefore, the univocal and unambiguous relationship they establish between signifier and signified, allows a high level of conceptualization of the texts in which terminological compound words occur. This makes it possible to understand the terminological content of a text by simply extracting the terminological entries from it, and verifying which sector of knowledge is most represented.

References

Gross G. et alii (1986), Typologie des noms composés, Rapport A.T.P. du C.N.R.S., Paris.

Gross G., Jung, R., Mathieu-Colas, M. (1987). Noms composés, rapport n° 5 du Programme de recherches coordonnées « informatique linguistique », CNRS, Paris.

Gross, M. (1975). Méthodes en syntaxe, Paris, Cantilène.

Gross, M. (1990). Grammaire transformationnelle du français : syntaxe de l'adverbe. Cantilène, Paris.

Harris, Z. S. (1946). "From Morpheme to Utterance", Language 22, pp. 161-83.

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

Mathieu-Colas, M. (1987). Variations graphiques de mots composés, rapport n° 4 du Programme de recherches coordonnées « informatique linguistique », CNRS, Paris.

Monteleone M., Elia A., De Bueriis G., Di Maio F. (2005). "Le poliremathiche 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.

Piot, M. (1988). Conjonctions de subordination et fixement, Langages n 90, Les expressions figées, Larousse, Paris.

Silberztein, M. (1989). Dictionnaires électroniques et reconnaissance lexicale automatique, thèse de doctorat, LADL, Université Paris VII

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.

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.