Constructive Linguistics for Computational Phraseology: the Esperanto Case

Federico Gobbo[1,2][0000−0003−1748−4921]

[1] University of Amsterdam, the Netherlands
[2] University of Turin, Italy
F.Gobbo@uva.nl
http://uva.nl/profile/f.gobbo/

Abstract. This paper presents the application of the constructive adpositional grammars (CxAdGrams) to phraseological units, through the special case study of Esperanto. Constructive linguistics is an approach to human language analysis that considers constructions, themselves being paradigms of language-in-use, as the first units. Unlike other constructional approaches, constructive linguists apply formalisms in understanding linguistic phenomena. The adpositional paradigm is the most developed formalism in constructive linguistics, which is understandable by humans and machine-readable at the same time. The term 'constructive' should also be understood in formal terms, as the adpositional paradigm is based on constructive mathematics, and in particular on topos-theory. From a theoretical perspective, CxAdGrams describe human languages in terms of constructions, described adpositional trees (in short, adtrees). This paper aims to explain why such an interpretation of constructions in terms of adtrees can be useful for a deeper understanding of phraseology. Esperanto is the case study chosen so to give an empirical base to CxAdGrams. In particular, we illustrate the problematisation of its phraseology as well as the advantages of Esperanto in setting up workable prototypes in a short time.

Keywords: Constructive Linguistics · Computational Phraseology · Adpositional Grammars · Adpositional Argumentation · Esperanto.

1 Introduction

Understanding how language is structured is one of the most fascinating and challenging endevour that human beings have ever done. Many approaches are possible, and many approaches were proposed throughout the flow of human history. Because of the computational turn, in the 21^{st} century our conceptualisation of language is changed and is still changing, and, for this reason, linguistics should propose robust theories that treat language in terms of information, to be understood by humans and read by machines at the same time.

Initially proposed in [14], constructive linguistics is a relatively new approach to human language that follows such informational tenet. It is important to note that, in this perspective, the word 'constructive' has both a mathematical and

EUROPHRAS 2019, pages 78–85,
Malaga, Spain, September 25-27, 2019.
https://doi.org/10.26615/978-2-9701095-6-3_010

a linguistic specific meaning at the same time. In short, on the one hand, constructive mathematics is a way to develop mathematics that strictly preserves the information content of any statement [2]. On the other hand, cognitive sciences show that humans are able to communicate as they can read intentions, i.e., infer what the listener is expecting from the speaker, and find patterns, i.e., they can categorise sensibilia mapping them into the mind; they learn intention-reading and pattern-finding by imitation of other humans [21]. In fact, humans use language to gain somebody's attention or to share their mental state. In order to do so, a human language can be described in terms of a map of social conventions of a specific speech community. This individual and collective process of categorisation leads to the emergence of linguistic *constructions*, which are patterns of form-meaning correspondence based on language usage. For this reason, human languages can be described as collections of constructions. We consider constructions as the hypernym of phraseological units and other linguistic phenomena. In general, (oral) discourses and (written) texts are split into units – such as sentences and phrases – which are instantiations of linguistic constructions; phraseological units are a specific type of such units. For the purposes of this paper, we will delve into phraseological units only.

Phraseological units are at the crossroad of grammaticalisation and lexicalisation, which are two complementary processes that can be found in any living human language [3]. While grammaticalisation is a syntactotelic process, lexicalisation is a synthetic process. In other words, grammaticalisation goes from the lexicon to the syntax, affecting lexical items both in their phonological material and in their meaning (which tends to be lost). Conversely, the process of lexicalisation makes constructions lose their flexibility and compositionality, and eventually, they acquire idiosyncratic content. The most extreme result of lexicalisation is the formation of idiomatic expressions, which are not analysable anymore, but should be taken as fixed. Therefore, under the perspective of constructive linguistics, phraseological units are in the middle of the continuum of constructions, where at one extreme we find idioms while at the other one we find word-playing, portmanteaus, dynamic metaphors, and, in general, creative language usage.

Let us show an example of a phraseological unit found in the middle of the continuum of constructions. The following quotation is from the political pamphlet *Gli Stati Uniti d'Europa* [United States of Europe] [16], written by one of the founding fathers of the European Union, the Italian antifascist intellectual Ernesto Rossi (author's English translation immediately below):

> Clemenceau diceva che la guerra è una cosa troppo seria per essere lasciata ai generali. Noi dobbiamo dire che la pace è una cosa troppo seria per essere lasciata ai diplomatici. [16, p. 96]

> [Clemenceau used to say that war is too serious a matter to be left to the generals. We should say that peace is too a serious matter to be left to the diplomats.]

The quotation shows the same construction – in this particular case, a phrase schema – in two different instantiations. The fixed part is *... say that ... is too serious matter to be left to ...* while the analysable parts for the respective sentences are the triples {Clemencau, war, generals} and {we, peace, diplomats}. In the next section, we illustrate how such a phraseological unit can be expressed in terms of adpositional trees.

2 Adpositional trees for phraseological units

Unlike purely constructionist approaches to language such as Radical Construction Grammar [4], constructive linguists do not avoid formalisms, instead they embrace them in constructive mathematical terms. The most developed paradigm in constructive linguistics is based on the concept of *adposition*. In this context, the term has to be intended in two different ways at the same time.

The first way to intend adpositions is linguistic. However, it should be underlined that, here, an adposition is not only a mere hypernym of prepositions, postpositions, and the like, but also and mainly a generalisation of functional words that connect lexemes and other semantically loaded material. The second way to intend adpositions is mathematical. Adpositions represent purely structural information, and they are placed as hooks under the upmost root that sustain the trees that represent constructions.

So far, the adpositional paradigm has been applied in various branches of human languages, generating constructive adpositional grammars (CxAdGrams), which are abstract and general and language-dependent at the same time. In the field of morphology and syntax, CxAdGrams were applied to purely constructional analysis, adapting the Tesnerian notion of valency, actant and grammar character to the key notion of adposition.[3] In the field of semantics and pragmatics, CxAdGrams were applied to discourse analysis of therapeutic conversations, through the representation of Searle's speech act theory of social world construction [17] in terms of pragmatic adtrees [14]. Up to the author's knowledge, there is still no application of CxAdGrams in the field of phraseology.

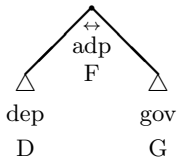

Fig. 1. The generic abstract adpositional tree in its standard form

[3] Strictly speaking, the adpositional paradigm does not generate dependency grammars, although there is a relation of ancestry between Tesnière's Structural Syntax and CxAdGrams [15].

Figure 1 shows the minimal, abstract adtree, in its standard form. Adpositions (adp) represent the relation between two linguistic elements. Linguistic relations are asymmetrical, and they are understood in terms of dependency (dep, conventionally on the left) on a governor (gov, on the right). Each element is tagged in terms of grammar characters: D and G respectively for dependants and governors, while F stands for 'final', as it is the result of their structural relation. Adtrees are recursive; the triangles $\triangle$ on the leaves are a convenient way to represent subtrees without indulging in details. Finally, adpositions convey information prominence between dependants and governors – in the form of the arrow $\leftrightarrow$ on top of the hook.[4] It is worth noting, that every adtree can be flattened, for instance for the purpose of coding, through trivial finite-state automata that do the linearisation. Figure 2 (immediately below) shows the generic abstract adtree in its linearised form.

$$\mathrm{adp}^{\leftrightarrow}_{\mathrm{F}}((\mathrm{dep})_{\mathrm{D}}, (\mathrm{gov})_{\mathrm{G}})$$

Fig. 2. The generic abstract adpositional tree in its linearised form

Let us see the phrase schema of Ernesto Rossi's example, previously stated, in terms of adtrees. For sparing space, Figure 3 represents only the instantiation by the triple {Clemencau, war, generals}.[5] Epsilons (ϵ) represent syntactic relations, i.e., where no morpheme is found. The right arrow $\rightarrow$ above *Clemencau* indicates that the information prominence is above the dependent instead of the governor in that particular subtree. The usefulness of triangles which hide non-relevant information for the analyst – but always retrievable, thanks to the constructive mathematical foundation – is immediate to the reader. For instance, in Figure 3 the morphological information of the word *generals* as well as the linguistic details of the verbal forms *is too serious matter to be left to* and *used to say* are of no interest as the purpose of this adtree is to put in evidence the phrase schema underlying this phraseological unit, i.e., in linguistic terms, what is grammaticalized, and hence fixed – the skeleton of the phrase schema – and what is conveying the lexical information, that is the triple {Clemencau, war, generals}.

The abstract grammar characters {D, G, F} shown in Figures 1-2 are instantiated as verbants, nominals, adjuncts, circumstantials, respectively {I, O, A,

[4] For more details on the constructive mathematical aspects of adpositional grammar, readers are invited to check Appendix B of the book presenting the mathematical foundation of CxAdGrams in terms of Grothendiek's toposes [14].

[5] It is worth noting that punctuation is included in CxAdGrams, being themselves adpositions between sentences. In such a way, potentially, a whole large text – like Dante's *Divina Commedia* – can be represented as a single, enormous adtree.

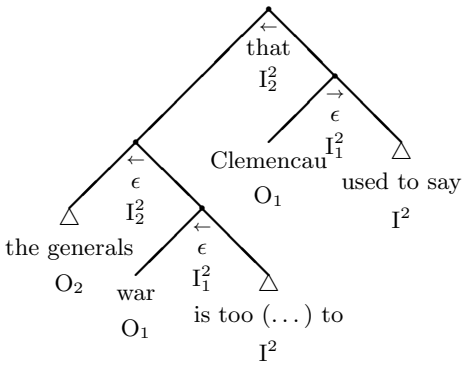

Fig. 3. The standard adtree of the example *Clemencau...*

E}.[6] In principle, any consistent part-of-speech tagging convention can be used with adtrees; it suffices to put the tags on the bottom of the leaves (such as O_1 under *Clemencau* and *war* in Figure 3) and of the hook (such as I_2^2 under *that*).[7]

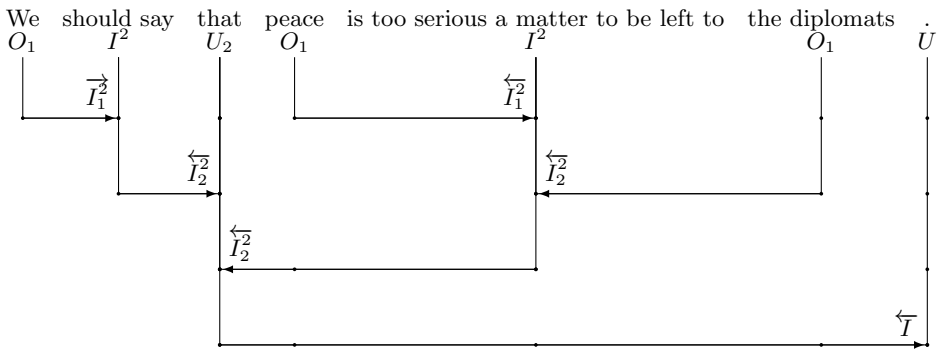

Fig. 4. The path-like adtree of the example *We should say...*

We present here a third way to represent adtrees, which we propose here to call "path-like" adtrees, in the absence of a better naming. This representation has the important advantage of respecting the linguistic word order, and therefore it can be useful for educational purposes, as shown in the Linguistic Atelier in Montessori primary schools in Milan, Italy [13]. Structurally speaking, this representation preserves the information, and thus it respects the fundamen-

[6] Such labelling of grammar characters is borrowed from the original Tesnière's Structural Syntax [19]. Unfortunately, the letters, which Tesnière took from Esperanto, were not kept in the English translation [20].

[7] Readers interested in delving into this particular convention can refer to [15, 14].

tal tenet of constructive linguistics, i.e., the foundation on constructive mathematics. In the example below, the fifth grammar character is needed, in order to indicate underspecified or unifying elements (hence, the vowel U), typically grammaticalised morphs or punctuation elements. Figure 4 shows the phrase schema presented before instantiated with the triple {we, peace, diplomats}. Of course, path-like adtrees do not show syntactic relations in terms of epsilons (ϵ) under the hook because they are driven by concrete linguistic material, i.e., by morphs.

3 Phraseology and Esperanto

Esperanto is the most interesting product of Interlinguistics, the branch of linguistics dealing with planned languages, i.e., languages that are written in their fundamental structural traits before even to be spoken [10]. Unlike all other planned languages proposed in the last two centuries, Esperanto succeeded in forming a stable community of language users, with a relevant critical mass, and so it shows emerging sociolinguistic traits that are a challenge for theoretical linguistics.[8] According to the corpus-based grammar of Esperanto by Gledhill [7], its high morphological regularity, especially in derivation, permits to drastically reduce the learning efforts, both for humans and for machines, even if this does not imply that in absolute terms Esperanto is simpler than other human languages [11]. While presenting its phraseology, Gledhill [7] notes that "many Esperantists are uncomfortable with the idea of variation and near-synonymy in the vocabulary of the language, but as Janton (1994) has pointed out[,] multiple vocabularies are an integral part of Esperanto's system of register and style." This may be a reason why in Esperanto studies phraseology is relatively an understudied aspect.

In order to reinforce its language project, Ludwik Lejzer Zamenhof in 1910 published *Proverbaro Esperanta*, a collection of Esperanto proverbs extracted from a comparative analysis of four major European languages (French, German, Polish, Russian) made by his father Mordechai Mark. That book can be considered the base of Esperanto phraseology. Because of the language ideology of ethnic neutrality surrounding Esperanto – which is rather complex [12] – some translations were not straightforward. Let us show one tricky example. Entry number 7 in the *Proverbaro* corresponds to the English phraseological unit 'it's Greek to me', which is construed around the idea of 'language of Otherness'. In particular, it contains four different proposals for expressing such phraseological unit in Esperanto.

- 7a (539 too [sic]). Ĝi estas por mi ĥina scienco.
- 7b. Ĝi estas por mi volapukaĵo.
- 7c. Nun finiĝas mia klereco.
- 7d. Venis fino al mia latino.

[8] For a discussion on the possible definition of such peculiar community of language practice, see at least [18].

If we take ethnic neutrality as the standpoint, the problem becomes obvious: you cannot blame Greeks for their "strange" language (following English), and analogously you don't blame Chinese (following the Spanish *me suena a chino*) or Arabic (following the Italian *per me è arabo*), because Esperanto speakers can be English, Greeks, Chinese, Arabs, Spanish and Italians alike. For this reason, proposal 7a, which refers to Chinese (i.e., *ĥina*) was discarded in practice. Proposal 7c literally means "now it arrived to the end of my knowledge", lowering too much the pragmatic force found in the phrase schema, because it does not involve a language of Otherness, and therefore it did not work either. Proposal 7d mentions Latin, but Latin cannot always play the role of Otherness: for example, the Dutch expression *Ik ben aan het eind van mijn latijn*, which is very similar to proposal 7c, both meaning literally "there is an end to my Latin", more or less, in Dutch is used to convey the information 'I have no energy anymore', which is completely different from a pragmatic point of view. For this reason, it survives only in the most prestigious register of Esperanto intellectuals. Proposal 7b actually won, as the blamed language of Otherness is Volapük, a language project planned before Esperanto which gained some success in the early days of Esperanto, but it did not work so well. Eventually Volapük entered the Esperanto culture as the language of Otherness – on Volapük, see at least [6].

In Esperanto, phraseological units are the result of the negotiation of meaning between speakers immersed most of their lives in other language environments (there are no Esperanto monolinguals). Zamenhof's proposal 7c shows that endogenous phraseological solutions are possible. In other terms, there are phaseological units referring specifically to the history, habits, ways of life of Esperanto speakers – as shown by Fiedler in her fundamental work [5]. On the other hand, many of the phraseological expressions found in colloquial Esperanto language use come from europeanisms, i.e., units that are commonly represented in most European languages. In his study on metaphors in Esperanto, Astori [1] shows the proposal by Hungarian Esperanto speakers proposed to introduce *dormi kiel lakto*, literally 'to sleep like milk' for the europeanism 'to sleep like a baby' (i.e., profoundly), did not work, being to specifically linked to the Hungarian *Weltanschauung*, Conversely, the europeanism, *dormi kiel ŝtono*, literally, 'like a stone' is of common use as an alternative expression in the colloquial Esperanto register.

4 A final remark

This position paper shows that CxAdGrams are apt to represent phraseological units, and that a first testbed for a consistent linguistic analysis could be done through Esperanto. The recent proposal of Adpositional Argumentation could analyse Ernesto Rossi's example as an argument from comparison framed into the Period Table of Arguments, with a considered added-value to the annotated corpus to be done [8, 9, 22].

References

1. Astori, D.: Metafore nell'esperanto. In: Astori, D. (Ed.) La metafora e la sua traduzione, pp. 133–148. Bottega del libro, Parma (2016)
2. Bridges, D., Richman, F.: Varieties of Constructive Mathematics. Cambridge University Press, Cambridge (1987)
3. Cabrera Moreno, J.C.: On the Relationships Between Grammaticalization and Lexicalization. In: Giacalone Ramat, A. and Hopper, P.J. (Eds.), The limits of grammaticalization. pp. 211–229. John Benjamins, Amsterdam (1998)
4. Croft, W.: Radical Construction Grammar. Syntactic Theory in Typological Perspective. Oxford University Press, Oxford (2001)
5. Fiedler, S.: Plansprache und Phraseologie Empirische: Untersuchungen zu reproduziertem Sprachmaterial im Esperanto. Peter Lang, Bern (1999)
6. Garvía, R.: Esperanto and its rivals: the struggle for an international language. Penn Press, Chicago (2015)
7. Gledhill, C.: The Grammar of Esperanto. A Corpus-based description. Lincom Europa, München, 2 edn. (2000)
8. Gobbo, F., Wagemans, J.H.M.: Building argumentative adpositional trees: Towards a high precision method for reconstructing arguments in natural language. In: Proceedings of the Ninth Conference of the International Society for the Study of Argumentation. pp. 408–420 (2019)
9. Gobbo, F., Wagemans, J.H.M.: A method for reconstructing first-order arguments in natural language. In: Proceedings of the 2nd Workshop on Advances in Argumentation in Artificial Intelligence (AI^3 2018). pp. 27–23 (2019)
10. Gobbo, F.: Interlinguistics, a discipline for multilingualism. Amsterdam University Press, Amsterdam (2015)
11. Gobbo, F.: Are planned languages less complex than natural languages? Language Sciences **60**, 36–52 (2017)
12. Gobbo, F.: Beyond the nation-state? the ideology of the esperanto movement between neutralism and multilingualism. Social inclusion **5**(4), 38–47 (2017)
13. Gobbo, F.: Language Games Children Play: Language Invention in a Montessori Primary School, pp. 1–14. Springer International Publishing, Cham (2019)
14. Gobbo, F., Benini, M.: Constructive Adpositional Grammars. Foundations of Constructive Linguistics. Cambridge Scholars Publishing, Newcastle upon Tyne (2011)
15. Gobbo, F., Benini, M.: Dependency and valency. from structural syntax to constructive adpositional grammars. In: In K. Gerdes, E. Hajiov and L. Wanner (Eds.), Computational Dependency Theory. pp. 113–135. IOS Press, Amsterdam (2013)
16. Rossi, E.: L'Europa di domani: Un progetto per gli Stati Uniti d'Europa. A cura di Mauro Rubino. Stilo editrice, Bari (2014),
17. Searle, J.R.: Making the Social World: The Structure of Human Civilization. Oxford University Press, Oxford (2010)
18. Stria, I.: Esperanto Speakers - an Unclassifiable Community? Wydawnictwo KUL (2015), instytut Pedagogiki na Katolickim Uniwersytecie Lubelskim Jana Pawła II w Lublinie. Księga Jubileuszowa
19. Tesnière, L.: Éleménts de syntaxe structurale. Klincksieck, Paris (1959)
20. Tesnière, L.: Elements of Structural Syntax. John Benjamins, Amsterdam (2015)
21. Tomasello, M.: Constructing a language. A Usage-Based Theory of Language Acquisition. Harvard University Press, Harvard (2003)
22. Wagemans, J.H.M.: Constructing a Periodic Table of Arguments. In: Argumentation, Objectivity, and Bias: Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation (OSSA) (2016)