The Hierarchical Organization of Syntax

Babak Ravandi (✉ b.ravandi@northeastern.edu)

Northeastern University

Valentina Concu

Universidad del Norte

Research Article

Keywords:

Posted Date: December 6th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1141072/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

The Hierarchical Organization of Syntax

Babak Ravandi^{1,2,†,*}

Valentina Concu^{3,*,†}

¹Network Science Institute, Northeastern University, Boston, USA

²Department of Physics, Northeastern University, Boston, USA

³Department of Foreign Languages, Universidad del Norte, Barranquilla, Colombia

Hierarchies are the backbones of complex systems and their analysis allows for a deeper understanding of their structure and how they evolve. We consider languages to be also complex adaptive systems. Hence, we analyzed the hierarchical organization of historical syntactic networks from German that were created from a corpus of texts from the 11th to 17th centuries. We tracked the emergence of syntactic structures in these networks and mapped them to specific communicative needs. We named these emerging structures communicative hierarchies. We hypothesise that the communicative needs of speakers are the organizational force of syntax. We propose that the emergence of these multiple communicative hierarchies is what shapes syntax, and that these hierarchies are the prerequisite to the Zipf's law. The emergence of communicative hierarchies indicates that the objective of language evolution is not only to increase the efficiency of transferring information. Language is also evolving to increase our capacity to communicate more sophisticated abstractions as we advance as a species.

*Authors equally contributed.

†Corresponding authors e-mail: bk.ravandi@gmail.com and vconcu@uninorte.edu.co

Introduction

Among all the known means of communications, the human language is the only one that developed syntax [1–3], which regulates how words combine into phrases and how phrases combine into larger phrases [4]. This regulatory strength enables speakers to make sense of and organize the underlying combinatory features of word association. It also allows them to construct a rich repertoire of sentences from a set of words within the boundaries imposed by semantics and pragmatics [5], “making infinite use of finite means” [6].

The current shape of syntax in any given language is the result of evolutionary processes that have been triggered by the communicative needs that speakers needed to satisfy over centuries [7]. The necessity to communicate time references, for instance, has prompted many languages to develop an effective system of verbal tenses and temporal adverbs [8]. Such a system makes possible for speakers to specify the time frame in which actions take place. In English, for instance, such a system counts today on different morphologically marked forms and syntactic constructions (“I went” vs “I have gone”) that speakers use based on the meaning they want to convey. Figure 1 shows syntactic changes around three common communicative needs at different stages in the history of German [9].

Current theories on why syntactic communication was preferred by evolution are based on analyzing the patterns of growth in the number of actions/signals and objects in language [10, 11]. Nowak et al. suggest that the crucial step that guided the transition from a non-

Figure 1

Figure 1: **The relationship between communicative needs and syntactic changes.** Communication has been often recognized to be a driving factor for language change in the syntactic inventory of a language. Hence, these changes always gravitate around specific communicative needs, they can be of distinctive types and affect one or more linguistics elements. **(A)** The verbs in Old High German were morphologically marked for the subject and no personal pronoun was needed to indicate who or what was doing a particular action. The loss of such morphological feature in Middle High German on one side, and the need to indicate the subject on the other, gives rise to the obligatory personal pronoun to accompany the main verb. **(B)** The emergence of the present perfect from Middle to Early New High German makes possible for speakers to use the passive also in the newly grammaticalized tense. In this way, the communicative goal to emphasize the semantic role of the “patient” can be expressed in the past using two different tenses. **(C)** The rise of the constructions with the verb *werden* and the infinitive verbs in Middle High German takes over a similar construction with the modal verb *sollen* to satisfy the communicative need to express future reference in Early New High German.

syntactic to a syntactic communication was an increase in the number of “relevant events” to which speakers needed to refer [10].

Solé [12] argues that the prerequisite for syntax is one of the fundamental principals of organizations that is common to all languages, the Zipf’s law [13]. This law indicates that in a ranked-ordered list of words’ frequencies, the frequency of words decays inversely with their rank. This means that, in any language, only a few words will be used quite often, while a large number of them will appear only once or twice in the same text. Syntax also shows the same organizational patten: only a few words posses the majorities of syntactic links, while a high number of them have only a few. According to Solé [12], is the Zipf’s law then what provides the connectedness between words “for free” and is, therefore, an essential precondition for the language transition into a syntactic means of communication

We propose a more detailed perspective from historical data to account for both the current structure and origin of syntax. As syntax connects words through asymmetrical relationships, the syntactic structures that have evolved throughout the centuries form multiple hierarchies that represent the preferred way in which communicative needs are satisfied. After tracking the emergence of these hierarchies in syntactic networks during a time span of eight centuries, we argue that the rise and evolution of these “communicative hierarchies” is what shapes syntax. Our analysis also suggests that the existence of these hierarchies is the prerequisite to Zipf’s law. Furthermore, we argue that the interplay between the hierarchical organization of syntactic networks and communicative needs could shed a new light on the evolution of language and the origin of syntax.

Results

We created syntax networks from a corpus of Middle (1050 – 1350) and Early New High German (1350 – 1650) texts for a total of 455,944 words distributed in 8 different centuries (Section SI1). This corpus has been used already to track the historical evolution of the periphrases with *werden* plus present participles and *werden* plus infinitive verbs [14]. All the eight networks (one for each century) were created following the principle of dependency grammar (Section SI2) and represent the sum of the sentence graphs found in the corpus. Hence, we refer to these networks as Aggregated Syntactic Networks, henceforth ASNs. Figure 2 demonstrates the relationship between dependency syntactic trees and an ASN as well as our choice to separate words based on the syntactic role of words in a sentence. In every ASN, all the nodes, which correspond to the lemma forms of the words, are labeled according to their grammatical function (Table SI1). Such a labeling allows to individuate the connective patterns of each node and how these patterns change over time. It also allows for analyzing which grammatical functions are the most influential or which one are involved in the rise of new structures and their related hierarchies.

Figure 2

Figure 2: From Syntactic Dependency Trees to Aggregated Syntactic Networks. A syntactic dependency tree represents the hierarchical structure of the constituents in any given sentence. However, to investigate the entire range of syntactic relationships in Middle and Early New High German, we aggregated the syntactic trees of all the sentences in which the target verb *werden* was found. This approach enables to shift focus from single sentences to a whole corpora during a particular century. Hence, an ASN is the representation of the sum of the syntactic relationships established between the sentences’ constituents in a given century. **(A-B)** Two syntactic dependency trees from the Middle High German corpus (11th century). The labels of the nodes (presented here in distinctive colors) reflect the grammatical roles, whereas the edges refer to the syntactic relationships between the words. **(C)** The resulting network by aggregating sentence A and B is no longer a tree because now two distinct paths exist from the root “AX werden” to the dependent “PP er”, as highlighted with yellow and orange colors. Therefore, the aggregated network can not be treated as a dependency syntactic tree anymore.

Hierarchies and Communicative Needs

A tree represents the most hierarchical graph structure, and the syntactic relationships that connect words in each sentence can be represented with a tree. As we merged the syntactic trees of multiple sentences to build ASNs, a central question we wanted to answer was: what happened to those single-sentence hierarchies? The study of trophic levels of organisms in ecological systems and their corresponding food chains has laid the foundations for analyzing hierarchical structures in graphs [15]. Thereafter, the generalization of trophic levels to complex networks paved the way to assess hierarchical characteristics of complex systems [16]. We show that the distribution of hierarchical levels in ASNs captures syntactic changes (Section SI3). By conducting a historical comparison of the hierarchical levels, we observe the rise of new hierarchies representing distinct communicative needs. Hence, we named these communicative hierarchies.

Through repetition and semantic weakening, words expand or shrink their associative capacity and their meaning’ repertoire [17]. Many instances of this type of syntactic change were captured in our ASNs (Section SI3). One of this changes is the rise of the modal verb *können* (*can* in English) in Early New High German as it replaces *mögen* (*to like* in English) to express mental and physical abilities. *Mögen* was the most frequent modal verb in Old High German and one of the most frequent ones in Middle High German [18]. Until the last centuries of Middle High German (14th century), *mögen* was used to express abilities as well as other communicative needs such as expressing preferences and obligations [18]. *Können*, which in Old and Middle High German was not yet an integral part of the group of modal verbs, was used instead to express a small range of intellectual capacities only. However, in the last two centuries of the Early New High German period, *können* becomes the preferred way to express physical and mental abilities [18].

The transition of the communicative need of expressing physical and mental abilities from *mögen* to *können* was captured by the hierarchical analysis in our ASNs, as showed in Figure 3. The vertical positions of nodes in Figure 3 represents the forward hierarchical levels of the words in our corpus from the 14th to the 17th centuries. *Mögen*, marked with green star, is constantly located at the top of hierarchies, together with other modal and auxiliary verbs. At the beginning of the Early New High German period, *können*, marked with the white star, starts to become part of the of the group of modal verbs, when the semantic restrictions that bounded this verb to limited communicative contexts began to be slowly lifted [18]. Consequently, *können* gradually started to replace *mögen* in more and more instances, until, in the 16th century, it becomes the preferred way to express abilities, both intellectual and physical [18]. The use of ASNs allowed us to capture the exact period in which the communicative hierarchy with *können* emerged, contributing to the strengthening of the multi-hierarchical organization of syntax. We believe that processes similar to this one is what shaped syntax in the first place as language was changing from a non-syntactical to a syntactic means of communication. We also believe that these hierarchies are what favored Zipf’s law in human languages.

Figure 3

Figure 3: **The Rise of *können* as a Modal Verb to Express Physical and Mental Abilities.** The vertical positions of nodes represent the forward hierarchical levels in their respective ASNs (the 3D layout is generated using the algorithm in [19]). The size of nodes indicate the number of syntactical connections a word has. *Mögen* is marked with the green star and *können* with the white star. **(a-b)** Through late Middle High German *mögen* is used to express abilities as well as other communicative needs such as expressing preferences and obligations. **(c-d)** In the last centuries of the Early New High German period, a transition happens and *können* becomes the preferred way to express physical and mental abilities.

It is also worth noticing that the use of ASNs goes beyond the traditional frequency based analysis for language change. *Können* is not as frequent in Early New High German [18] and its frequency is less than half of the instances compared to *mögen* in our corpus as well. Nonetheless, the hierarchical analysis of the ASNs ranks *können* almost as high as *mögen*. Traditionally, historical linguistic analyses rely on frequency to track the evolution of specific words or constructions [20]. Without relying on frequency, we could indeed identify *können* as one of the most influential nodes in our ASNs, although the instances of this verb were less than half compared to other influential nodes.

In essence, our approach captures multiple syntactic changes such as the rise of *können* as a modal verb (Section SI3) and its corresponding hierarchies. This suggests that similar processes have given rise to these linguistic hierarchies in the first place, when the human language was changing from a non-syntactic to a syntactic communication system [6] shaping syntax over time. The resulting multi-hierarchical structure is a reflection of the language effectively managing its duty to serve the communicative needs of the speakers. How heads are selected to carry out specific communicative needs is likely triggered by a combination of semantic and pragmatic factors that created the right environment for these hierarchies to rise.

Limitations

Satisfying the communicative needs of speakers is an extremely complex problem. To carry out this task, human languages have developed a broad varieties of resources ranging from phonology and morphology to syntax and semantics. The main focus of this study is on syntax and its interplay with semantics. Hence, our ASNs do not capture other linguistic aspects such as morphology or phonology. The main reason to focus on syntax is its structural uniqueness and characteristics as well as the type of data that is available to conduct historical linguistic analysis.

Moreover, the list of communicative hierarchies that we conceptualized in this work is not finite and is bounded to our corpus selection to analyse language evolution and change. Our corpus selection is a representative of the language around the verb *werden*. The reasons be-

hind this choice are explained in Section SI1. To identify a broader range of communicative hierarchies, ASNs on more languages and their evolution over time should be studied. Also, our study revealed a multi-hierarchical structures behind the organization of languages through time. This does not convey that all aspects of language also organize through hierarchies as well.

Lastly, we have to acknowledge that working with historical data is challenging, since the availability of corpora is not as high as it is for modern languages. Also, available algorithms to parse language syntax only work with modern languages, so we had to manually build the syntactic trees for each sentence.

Discussion

Hierarchies are the invisible backbones of complex systems around which they self organize [21–24]. Languages are also complex systems that change and evolve to satisfy the communicative needs of the speakers over time [25–27]. For this reason, the grammar by which we speak today is considered as “epiphenomenal of communication” [7]. This work also gravitates around the socio-individual needs of speakers to communicate with the other members of the society. Hence, we broke down communication into specific “communicative needs”, which allowed us to take advantage of the interplay between semantics and syntax. We mapped these communicative needs to specific hierarchies in our ASNs and we tracked their evolution through a time span of eight centuries. By doing that, a central question arose: what are the underlying mechanisms behind the emergence of those communicative hierarchies that we were able to identify from the hidden hierarchical characteristics of ASNs? The clues to answer this question lay in the global characteristics of the ASNs and their relation to the socio-individual needs of speakers to connect and communicate with each other.

We propose that the emergence of communicative hierarchies is a driving force in what shapes syntax, and the multi-hierarchical shape of syntax could provide a better way to understand the emergence of Zipf’s law in human language. It is often considered that the general objective of language evolution is to maximize information transfer while minimizing the effort for the hearer [13]. However, the observed multi-hierarchical structure could provide a more precise explanation on the general goal of language being optimized through evolution. The human language is the battleground of communicating an ever-widening range of abstractions. The current communicative needs that we found in ASNs are the manifestation of the need to communicate abstract ideas that already found their way into the structure of our language. The more advanced human language gets, the more complex abstractions will need to be communi-

cated. This means that the language will experience an increase in communicative hierarchies, which means a more hierarchical organization.

The multi-hierarchical structure we observed could also be described as polyarchy. Polyarchy as a term was coined in political science and it is a type of democracy in which the power is split among multiple entities. Since we saw a resemblance between the hierarchical organization of our syntactic networks and polyarchal governments, we decided to explore some possible connection between language change and political science in SI 3.4. Such exploration could be interpreted as an attempt to find a way to scientifically predict the trajectory of language change through centuries.

Acknowledgments

The authors acknowledge Csaba Both, Ph.D. student in the Network Science Institute at Northeastern University, Boston for his help and guidance on visualizing 3D layouts of the communicative hierarchies.

Author contributions

B.R. and V.C. conceived and designed the project, interpreted the results, wrote the manuscript, and bridged the concepts from network science and complex systems theory to linguistics. B.R. performed data modeling, statistical analysis, data query and integration. V.C. created the historical syntactic networks, performed linguistic analysis, and validated the results from the linguistics perspective.

References

- [1] Wilson, E. O. Animal communication. *Scientific American* **227**, 52–63 (1972).
- [2] Hauser, M. D. *The evolution of communication* (MIT press, 1996).
- [3] Bickerton, D. *Language and species* (University of Chicago Press, 1992).
- [4] Van Valin, R. D., van Valin Jr, R. D., van Valin Jr, R. D., LaPolla, R. J. & LaPolla, R. J. *Syntax: Structure, meaning, and function* (Cambridge University Press, 1997).
- [5] Wechsler, S. *Word meaning and syntax: Approaches to the interface*, vol. 9 (Oxford Surveys in Syntax & Mor, 2015).
- [6] Chomsky, N. *Aspects of the Theory of Syntax*, vol. 11 (MIT press, 2014).

- [7] Hopper, P. Emergent grammar. In *Annual Meeting of the Berkeley Linguistics Society*, vol. 13, 139–157 (1987).
- [8] Bybee, J. L., Perkins, R. D., Pagliuca, W. *et al.* *The evolution of grammar: Tense, aspect, and modality in the languages of the world*, vol. 196 (University of Chicago Press Chicago, 1994).
- [9] Leuschner, T., Mortelmans, T. & De Groodt, S. *Grammatikalisierung im Deutschen*, vol. 9 (Walter de Gruyter, 2005).
- [10] Nowak, M. A., Plotkin, J. B. & Jansen, V. A. A. The evolution of syntactic communication. *Nature* **404**, 495–498 (2000). URL <https://doi.org/10.1038/35006635>.
- [11] Ferrer i Cancho, R., Riordan, O. & Bollobás, B. The consequences of zipf’s law for syntax and symbolic reference. *Proceedings of the Royal Society B: Biological Sciences* **272**, 561–565 (2005). URL <https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2004.2957>. <https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.2004.2957>.
- [12] Solé, R. Syntax for free? *Nature* **434**, 289–289 (2005).
- [13] Cancho, R. F. i. & Solé, R. V. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences* **100**, 788–791 (2003). URL <https://www.pnas.org/content/100/3/788>. <https://www.pnas.org/content/100/3/788.full.pdf>.
- [14] Concu, V. Werden and the periphrases with present participles and infinitive verbs: A diachronic corpus analysis. *Journal of Germanic Linguistics* **138**, 195 (2022).
- [15] Lindeman, R. L. The trophic-dynamic aspect of ecology. *Ecology* **23**, 399–417 (1942).
- [16] Moutsinas, G., Shuaib, C., Guo, W. & Jarvis, S. Graph hierarchy: a novel framework to analyse hierarchical structures in complex networks. *Scientific Reports* **11**, 13943 (2021). URL <https://doi.org/10.1038/s41598-021-93161-4>.
- [17] Bybee, J. From usage to grammar: The mind’s response to repetition. *Language* 711–733 (2006).
- [18] Diewald, G. *Die Modalverben im Deutschen: Grammatikalisierung und Polyfunktionalität* (2012).

- [19] Dehmamy, N., Milanlouei, S. & Barabási, A.-L. A structural transition in physical networks. *Nature* **563**, 676–680 (2018). URL <https://doi.org/10.1038/s41586-018-0726-6>.
- [20] Hopper, P. J. & Traugott, E. C. *Grammaticalization* (Cambridge University Press, 2003).
- [21] Wu, J. *Hierarchy Theory: An Overview*, 281–301 (Springer Netherlands, Dordrecht, 2013). URL https://doi.org/10.1007/978-94-007-7470-4_24.
- [22] Corominas-Murtra, B., Goñi, J., Solé, R. V. & Rodríguez-Caso, C. On the origins of hierarchy in complex networks. *Proceedings of the National Academy of Sciences* **110**, 13316–13321 (2013). URL <https://www.pnas.org/content/110/33/13316>. <https://www.pnas.org/content/110/33/13316.full.pdf>.
- [23] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002). URL <https://www.science.org/doi/abs/10.1126/science.1073374>. <https://www.science.org/doi/pdf/10.1126/science.1073374>.
- [24] Trusina, A., Maslov, S., Minnhagen, P. & Sneppen, K. Hierarchy measures in complex networks. *Phys. Rev. Lett.* **92**, 178702 (2004). URL <https://link.aps.org/doi/10.1103/PhysRevLett.92.178702>.
- [25] The “Five Graces Group” *et al.* Language is a complex adaptive system: Position paper. *Language Learning* **59**, 1–26 (2009). URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9922.2009.00533.x>. <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9922.2009.00533.x>.
- [26] Cameron, L. & Larsen-Freeman, D. Complex systems and applied linguistics. *International journal of applied linguistics* **17**, 226–239 (2007).
- [27] Ellis, N. C. & Larsen-Freeman, D. *Language as a complex adaptive system*, vol. 11 (John Wiley & Sons, 2009).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TheHierarchicalOrganizationofSyntaxSI.pdf](#)