SUPPLEMENTARY INFORMATION

The Hierarchical Organization of Syntax

Babak Ravandi[1,2,†,*] Valentina Concu[3,*,†]

[1]Network Science Institute, Northeastern University, Boston, USA
[2]Department of Physics, Northeastern University, Boston, USA
[3]Department of Foreign Languages, Universidad del Norte, Barranquilla, Colombia

Contents

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

1 Data Collection

Our corpus selection is a representative of the language around the verb *werden*. *Werden* can be combined with a large variety of elements depending on the grammatical functions it is carrying out in a sentence. The functions of *werden* in Modern German are represented in Figure SI1. This section aims to explain why our corpus selection gravitates around this verb.

The texts from which we have collected the data come from two different database, one for Middle High German (Referenzkorpus Mittlehochdeutsch) [1] and one for Early New High German (Bonner Frühneuhochdeutschkorpus) [2]. The selection of the text were based on both internal and external textual features [3]. As for the first type, these features were based on syntactic, lexical, and semantics aspects. For instance, since poetic metric has been shown to affect and, sometimes, even "distort syntactic structures," [4] we selected prose texts only. Further, in order for us to concentrate on texts with similar topics, we limited our selection to religious, legal, and literary works. The aforementioned criteria (text type and genre - syntactic, lexical, and semantic aspects) allow us to ensure "corpus homogeneity" [3] and make the results from Middle and Early New High German texts comparable with each other. The external features that influenced our selection were chronological and geographical origin. To work on a variety of texts that could best represent the language periods we are studying, we selected texts from all the attested five different dialect area and from all the centuries traditionally included in the Middle and Early New High German periods. Specifically, the Middle High German corpus consists of 30 different texts form the 11th, 12th–13th, and early 14th centuries in West Middle German, East Middle German; West Upper German, East Upper German, and North Upper German. In the same way, the corpus for Early New High German includes 20 prose texts from the late 14th, 15th, 16th, and 17th centuries in West Middle German, East Middle German, West Upper German, East Upper German, and Nord Upper German. This corpus was already used to track the development of *werden* as the auxiliary for the expression of future references [5]

The data were collected from the aforementioned annotated online corpora and the instances of*werden* were manually parsed. This process is highly demanding in terms of time and limits the quantity of data that can be included for analysis. However, manually parsing allows us to have homogeneity across the corpora, which makes the comparison between Middle and Early New High German consistent.

We created tow different databases, one for Middle High German and one for Early New High German. All the words are listed in their lemma form. Each database contains all the tokens of *werden* in the following combinations: [WERDEN + past participle], [WERDEN +

present participle], [WERDEN + adjective], [WERDEN + infinitives], [WERDEN + nouns], and [modal verb + WERDEN]. The analysis did not include sentences with werden in which the annotations of one of the elements was missing. The corpora display missing words and annotation as [!] in the Middle High German corpus, and as "unbekannt" (unknown) in the Early New High German corpus, as shown in Figure 2.

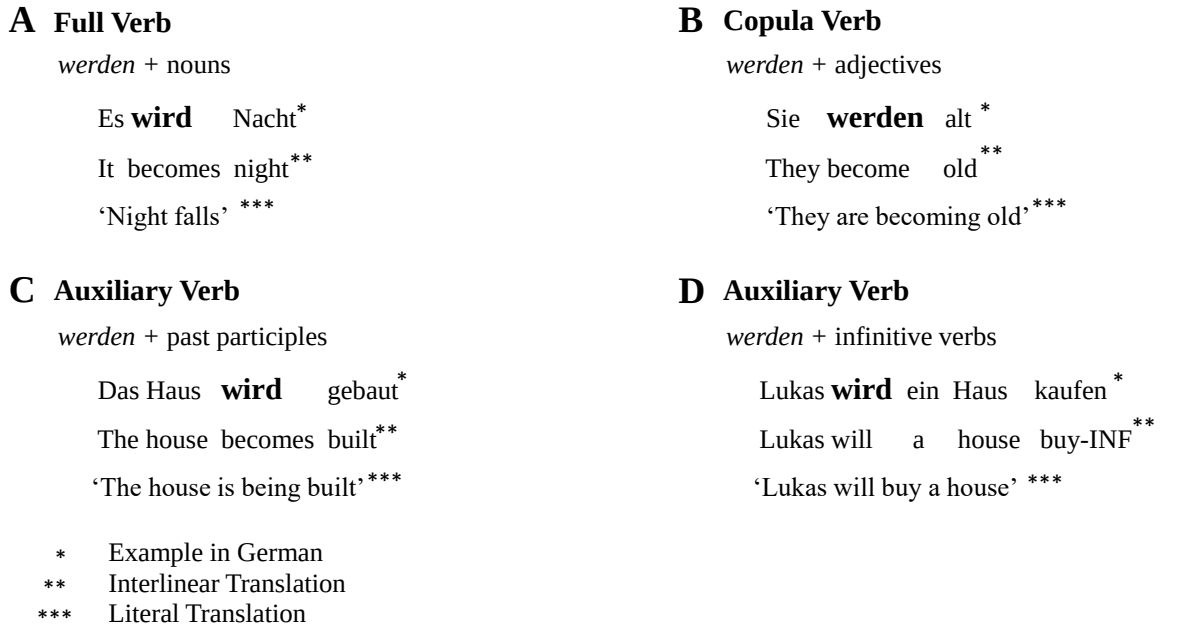

Figure SI 1

Figure SI 1: **Functions of *Werden* in Modern German.** *Werden* can be combined with a large variety of elements depending of the grammatical functions it is carrying out in a given sentence. *Werden* acquired the capacity to carry out these functions due to a process of *Desemantisierung* (desemantization) that allowed this verb to appear in combinations with an increased type of elements throughout the centuries. In the examples **A-D**, the first line represents a sentence in Modern German. The second line contains the interlinear morpheme-by-morpheme glosses (translation of each word). The third line is the literal translation of the sentence in English. The functions of *werden* are: **(A) Full Verb** with its own meaning of 'to become'; **(B) Copula Verb** with nouns and adjectives (connecting a subject with a subject complement); **(C-D) Auxiliary Verb** with past participles and infinitive verbs.

A

unte dero sententia **uuart** reprobate *

and of their [!] became [!] **

'and of their…..was….' ***

Hoheliedkommentar, 11v,05 ****

B

ze vile [!!] **worden** sint die [!!] *

too many [!] became are the [!] **

'too many had been…the' ***

Rheinfränkische Interlinearversion der Psalmen, 3v,4 ****

* Example from the Middle High German Annotated Corpus
** Interlinear Translation
*** Literal Translation
**** Source and Location in Text
[!] Missing Annotation in Corpus
[!!] Missing Word in Manuscript

Figure SI 2: **Missing Words and Annotations in the Corpus.** The Middle and Early New High German corpora use different levels of annotations providing translation, grammatical functions, inflections, and more. However, the corpus has missing words or annotations that affects the data collection and the related creation of the networks. **(A)** The verb *werden* ('*uuart*') is used in combination with the words *'sententia'*'sententia' and '*rebrobate*', but no annotations are provided for any of them. Hence, it is not possible to establish how the target verb was used in this instance. Therefore, we filtered all sentences with this condition in our selection. **(B)** The missing words and annotations in this instance refer to elements not directly connected with *werden*. The presence of the auxiliary verb *sein* (to be - '*sint*') *werden* in the participle form ('*worden*'), and the definite article *die* suggest that the target verb has been used as a copula and can be, therefore, included in our selection. Accordingly, we included sentences with this type of missing annotations in our data as they do not interfere with identifying the grammatical function of *werden*.

2 Network Creation and Properties

2.1 Syntactic Trees

Syntactic trees are commonly used in dependency grammar to map the relationships between words. Dependency grammar is based on the notion of "grammatical relation" according to which the relationship between a head and a dependent are established in a given sentence. [6] The head is the word in the phrase that is grammatically the most important since it determines the syntactic category of of a larger constituent and its the central organizing word (e.g., the primary noun in a noun phrase, or verb in a verbal phrase). The remaining words in the constituent are either directly, or indirectly, dependent on their head. [6] Furthermore, "the head-dependent relationship is made explicit by directly linking heads to the words that are immediately dependent on them." [6]

The relationships established between the head and its dependent "allows us to further classify the kind of grammatical relations, or grammatical functions, in terms of the role that the dependent plays with respect to its head. [6] Dependency grammar uses "dependency trees" to structuralize and visualize the relationship between a head and its dependents. Figure 3

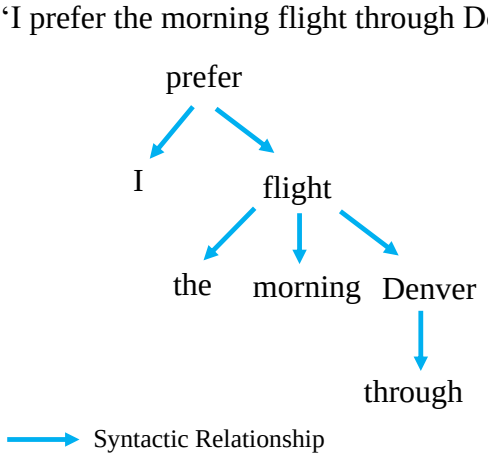

Figure SI 3: **Dependency Syntactic Tree.** Syntactic trees map the relationships between words in a sentence and are based on the notion of “grammatical relation.” According to this notion, the relationships between a head and a dependent are established in a given sentence. The root is selected based on the concept of a governing element that corresponds to the most grammatically important element among the constituents. The governing element in this sentence is the verb “prefer” on which both the pronoun “I” and the noun “flight” depend. At the same time, the noun “flight” functions as the head with three dependents: the definite article “the”, and the nouns “morning” and “Denver”. “Denver” has also another dependent, namely the preposition “through (modified and used with permission from [6]).

shows a syntactic tree for a sentence in Modern English. According to Jurafsky and Martin, [6] dependency structures as the one shown in Figure 3 are direct graphs that respect the following constraints:

1. There is a single designated root node that has no incoming arcs.

2. With the exception of the root node, each vertex has exactly one incoming arc.

3. There is a unique path from the root node to each vertex in the graph.

We followed the aforementioned constraints to build individual syntactic dependency trees and consequently our ASNs.

2.2 Building Aggregated Syntactic Networks

In this section, we will describe the network mapping. As previously mentioned, the rules used to build such networks are inspired by the principles of dependency grammar. [6] There is a total of three rules in the current study, and these reflect the syntactic relations between the elements in a sentence. A rule is classified as nominal phrase (NP) if it has a noun or a pronoun as the head. A rule is classified as verbal phrase (VP) if it has a verb as the head. A rule is classified as prepositional phrase (PP) if it has a preposition as the head. All the source and target nodes are listed in their lemma form which corresponds to the canonical word form

(for instance, the singular form of a noun is the lemma for a plural form, or the infinitive form of a verb is the lemma of an inflected form of a verb). Each node in the network represents the lemma of a given word, We have also assigned to every node in the source and target columns a label according to its grammatical role. The complete list of grammatical roles is displayed in Table SI1.

Table SI 1: Grammatical Roles Used in ASNs.

Label	Grammatical Role	Label	Grammatical Role
AD	Adverb	PR	Preposition
AJ	Adjective	PP	Personal Pronoun
AR	Article	PS	Possessive Pronoun
AX	Auxiliary	PCPR	Present Participle
CJ	Coordinating Conjunction	PCPS	Past Participle
DM	Demonstrative Pronoun	RX	Reflexive Pronoun
IV	Infinitive Verb	RPO	Relative Pronoun
MV	Modal Verb	SC	Subordinating Conjunction
N	Noun	V	Verb
PK	Particle		

After assigning to each word its grammatical role, we created the network for each sentence in which the target verb *werden* was found. Afterwards, we merged all the sentences in one network for each one of the eight centuries included in our analysis. The resulting networks, which we decided to name Aggregated Syntactic Networks (ASN), are not trees anymore, as multiple paths now exist between heads and dependents. Table SI2 presents the general network characteristics of ASNs for each historical period.

To further investigate the difference between the characteristics of sentence dependency trees and ASNs, we compared the depth of trees with the lengths of shortest distances in ASNs. Someone may raise the point that an ASN is a dependency tree that is growing in width and that the merging trees may not reflect the global syntactic characteristics of our corpus. However,

Table SI 2: Characteristics of Aggregated Syntactic Networks

Corpus	**Century**	**Nodes**	**Edges**	$< k >$	$< c >$
Middle High German	11th Century	411	723	3.52	0.06
	12th Century	434	803	3.7	0.05
	13th Century	697	1325	3.8	0.05
	Early 14th Century	686	1321	3.85	0.14
Early New High German	Late 14 Century	751	1585	4.22	0.08
	15th Century	735	1422	3.87	0.05
	16th Century	1187	2446	4.12	0.07
	17th Century	1561	3076	3.95	0.07

as presented in Figure SI4, we observe a structural difference in ASNs starting from the 16th century. The network diameter and average pair-wise shortest path length suddenly are as twice as the depth of largest sentence in our corpora. This phenomenon can be captured by ASNs and requires further investigation.

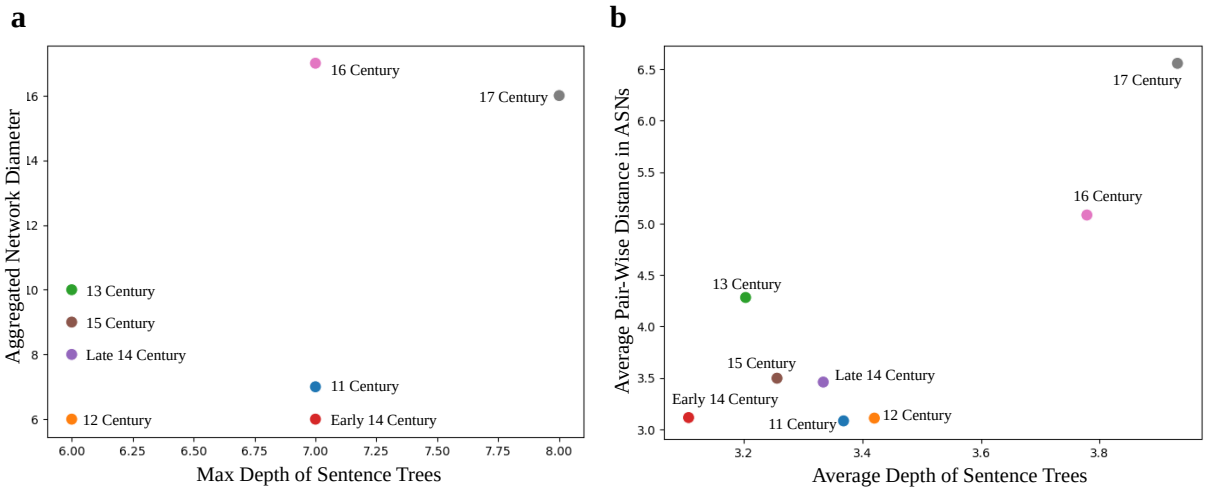

Figure SI 4: **Individual Tree's Depth vs Diameter of ASNs.**

2.3 Degree Distribution of Aggregated Syntactic Networks

There are many linguistics phenomena that follow a power-law distribution as initially discovered by Zipf's law [7]. To evaluate whether power-law distribution is a statistically plausible in our ASNs, we utilized the semi-parametric bootstrap approach combined with goodness-of-fit tests based on the Kolmogorov-Smirnov (KS) statistic [8]. KS statistic and the likelihood ratio test (LRT) showed that except for 17th century, the power-law model is favored over the alternatives models including lognormal, exponential, stretched exponential, and truncated power law [9]. Figures SI5 and SI6 present the details, the calculated p-value for the estimated power-law fit is > 0.01 (except for 17th century) that signals power-law model is favored over the alternatives by the LRT, hence we have a strong support for the data following a power-law distribution. In the case of 17th century, the test does pass with starting degree 13 with $\alpha = 1.91$ and $p = 0.72$, however that is not a consistent behaviour with the previous centuries.

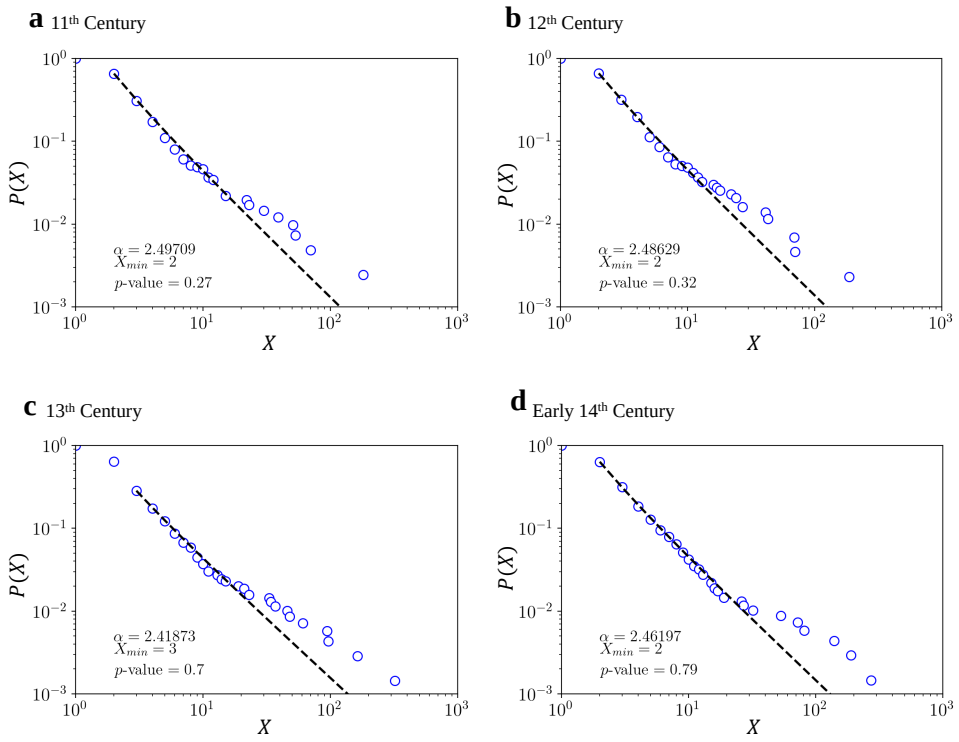

Figure SI 5: Degree distribution of ASNs in Middle High German.

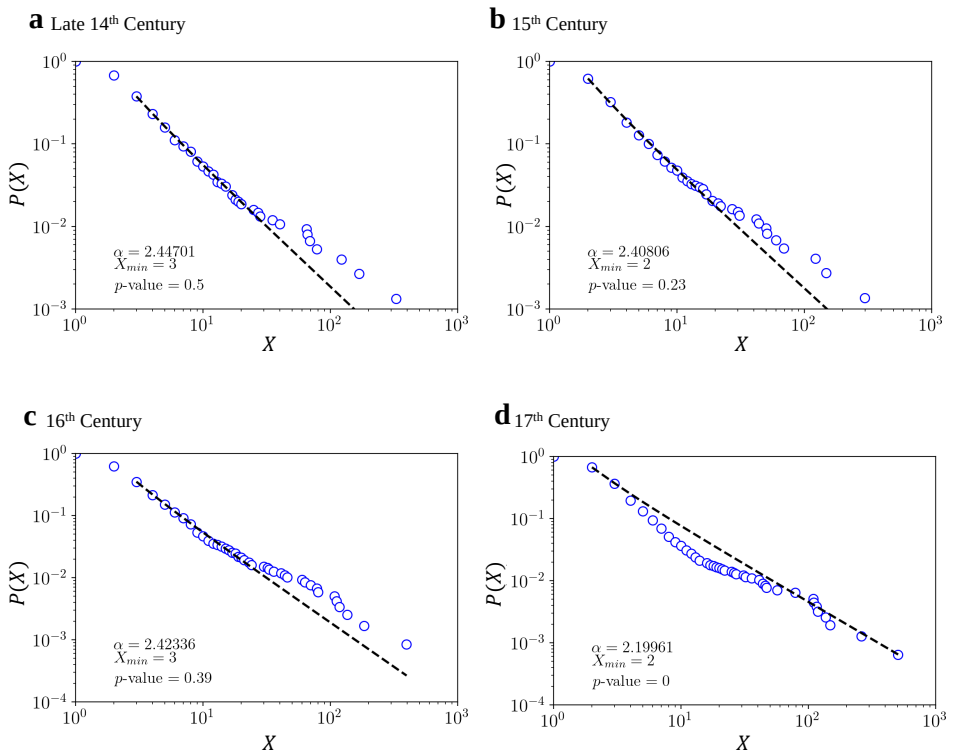

Figure SI 6: Degree distribution of ASNs in Early New High German.

3 More on Hierarchies

3.1 Communicative needs

Communication has been recognized as a driving force for language evolution and change, and the grammar and the syntax of a language have been described as "epiphenomenal" of communication [10]. In our work, communication also plays an important role. More specifically, we focused on what we called communicative hierarchies," which allowed us to take advantages of the interplay between semantics and syntax. In our work, a communicative need corresponds to the resulting meaning expressed by the syntactic head and its dependants in any given tree. For instance, the trees with the auxiliary *werden* as the syntactic head that has a past participle as a direct dependent is used express the passive voice. It serves the communicative need of focusing on the patient and not on the active subject of the sentence. Hence, we used the syntactic heads to track down what kind of meanings could be mapped to specific hierarchies in our ASNs in a time spam of eight centuries.

Table SI3 contains the list of communicative hierarchies that we identified in our ASNs. The chart shows the related syntactic heads, together with the historical stage, Middle High German (MHG) and Early New High German (ENHG) in which they were found or emerged. Note that some heads are associated with multiple communicative needs, while some of them acquired a communicative needs later on, like in the case of future references (transition from *sollen* to *werden*) or the expression of mental and physical capabilities (transition from *mögen* to *können*).

3.2 Further Validation

3.2.1 Grammaticalization of the German Present Perfect

Attestations of the periphrastic constructions with the auxiliaries *haben* and *sein* plus a past participle can be found already in Old High German [11]. At this point, both auxiliaries and past participles were two independent units that expressed possession [12]. The grammaticalization of this combination as the modern present perfect that could express, among others, past references, started ind Middle High German and culminates in Early New High German [11]. Such process is captured by two different measures in our analysis. The grammaticalization of the present perfect allowed this construction to be used in the passive voice with the auxiliary verb *werden*. Until the last centuries of the Middle High German period, the passive voice plus past references appeared only in the simple past. The use of the present perfect in the last centuries affects the average path length of the networks, especially in the last two centuries

Table SI 3: Identified Communicative Hierarchies.

Communicative Need	Transition of Syntactic Heads	
	From	To
Express preferences	MV mügen*,**	
Express mental and intellectual capacities	MV mügen*	MV können**
Express obligations	MV mügen*	MV sollen** – MV müssen**
Express being powerful	MV mügen*	
Future references	MV sollen*	AX werden**
Express Needs	MV sollen*,**	
Ask for Permission	MV sollen*	MV dürfen**
Express change of state	V werden*,**	
Passive voice	AX werden*,**	
Desires and Wishes	MV wellen*,**	

* Middle High German
** Early New High German

of the Early New High German period, as shown in Figure SI4. The aforementioned syntactic change is also captured by the forward hierarchical levels. *Sein*'s value increases indeed in the last two centuries of the Early New High German period (From $FHL = 2.313$ to $FHL = 2.934$).

3.2.2 The Rise of *werden* as the Auxiliary for Future References

Sollen was used in Old and Middle High German as a way to express future references [5] In Early New High German, this function is slowly taken over by the verb *werden* [5] The weakening of the semantic features of *werden* between the late 14th and the 17th centuries and the related increase in frequency of the combination with infinitive verbs is also captured by the forward hierarchical levels. The infinitive verbs increase their hierarchical levels, especially in the last centuries of the Early New High German period.

3.2.3 The Fall of the periphrastic constructions of *werden* plus present participles

One of the syntactic changes captured in our ASNs is the disappearances of the periphrastic construction of *werden* plus the present participles. This construction indicated non-mutative and non-terminative actions, conveying a meaning that was close to the semantic of any give verb in its finite form: habund bist vs. hâst or redened war vs. redete – 'you are having' vs. 'you have' or 'you were speaking' vs. 'you spoke' [12]. It was exactly this redundancy in meaning that caused the decline of this periphrasis, which also prompted the disappearance of

the constructions of *werden* plus present participle [12]. This process is captures also by the forward hierarchical levels, as the present participles decreases their values in the Early New High German period, exactly when is disappearance has been attested.

3.3 Distribution of Hierarchical Levels

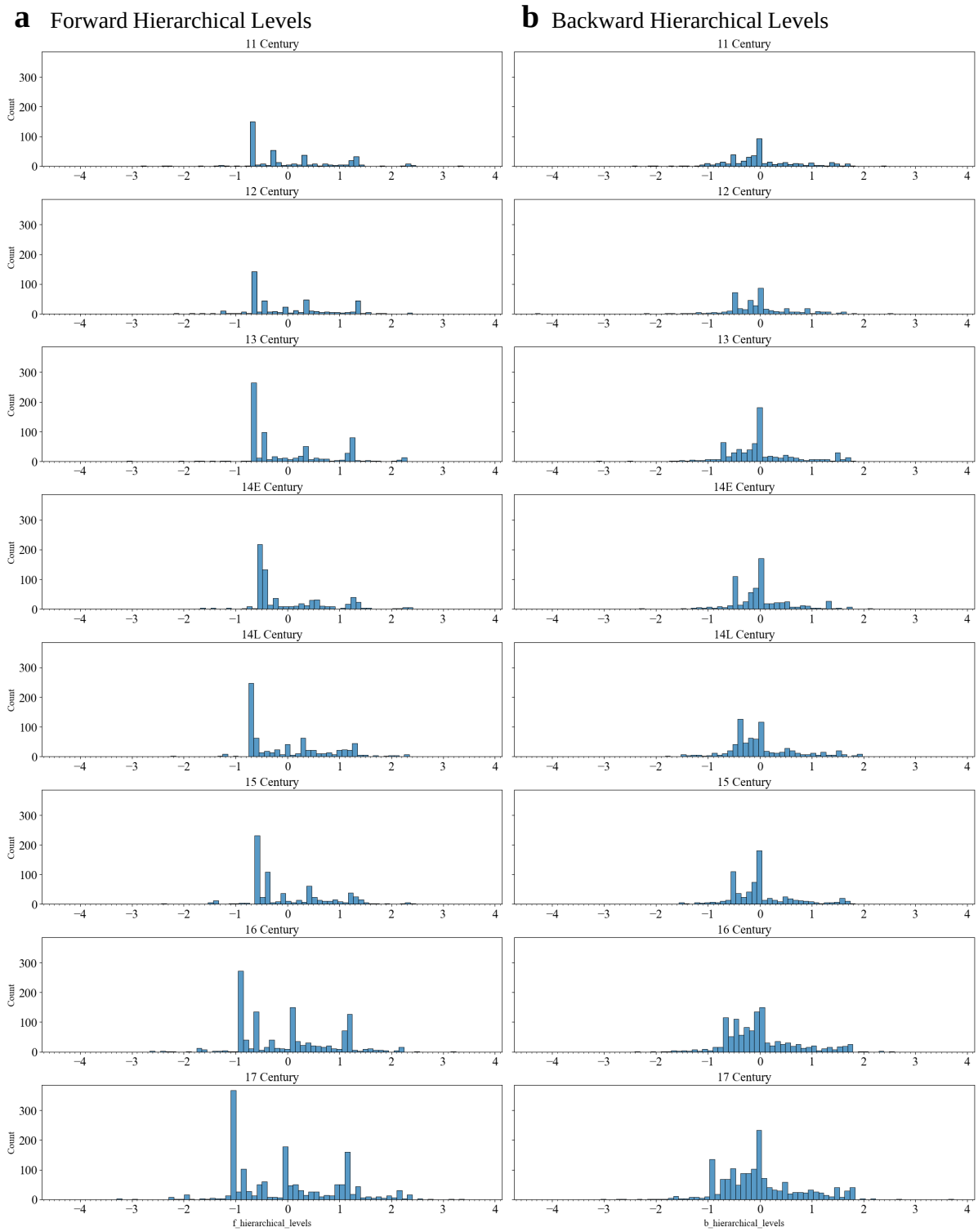

Figure SI 7: **Distributions of forward and backward hierarchical levels from 11th to 17th century.**

3.4 Possible Connection with Political Science

We observe that our ASNs have multiple heads, revealing a multi-hierarchical organization. Another term used to describe an entity with multiple hierarchies is polyarchy. Polyarchy has been defined in political science as a type of modern democratic representative government for large scale states or countries and means "rule by the many" [13]. Polyarchy itself can be quantified through two measurements: liberalization and inclusiveness [14]. Liberalization corresponds to public contestation, which translates into the degree by which a government allows any type of political opposition. Inclusiveness accounts for the ability of citizen to participate in controlling and contesting the conduct of the government. The more a society allows public contestation and the participation of its citizens to its political life, the more polyarchal will be, as illustrated in Figure SI8a. Since languages and societies have co-evolved and affected each others throughout the centuries [15], we suspect that those processes that are involved in the democratization of a society could also be behind the language change. If so, these processes could provide us a conceptual framework to understand the trajectory of language evolution. Hence, we mapped liberalization and inclusiveness to democracy coefficient and hierarchical incoherence respectively.

The democracy coefficient of complex networks indicates the degree by which the influencers at the top of the hierarchies influence each other. In our ASNs, it translates into the ability of communicative hierarchies to share the expression of multiple communicative needs. The hierarchical incoherence indicates how neatly a graph can be partitioned into discrete levels,

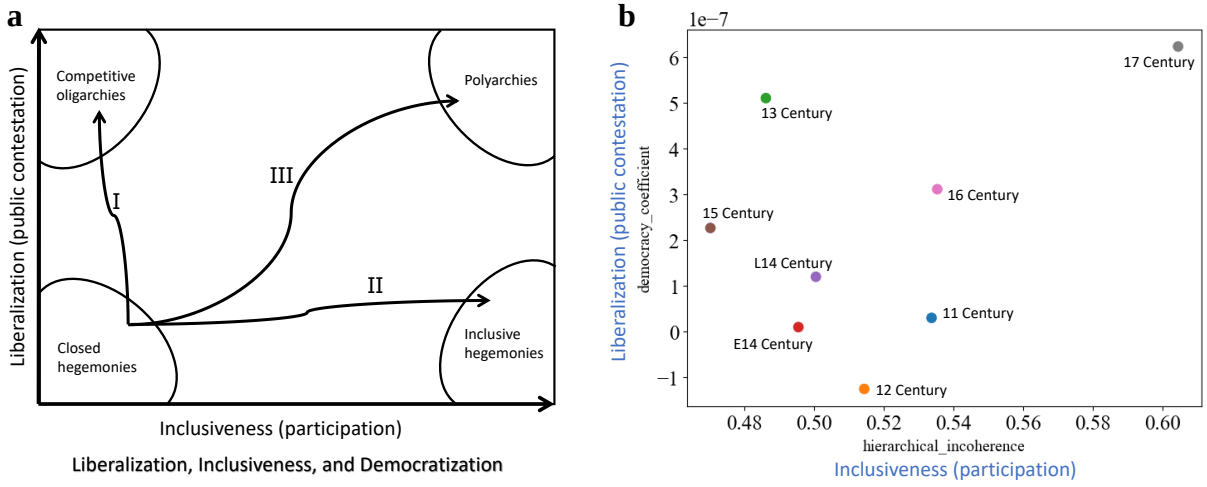

Figure SI 8: **Trajectory of Language Change.** (**a**) Dahl indicates democratization of societies increase inclusiveness and public contestation, hence moving toward polyarchies. (**b**) The trajectory of language change with respect to macro-level hierarchical characteristics of ASNs signals that the same processes behind democratization may also be behind the change in language.

and it is derived from the hierarchical levels of nodes. We note that the number of words in our corpora selection is equal for all centuries (Section SI1), however the distribution of forward hierarchical levels tend to form more and more distinct strata of words as the language evolves (Figure SI7). This signals that the functions of words in language are becoming more specific and, as a consequence, they are acquiring a more defined place in the communicative hierarchies. Figure SI8 presents the trajectory of democracy coefficient and hierarchical incoherence of ASNs, we observe that both measures tend to increase as the language evolved.

As human beings started to settle down and organized themselves in fixed communities, the number of the members of these communities also increases. It is at this point that the first social hierarchical structures emerged [13]. We can assume that the hierarchical structures that are found in ASNs also emerged as the number for words available to the speakers also increased. As already mentioned, Dahl claims that polyarchal democracies are democratic government on the large scale of the nation-state country and are the most effective way to govern, allowing to provide political representations as the number of citizens grows in a country [13]. In the same way, the multi-hierarchical structure of the language that we observed in our ASNs could be the most effective way for the language to handle the always increasing number of words and maintain its capability to satisfy the communicative needs of the speakers.

References

[1] Klein, T., Wegera, K.-P., Dipper, S. & Wich-Reif, C. Referenzkorpus mittelhochdeutsch (1050–1350), version 1.0 (2016).

[2] Schmitz, H.-C., Schröder, B. & Wegera, K.-P. Das bonner frühneuhochdeutsch-korpus und das referenzkorpus ‚frühneuhochdeutsch (2011).

[3] Cavaglià, G. Measuring corpus homogeneity using a range of measures for inter-document distance. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)* (European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain, 2002). URL `http://www.lrec-conf.org/proceedings/lrec2002/pdf/232.pdf`.

[4] Somers, K. & Dubenion-Smith, S. The intersection between syntax and meter in the old saxon. *Amsterdamer Beiträge zur älteren Germanistik* **72**, 83–134 (2014).

[5] Concu, V. Werden and the periphrases with present participles and infinitive verbs: A diachronic corpus analysis. *Journal of Germanic Linguistics* **138**, 195 (2022).

[6] Jurafsky, D. & Martin, H. J. *Speech and language processing* (XXX, 2009).

[7] 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.

[8] Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Review* **51**, 661–703 (2009). URL https://doi.org/10.1137/070710111. https://doi.org/10.1137/070710111.

[9] Alstott, J., Bullmore, E. & Plenz, D. powerlaw: A python package for analysis of heavy-tailed distributions. *PLOS ONE* **9**, 1–11 (2014). URL https://doi.org/10.1371/journal.pone.0085777.

[10] Hopper, P. Emergent grammar. In *Annual Meeting of the Berkeley Linguistics Society*, vol. 13, 139–157 (1987).

[11] Concu, V. Toward the german present perfect as an emergent structure. *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis* **21**, 111–164 (2016).

[12] Kuroda, S. Michail l. kotin: Die werden-perspektive und die werden-periphrasen im deutschen. frankfurt am main, ua: Peter lang 2003 (= danziger beiträge zur germanistik 6). *Neue Beiträge zur Germanistik* **124**, 137–140 (2005).

[13] Dahl, R. A. *On democracy* (Yale university press, 2020).

[14] Dahl, R. A. *Polyarchy: Participation and opposition* (Yale University Press, 2008).

[15] Hopper, P. J. & Traugott, E. C. *Grammaticalization* (Cambridge University Press, 2003).