DE GRUYTER OPEN

Yearbook of the Poznań Linguistic Meeting 1 (2014), pp. 89–102
DOI: 10.1515/yplm-2015-0005

An application of graph theory to linguistic complexity

Alexander Piperski
Russian Academy of National Economy, Moscow
apiperski@gmail.com

Abstract

This article introduces a new measure of linguistic complexity which is based on the dual nature of the linguistic sign. Complexity is analyzed as consisting of three components, namely the conceptual complexity (complexity of the signified), the formal complexity (complexity of the signifier) and the form-meaning correspondence complexity. I describe a way of plotting the form-meaning relationship on a graph with two tiers (the form tier and the meaning tier) and apply a complexity measure from graph theory (average vertex degree) to assess the complexity of such graphs. The proposed method is illustrated by estimating the complexity of full noun phrases (determiner + adjective + noun) in English, Swedish, and German. I also mention the limitations and the problems which might arise when using this method.

Keywords: linguistic complexity; morphology; graph theory; average vertex degree.

1. Introduction

Linguistic complexity has been extensively studied for more than two decades, starting with Nichols (1992) and McWhorter (2001), who incited a lively scholarly debate on this topic. The equi-complexity hypothesis, which was commonplace in the 20th century, has now been almost completely debunked (Shosted 2006; Sampson et al. 2009; Trudgill 2012). Since we assume that different languages exhibit different degrees of complexity, it is important to measure complexity and express it numerically. Otherwise, any discussion of complexity would suffer from vagueness and inexactness.

2. Grammatical complexity and form-meaning correspondence

Quantifying grammatical complexity has proven to be a difficult task. Even though it is clear that different languages exhibit different degrees of complexi-

ty, it is not easy to measure these differences exactly. Most attempts at estimating overall complexity take into account a fixed number of arbitrarily chosen features belonging to different levels of language structure (cf. Nichols 1992). Some other studies have tried to quantify the complexity of a certain linguistic level, such as phonology (Atkinson 2011) or morphology (Nichols et al. 2006), but they face the same problem, namely arbitrariness of the parameters chosen for the estimation. For this reason, I propose a measure of linguistic complexity which is based not on the arbitrarily chosen features, but rather on the most fundamental properties of human language.

Probably the most important characteristic of language is the dual nature of its signs which consist of the signifier and the signified (Saussure 1916: 97–101, etc.). This makes it possible to describe language proper as "a [...] correspondence between an infinite set of meanings and an infinite set of texts" (Mel'čuk and Pertsov 1987: 12). Actually, no general linguistic theory can dispense with a statement like this, cf.

> The grammar of a language, as a model for idealized competence, establishes a certain relation between sound and meaning – between phonetic and semantic representations. We may say that the grammar of the language L generates a set of pairs (*s*, *I*), where *s* is the phonetic representation of a certain signal and *I* is the semantic interpretation assigned to this signal by the rules of the language.
>
> (Chomsky 2006: 103)

It seems reasonable to assume that linguistic complexity is actually the complexity of this correspondence relation. However, the complexity of linguistic form and linguistic meaning themselves should not be neglected, too.

The level of linguistic structure where form and meaning interact most closely is morphology, because this is exactly the module that is concerned with mapping abstract meanings to concrete morphemes. I argue that the complexity of this mapping can be measured using methods from graph theory.

3. Graph theory and grammatical complexity

The morphological structure of a language can be represented using two tiers:

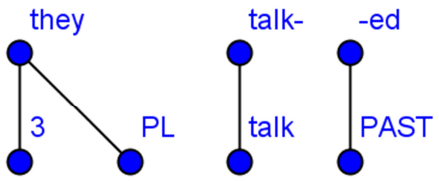
(1) the form tier;
(2) the meaning tier.

The form tier contains individual morphemes of a word, sentence, utterance or text, whereas the meaning tier contains grammatical and lexical meanings. It is not easy to define a minimum unit of meaning, and for practical reasons I consider it equal with a minimum element of an interlinear gloss. It is definitely a makeshift, but any other formal definition of meaning would be less operational.

Let us take the English sentence in (1) as an example:

(1) They talked

This sentence has three morphemes on the form tier (*they*, *talk*-, *-ed*) and four items on the meaning tier (3rd person, plural, past tense and the lexical meaning 'to talk'). This English example illustrates that there is no one-to-one correspondence between form and meaning. In this case, *they* expresses both person and number. This can be shown in a graph where vertices represent the elements on the form tier and on the meaning tier and edges mark correspondences between the two tiers:



Graph 1. English *They talked.*

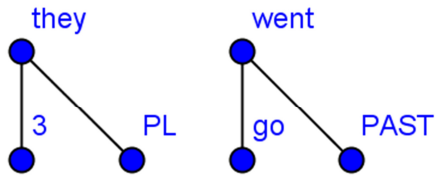
The correspondence relation between the meaning tier and the form tier can be defined as follows:

> Vertex A on the form tier corresponds to vertex B on the meaning tier iff the knowledge of B is required for producing the correct form of A

In Graph 1, the speaker has to know that the subject of the sentence is 3rd person plural in order to produce the correct form of the pronoun (in this case, *they*), and this means that the vertex *they* is connected with the vertices '3' and 'PL'.

The structure of the correspondences can be different not only across languages, but also with different words in the same language. Graph 2, which represents the English sentence (2), illustrates this.

(2) They went

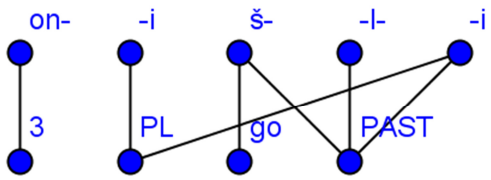


Graph 2. English *They went.*

Obviously, the same meaning is expressed in other ways in other languages. In (3) is the Russian sentence which means the same.

(3) Russian:
oni šli 'they went'

The form-meaning correspondences for (3) are shown in Graph 3:



Graph 3. Russian *Oni šli.*

In this case *š-* corresponds to 'go' and to 'PAST' (it is typical of Russian that the speaker needs to know the tense of the verb in order to choose the correct stem), and *-i* corresponds to 'PL' as well as 'PAST' since the set of endings is different

in the present and in the past, and tense should be taken into account when choosing the ending. In (2) and (3), the meaning tier is the same, but the form tier has more vertices in Russian (5 vs. 2), and the Russian structure of the correspondences between the two tiers seems to be more complex.

3.1. Graph-theoretic measures of grammatical complexity

Generally speaking, the number of vertices on each of the tiers and the structure of edges connecting the tiers is the manifestation of grammatical complexity. The most straightforward measure of graph complexity is counting the vertices. By counting the vertices on the form tier, we can measure the formal complexity of a linguistic unit (word, sentence, utterance, or text): clearly, the more morphemes the unit contains, the more complex it is. By counting the vertices on the meaning tier, we get a measure of the conceptual complexity of a linguistic unit: the more grammatical meanings the speakers have to express, the more complex the structure is.

It is also possible to quantify the complexity of the correspondence between vertices on the two tiers. A good measure of complexity for this purpose is the **average vertex degree**, i.e. the mean number of edges incident to a vertex in the graph (Bonchev and Buck 2005). Denoting the vertex degree of each vertex ($i = 1, 2, \ldots, n$) as a_i, it is possible to express the average vertex degree $\bar{a}_i$ as the total of all vertex degrees divided by the number of vertices (n):

$$\bar{a}_i = (a_1 + a_2 + \ldots + a_n) / n$$

Since each edge connects two vertices, the total of all vertex degrees is twice the number of edges (E), and the formula for average vertex degree can be simplified:

$$\bar{a}_i = 2 \times E / n$$

For Graph 1 representing the English sentence *They talked*, $\bar{a}_i = 2 \times 4 / 7 = 1.14$, and for Graph 2 (*They went*) $\bar{a}_i = 2 \times 4 / 6 = 1.33$. Thus, we can compare the complexity of these two sentences using three graph-theoretic measures described above:

– these sentences are equally conceptually complex because each of them expresses four distinct meanings;

– the sentence *They talked* is formally more complex because it contains more morphemes;
– the sentence *They went* is more complex because its average vertex degree is higher (i.e., the correspondences between form and meaning are organized in a more complex way than in the sentence *They talked*).

It might be tempting to derive a single complexity measure from these three measures, but it is unclear how it can be computed. For this reason, it would be more reasonable to say that the grammatical complexity of any linguistic unit has three dimensions:

– conceptual complexity (number of vertices on the meaning tier);
– formal complexity (number of vertices on the form tier);
– correspondence complexity (average vertex degree of the graph).

3.2. Limitations and complications

3.2.1. Segmentation issues

It is not always easy to segment a linguistic unit into morphemes. The problems of this kind have been extensively treated by morphologists, but in many cases no universally accepted solution has been met. For instance, when dealing with the German definite article, it is unclear whether it consists of two parts or not. The possible analyses of some forms are presented in (4) and (5).

(4) German:

d-er	d-em	d-ie
DEF-NOM.SG.MASC	DEF-DAT.SG.MASC	DEF-NOM.SG.FEM
das		
DEF-NOM.SG.NEUT		

(5) German:

der	dem	die
DEF.NOM.SG.MASC	DEF.DAT.SG.MASC	DEF.NOM.SG.FEM
das		
DEF-NOM.SG.NEUT		

The analysis presented in (4) seems less satisfactory because it requires stipulating the existence of endings which are unique to the definite article (e.g., *-as* for NOM.SG.NEUT). An additional argument for the second analysis is the length of morphemes proposed in (4): in most cases, the root of the German definite article consists of one phoneme, whereas the endings are two phonemes long. It is not impossible that the endings are longer than the root but still indicative. For these reasons, the definite articles of Germanic languages treated in Section 4 will always be left unsegmented, even though this solution is not undebatable.

The indefinite Germanic articles cannot be treated uniformly. For instance, the Norwegian indefinite article is clearly unsegmentable:

(6) Norwegian (Bokmål):

en	ei	et
DEF.SG.MASC	DEF.SG.FEM	DEF.SG.NEUT

However, the indefinite article of German inflects very much the same way as adjectives do, and this makes it possible to segment the indefinite article into a root and an ending:

(7) German:

ein-e	ein-er	ein-em
INDEF-NOM.SG.FEM	INDEF-DAT.SG.FEM	INDEF-DAT.SG.NEUT

The examples presented above show that there is no general solution for segmentation. However, if one tries to count edges and vertices on a form-meaning graph, it is necessary to make the segmentation procedure as transparent and well-argued as possible.

3.2.2. Zero markers

The presence or absence of zero markers is also a problem for plotting form-meaning graphs. If zero markers are accepted, the metric can capture no complexity difference between two languages like Italian and English:

(8) Italian:

tavol-a	tavol-e
table-SG	table-PL

(9) English:
table-Ø table-s
table-SG table-PL

However, this solution does not seem satisfactory because Italian with its overt number marking still seems more complex than English. For this reason, I assume that the complexity graphs for (8) and (9) look as follows:

Graph 4. Complexity graphs for English and Italian singular and plural nouns.

Thus, the vertex for a category that has no overt marker is present on the meaning tier, but remains unattached to any vertex on the form tier. The graph for English *table* with 1 form vertex and an average vertex degree of 0.67 is less complex than the graph for Italian *tavola* with 2 form vertices and an average vertex degree of 1, which conforms to the intuition that the English example has a less complex structure than the Italian one.

3.2.3. Cross-linguistic comparisons

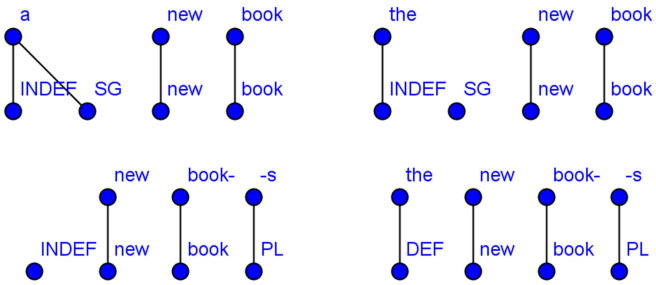
It is quite difficult to use the graph-theoretic method for comparing the overall grammatical complexity of different languages. First, it is unclear what texts can be taken for comparison because they should representative of languages as a whole. Second, it is not always easy to find equivalent texts in different languages, because we would have to deal either with cumbersome artificially constructed texts or with translations of literary or legal texts.

Therefore, the application of graph-theoretic method has to be limited not to comparing languages as a whole, but to comparing rather grammatical constructions in different languages. To show this, I turn to analyzing the structure of the full noun phrase (determiner + adjective + noun) in three Germanic languages.

4. Case study: Noun phrase in three Germanic languages

4.1. English

In English, the noun phrase has two grammatical features, namely definiteness and number. There are four possible types of noun phrases, since there are two possible values for definiteness and two possible values for number. This means that there are four vertices on the meaning tier for an English noun phrase: definiteness, number, lexical meaning of the adjective and lexical meaning of the noun. On the form tier, an English noun phrase can have at most four vertices (article, adjective stem, noun stem, noun ending). The correspondences are in all but one cases one-to-one, though some meaning vertices remain unattached to form vertices:



Graph 5. The structure of the English noun phrase.

For these graphs, the values of the parameters are as in Table 1.

Table 1. The structure of the English noun phrase.

	Definiteness	Number	Vertices (form)	Vertices (meaning)	Vertices (total)	Edges	Average vertex degree
a new book	INDEF	SG	3	4	7	4	1.14
the new book	DEF	SG	3	4	7	3	0.86
new books	INDEF	PL	3	4	7	3	0.86
the new books	DEF	PL	4	4	8	4	1.00

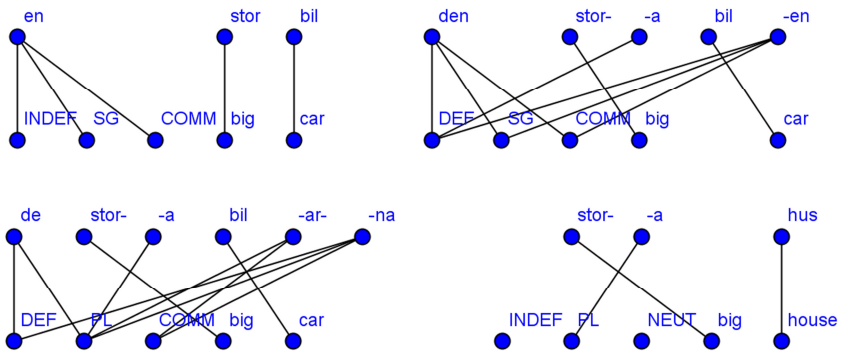
On average, the English noun phrase has 3.75 form vertices, 4 meaning vertices and 3.5 edges. The average vertex degree of such a graph would be 0.90.

4.2. Swedish

In Swedish, there are more noun phrase types than in English since a gender distinction (common vs. neuter) is added to the definiteness and number distinction. Furthermore, some common gender nouns are overtly marked for common singular (cf. *flick-a* 'girl-COMM.SG' vs. *flick-or* 'girl-COMM.PL', *gubb-e* 'old.man-COMM.SG' vs. *gubb-ar* 'old.man-COMM.PL'). However, since such words are in a minority, they can be neglected for the purposes of our analysis.

One more remark should be made about neuter nouns. Most of them have no overt ending in plural (*hus* 'house.SG' = *hus* 'house.PL'). There is a small number of neuters which do have a plural ending (*äpple* 'apple.SG' ≠ *äpple-n* 'apple-PL'), but they are not numerous and are also left out of the analysis.

Combining the possible feature values, we arrive at the total of eight possible types of the noun phrase (2 values of definiteness × 2 genders × 2 numbers). Some of these types are represented in Graph 6, and all of them are listed in Table 2.



Graph 6. The structure of the Swedish noun phrase.

On average, the Swedish noun phrase has 4.38 form vertices, 5 meaning vertices and 5.88 edges. The average vertex degree of such a graph would be 1.25.

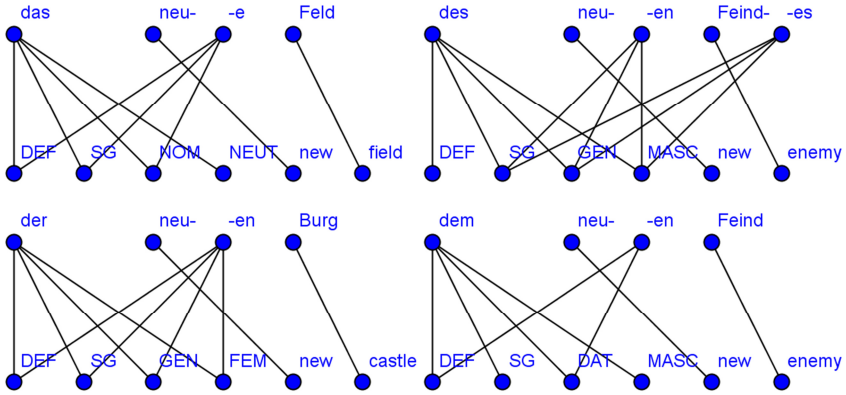
Table 2. The structure of the Swedish noun phrase.

	Definiteness	Number	Gender	Vertices (form)	Vertices (meaning)	Vertices (total)	Edges	Average vertex degree
en stor bil	INDEF	SG	COMM	3	5	8	5	1.25
den stora bilen	DEF	SG	COMM	5	5	10	9	1.80
ett stort hus	INDEF	SG	NEUT	4	5	9	8	1.78
det stora huset	DEF	SG	NEUT	5	5	10	9	1.80
stora bilar	INDEF	PL	COMM	4	5	9	5	1.11
de stora bilarna	DEF	PL	COMM	6	5	11	10	1.82
stora hus	INDEF	PL	NEUT	3	5	8	3	0.75
de stora husen	DEF	PL	NEUT	5	5	10	8	1.60

4.3. German

The German noun phrase has four grammatical features (definiteness, number, case and gender). Since genders are not distinguished in plural, there are only four possible combinations of number and gender (SG.MASC, SG.FEM, SG.NEUT and PL) and 32 feature combinations in total (= 2 values of definiteness × 4 combinations of number and gender × 4 cases). There are 13 types of relations between form and meaning which are represented in Table 3 (for those structures which occur with more than one feature combination, all the feature combinations are listed, but only one example is given). Some of these structures are also illustrated in Graph 7.

On average, the German noun phrase has 4.69 form vertices, 6 meaning vertices and 9.75 edges. The average vertex degree of such a graph would be 1.82.



Graph 7: The structure of the German noun phrase.

4.4. Noun phrases in three Germanic languages: A comparison

The comparison of values for English, Swedish, and German makes it possible to establish the following complexity hierarchy:

English < Swedish < German

Fortunately, this hierarchy is the same for all three components of complexity, i.e. conceptual complexity, formal complexity and correspondence complexity. Because of this, it is possible to say that the English noun phrase is less complex than its Swedish counterpart, and both of them are less complex than the German noun phrase. This conforms to the intuitive impression of the structure of these languages, which shows that the graph-theoretic measure is a good way to formalize the linguists' intuition.

5. Conclusions

A graph-theoretic approach can be useful for quantifying grammatical complexity. A method proposed in this paper distinguishes three components of grammatical complexity:

Table 3. The structure of the German noun phrase.

	Definiteness	Number	Case	Gender	Vertices (form)	Vertices (meaning)	Vertices (total)	Edges	Average vertex degree
das neue Feld	DEF def def	SG sg sg	NOM nom nom	MASC fem neut	4	6	10	9	1.8
des neuen Feindes	DEF def	SG sg	GEN gen	MASC neut	5	6	11	12	2.18
der neuen Burg	DEF def	SG sg	GEN acc	FEM neut	4	6	10	10	2.00
dem neuen Feind	DEF def def	SG sg sg	DAT dat dat	MASC fem neut	5	6	11	12	2.18
den neuen Feind	DEF def	SG sg	ACC acc	MASC fem	4	6	10	10	2.00
die neuen Feinde	DEF def def	PL pl pl	NOM gen acc		5	6	11	8	1.45
den neuen Feldern	DEF	PL	DAT		6	6	12	10	1.66
eine neue Burg	INDEF indef indef indef indef	SG sg sg sg sg	NOM nom nom gen acc	MASC fem neut gen neut	5	6	11	12	2.18
eines neuen Feldes	INDEF indef	SG sg	GEN gen	MASC neut	6	6	12	14	2.33
einer neuen Burg	INDEF indef indef	SG sg sg	DAT dat dat	MASC fem neut	5	6	11	10	1.82
einen neuen Feund	INDEF indef	SG sg	ACC acc	MASC fem	5	6	11	11	2.00
neue Felder	INDEF indef indef	PL pl pl	NOM gen acc		4	6	10	6	1.20
neuen Feldern	INDEF	PL	DAT		5	6	11	7	1.27

– conceptual complexity
– formal complexity
– form-meaning correspondence complexity

It is problematic to use this method for comparing the overall complexity of unrelated languages, but it can turn out useful for comparing similar grammatical phenomena cross-linguistically.

References

Atkinson, Q.D. 2011. "Phonemic diversity supports a serial founder effect model of language expansion from Africa". *Science* 332 (6027). 346–349.
Bonchev, D. and G.A. Buck. 2005. "Quantitative measures of network complexity". In: Bonchev, D. and D.H. Rouvray (eds.), *Complexity in chemistry, biology, and ecology.* New York: Springer. 191–236.
Chomsky, N. 2006. *Language and mind*. Cambridge: Cambridge University Press.
McWhorter, J.H. 2001. "The world's simplest grammars are creole grammars". *Linguistic Typology* 5(2–3). 125–166.
Mel'čuk, I.A. and N.V. Pertsov. 1987. *Surface syntax of English: A formal model within the meaning-text framework*. Amsterdam/Philadelphia: John Benjamins.
Nichols, J. 1992. *Linguistics diversity in space and time*. Chicago: University of Chicago Press.
Nichols, J., J. Barnes and D.A. Peterson. 2006. "The robust bell curve of morphological complexity". *Linguistic Typology* 10(1). 96–106.
Sampson, G., D. Gil and P. Trudgill (eds.). 2009. *Language complexity as an evolving variable*. Oxford: Oxford University Press
Saussure, F. de. 1916. *Cours de linguistique générale*. Paris: Éditions Payot & Rivages.
Shosted, R. 2006. "Correlating complexity: A typological approach". *Linguistic Typology* 10(1). 1–40.
Trudgill, P. 2012. "On the sociolinguistic typology of linguistic complexity loss". In: Seifart, F. et al. (eds.), *Potentials of language documentation: Methods, analyses, and utilization.* (Language Documentation & Conservation Special Publication No. 3.) 90–95.

Address for correspondence
Alexander Piperski
Center of Sociolinguistics
School of Humanities
Russian Academy of National Economy
82/84 Prospekt Vernadskogo
119571 Moscow
Russia
apiperski@gmail.com