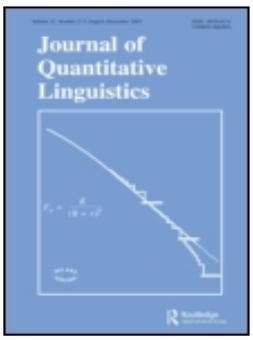

Journal of Quantitative Linguistics

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: http://www.tandfonline.com/loi/njql20

Automatic Language Classification by means of Syntactic Dependency Networks

Olga Abramov & Alexander Mehler

To cite this article: Olga Abramov & Alexander Mehler (2011) Automatic Language Classification by means of Syntactic Dependency Networks, Journal of Quantitative Linguistics, 18:4, 291-336, DOI: 10.1080/09296174.2011.608602

To link to this article: https://doi.org/10.1080/09296174.2011.608602

Journal of Quantitative Linguistics
2011, Volume 18, Number 4, pp. 291–336
http://dx.doi.org/10.1080/09296174.2011.608602

Routledge
Taylor & Francis Group

Automatic Language Classification by means of Syntactic Dependency Networks*

Olga Abramov[1] and Alexander Mehler[2]
[1]Bielefeld University, Germany; [2]Goethe-University Frankfurt, Germany

ABSTRACT

This article presents an approach to automatic language classification by means of linguistic networks. Networks of 11 languages were constructed from dependency treebanks, and the topology of these networks serves as input to the classification algorithm. The results match the genealogical similarities of these languages. In addition, we test two alternative approaches to automatic language classification – one based on *n*-grams and the other on quantitative typological indices. All three methods show good results in identifying genealogical groups. Beyond genetic similarities, network features (and feature combinations) offer a new source of typological information about languages. This information can contribute to a better understanding of the interplay of single linguistic phenomena observed in language.

1. INTRODUCTION

Research on language classifications has passed through two main stages of development. Early typologists tried to classify language as a whole (holistic stage). Classifications worked out in this stage were of great importance up until now, though they are insufficient for several reasons. The main criticism to holistic (predominantly morphological) typologies is that they focus only on some aspects of language and disregard the others. In fact, it turned out to be rather challenging to integrate the whole language into a single typology by considering all linguistic levels (morphology, syntax, etc.). This is the reason why the later typologists

*Address correspondence to: Olga Abramov, Bielefeld University, Faculty of Linguistics and Literature, Universitaetsstr. 25, 33615 Bielefeld.
E-mail: olga.abramov@uni-bielefeld.de; E-mail: mehler@em.uni-frankfurt.de

gave up the goal of establishing a holistic classification and focused instead on sub-parts of language (partial stage). Most of the research in contemporary typology adapts the partial approach (Masayoshi & Bynon, 1995). However, many questions concerning the macro level of language (i.e., the interplay of its parts) remain unanswered. The demand to understand the macro level or the general form of language is formulated by Sapir (1921) as follows:

> It must be obvious to anyone who has thought about the question [of the general form of the language] or who has felt something of the spirit of a foreign language that there is such a thing as a basic plan, a certain cut, to each language. This type or plan or structural "genius" of the language is something much more fundamental, much more pervasive, than any single feature of it that we can mention, nor can we gain an adequate idea of its nature by a mere recital of the sundry facts that make up the grammar of the language.

To interpret the macro level of language in the sense of 19th century linguists means to refer to its inner character (see Masayoshi & Bynon, 1995). Structuralists also speak of a "general structure" or "scheme" of a language. The availability of such a scheme, plan or genius should help to bridge the gap between the macro level (holistic) and micro level (partial) studies in typology. Knowing the "general structure" in advance, we will easily be able to predict or verify the findings made by micro level partial methods (Altmann & Lehfeldt, 1973).

But how can we get at this "general structure" of languages? This could be done, for example, by including all characteristics observed in a language into a single model. However, it would be rather a hard task or even impossible since not all features are known for every language in the world. Alternatively, one could try to extract an imprint of a language by inducing the regularities from a large amount of natural language data. This is done in the present article. We test three different approaches to account for this task. One of them is sketched in the following.

More specifically, we use dependency treebanks of 11 languages as a resource to extract language networks (Ferrer i Cancho et al., 2004). The main assumption is that the topology of these networks reflects similarities among languages. We make a classification of language networks based on their topology to test this assumption.

The whole procedure (summarized in Figure 1) represents an application of quantitative network analysis (QNA) (Mehler, 2008) to syntactic networks that are derived from dependency treebanks. In this scenario, vertices of the networks represent word forms of the treebank, and edges represent dependency relations between them. First a treebank is parsed sentence by sentence, and then subsequently nodes and edges are added to the network (I–III). When a word form is already in the network an additional edge will be attached to it. After parsing the treebank we get a graph representation of each language. Next, QNA is applied to characterize the topologies of these graphs. Twenty-one coefficients of network theory that measure, for example, the amount of clustering, path distances, centrality etc. are calculated for each language graph (IV). Coefficient values form a vector representing a language. These vectors are clustered in order to see whether languages can be distinguished by means of their network topology (V).

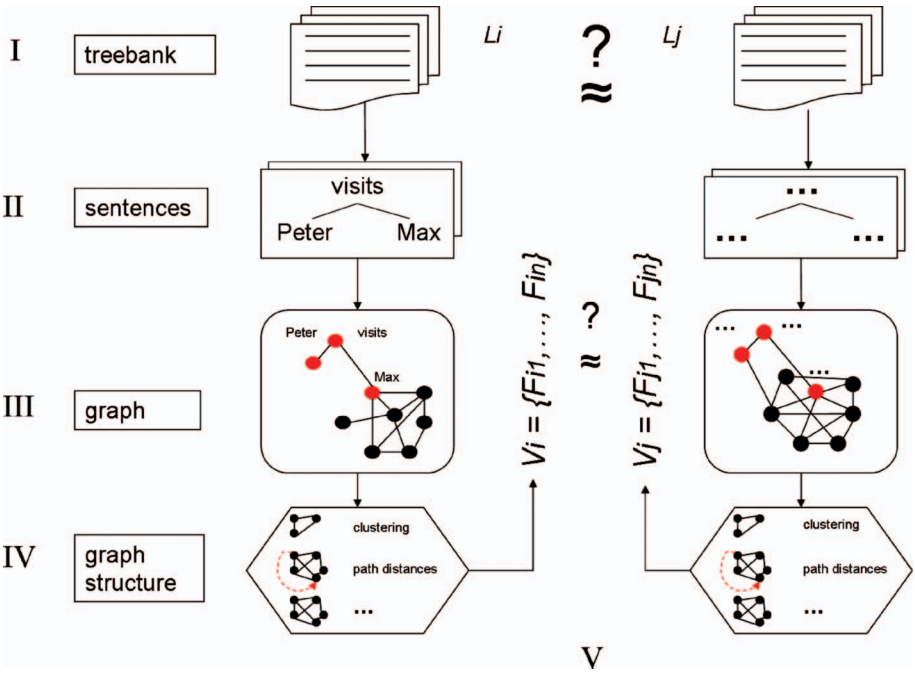

Fig. 1. Applying QNA to GSDNs: short description of the approach.

We test the resulting classification against the genealogical distribution of these languages according to language families. The question here is, whether the network topology bears information about the genetic origin of the languages under consideration. Further we explore the typological gain of single network indices applied here.

The paper is organized as follows: in Section 2 we relate our work to research on automatic language classification. In Section 3 we describe the procedure of generating networks from treebanks. Section 4 presents a discussion of how the network characteristics of QNA are related to the properties of language. Two alternative approaches to automatic language classifications, one based on *n*-grams (Section 5.1) and the other based on quantitative typological indices (Section 5.2) from Altmann and Lehfeldt (1973) are presented in Section 5. The experimental procedure is described in Section 6. The results are presented and discussed in Section 7. Finally, in Section 8 we give a conclusion.

2. RELATED WORK

Ferrer i Cancho et al. (2004) were the first to study the properties of syntactic networks based on data from dependency treebanks of three languages. They could show that the topology of these networks is not random. Rather these networks all fit the small world model (SWM) of Watts and Strogatz (1998). The work of Ferrer i Cancho et al. (2004) helped to shed light on the relation between the degree distributions of syntactic networks and the Zipfian distribution of word frequencies.

Liu (2008) looked at the topology of language networks of a single language consisting of two different text types. He confirms the fit to the small-world model for both text types (networks) which he relates to the Zipfian law of natural language. He also observed small differences in the values of the coefficients for treebanks representing different text types. Liu et al. (2010) studied the question whether local differences in syntactic annotation scheme (different representation of co-ordinating constructions) influences the global structure of dependency networks. We found out that global properties of being a small-world and scale-freeness are not significantly influenced by local syntactic changes. However, other network properties like centrality are more sensitive to local changes of particular syntactic constructions. Liu et al. (2010)

argued that we need to find other global statistical properties, which better reflect local changes in the network. The present paper examines 21 different network indices with respect to their potential in distinguishing language networks.

Minkov and Cohen (2008) performed a graph walk based on named entity extraction (or ''named entity co-ordinate term extraction'') using directed weighted labelled global syntactic dependency networks (GSDN).[1] They have shown that sequences of labelled dependency paths bear information about word similarities allowing the detection of city and person names.

Mehler (2008) introduced quantitative network analysis (QNA) as an approach to classify complex networks in terms of their topology. QNA combines complex network theory with unsupervised machine learning to model classifiers of networks that explore only their structure. This is exemplified by classifying social and linguistic networks – the latter derived on the textual and lexical level – where all these networks are derived from special wikis. The classification shows not only that these networks can be distinguished ontologically, but also functionally in terms of communication areas.

Mehler et al. (2010) further applied QNA to classify languages genealogically. As a data source, Mehler et al. (2010) utilized the category system of Wikipedia that is available for many languages in the world. They have shown that languages can be classified into language families by exploring the topology of the Wikipedia category systems of the corresponding languages to be classified.

In the present article, we apply QNA to a data driven language classification by means of syntactic networks. We aim to find out whether the network structure induced from dependency treebanks provides any information about the relatedness of languages in analogy to the classification presented in Mehler et al. (2010). That is, while Mehler et al. (2010) used social ontologies and, thus, semantic networks to classify languages, we will now use syntactic networks for the same task.

Further studies related to language classifications are subsumed to the field of language tree reconstructions. This research field attracts researchers from different sciences: physicians, biologists and linguists (especially historical-comparative linguists). The leading assumption here

[1]This notion goes back to Ferrer i Cancho et al. (2004) and will be explained in more detail below.

is that all languages originate from a single proto-language which had been split apart in smaller pieces or language families.

> If two or more languages share a feature which is unlikely to have occurred spontaneously in each of them, this feature must have arisen once only, when these languages were one and the same.
>
> (Anttila, 1972)

Different methods were proposed to recover genetic relationships of languages, which are mostly based on lexicostatistics (e.g. Swadesh, 1952; Batagelj et al., 1992; Bryant et al., 2005). That is, the number of common basic words (cognates[2]) determines the degree of distance between languages. The more words two languages share, the closer this genetic relationship. Genetic relationships are represented in terms of trees going up to the proto-language. In a nutshell, the lexicostatistical approach determines the proportion of the most basic vocabulary shared by two languages (Warnow et al., 1996). The validity of this method is widely questioned since it disregards many factors in language. Alternative approaches include additional information like phonology, morphology etc. to calculate genetic trees.

Warnow et al. (1996), for example, proposed a combined approach to language tree reconstruction using cognates, morphological and phonological features to reconstruct the tree.

Batagelj et al. (1992) enhanced the cognate-based method by providing simple distance metrics to measure the similarity between cognates. For example, they calculated the number of steps (insertions, deletions, etc.) needed to transform one form of a cognate from one language into another form of the same cognate from the other language. They clustered 65 languages based on these counts and achieved good classification results comparable with results achieved applying the historical explorative reconstruction methods. A more elaborated approach using normalized edit distances and graph walks is proposed by Blanchard et al. (2009) who extend the sample of Swedish including Austronesian languages.

[2]Cognates are pairs of words from different languages that originate from the same ancestor language. The common origin is determined by regular phonetic change from one language to another and by related meaning of the two words. Borrowed words are not cognates (Kruskal et al., 1992).

Holman et al. (2008) presented an approach to automatically classifying languages based on word-lists, which are not restricted to cognates. They developed several techniques to identify the most stable words that improve the classification. That way, they reduced the word space from 100 to 40 word features. They also reported that combining word-list based features with typological features from the world atlas of language structures (WALS) (Haspelmath et al., 2005) can improve the outcome of the classification. Other algorithms to automatically reconstruct language relationships, which are based to a large extent on phonetics, are reviewed in Kondrak (2002).

Daumé III (2009) have shown that methods in language reconstruction can be enhanced by including areal information. Daumé III (2009) identified those typological features from WALS which are shared among *arealy* related languages and used this information to improve the reconstruction of genetic trees.

The above approaches deal with selected features from several levels of linguistic representation. However, a general classification of languages according to Altmann and Lehfeldt (1973) is one that captures as many levels as possible. Since it is hardly possible to obtain all features from all linguistic levels, another possibility is to collect as many features from different linguistic levels as possible and to classify the languages hoping that the selected features recover the genetic relationships between them. Then, proceeding top-down, we can look at single features in this set, and ask for their contribution to the overall result.

The present article accounts for this task presenting three approaches (one network based, one *n*-gram based, and one quantitative typological) to language classification and analyses single features with respect to their discriminative potential.

3. FROM TREEBANKS TO SYNTACTIC DEPENDENCY NETWORKS

For the purpose of the present study, we used 11 treebanks annotated with syntactic structure according to dependency grammar (Tesnière, 1959). We selected the dependency grammar (DG in the following) since it allows for cross-language comparisons irrespective of the word order of a particular language. For every language we extracted a global syntactic dependency network (GSDN) (Section 3.2) as proposed by Ferrer i

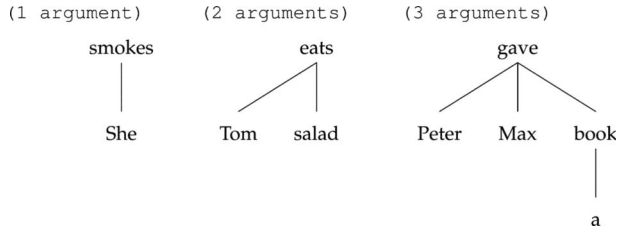

Fig. 2. The DG representation of sentences based on the *valency* of the verb.

Cancho et al. (2004) and compared languages according to these networks. The following section describes the treebanks and the extraction procedure.

3.1 Data Source: Dependency Treebanks

The treebanks used in this study are listed in Table 1. In this collection 11 languages are available. As can be seen from Table 1, the treebanks differ with respect to the dependency grammar used for annotation. In some cases punctuation marks are included as parts of the dependency trees, in others not. The representation formats used to represent the dependency trees are also different which makes the access to the treebanks within a single interface rather challenging. There are seven different formats that had to be transformed into a unique representation.

The treebanks were transformed into a graph based XML data model called eGXL[3] that was designed to represent graph and tree relations. Of course, the heterogeneity of the data makes comparisons of languages a hard task. On the other hand, quantitative methods are less prone to small differences, than direct comparisons of, for example, single sentences. However, systematic divergence might also confuse the quantitative result. Some treebanks, for example, include punctuation marks in the dependency trees, others do not, which might result in biases when comparing the networks. For this reason, we normalized the networks excluding punctuation marks (like commas, full-stops and other special characters). That is, all dependency links containing a punctuation mark were removed from the network. The

[3]This model is based on the Graph eXchange Language GXL (Holt et al., 2006). Pustylnikov and Mehler (2008); Pustylnikov et al. (2008) adapted this format in order to model syntactic trees. See the TreebankWiki (http://ariadne.coli.uni-bielefeld.de/wikis/treebankwiki/) for all details on the conversion process.

Table 1. 11 Dependency Treebanks.

Treebank	Language	$\|V\|$	$\|E\|$	Punctuation included	Format used	Reference
Alpino Treebank v. 1.2	Dutch	28,475	102,184	yes	CoNLL	van der Beek et al. (2002)
Danish Dependency Treebank v. 1.0	Danish	19,133	50,858	yes	TIGER-XML	Kromann (2003)
Dependency Grammar Annotator	Romanian	8,867	23,901	no	simple XML	Hristea and Popescu (2003)
Russian National Corpus	Russian	58,283	177,942	no	RNC-XML	Boguslavsky et al. (2002)
Slovene Treebank v. 0.4	Slovene	8342	20,453	yes	TEI	Džeroski et al. (2006)
Talkbanken05 v. 1.1	Swedish	25,097	126,526	yes	TIGER-XML	Nivre et al. (2006)
Turin University Treebank v. 0.1	Italian	7984	24,269	no	TUT format	Bosco et al. (2000)
CESS – Catalan Dependency Treebank	Catalan	38,882	215,308	yes	CoNNL	Civit et al. (2004)
Cast3LB – Spanish Dependency Treebank	Spanish	17,101	56,911	yes	CoNNL	Civit and Martí (2005)
Prague Dependency Treebank 2.0	Czech	146,504	696,379	yes	PDT	Hajič (1998)
BulTreeBank	Bulgarian	32,421	95,698	yes	CoNLL	Osenova and Simov (2004)

procedure of creating the dependency networks is described in the following section.

3.2 Extracting Global Syntactic Dependency Networks from Treebanks

The notion of GSDN goes back to (Ferrer i Cancho et al., 2004) who defined a GSDN as "a set of n words $V=\{s_i\}$ $(i=1,\ldots,n)$ and an adjacency matrix $A=\{a_{ij}\}$. If a link goes from the modifier s_i to the head s_i then $a_{ij}=1$ (and $a_{ij}=0$ otherwise)."[4] In this case, links go from the modifier to the head; of course, this can be changed the other way round. According to this definition, GSDNs are simple directed graphs, however, complex network theoretic measures applied in this article to characterize GSDNs treat them as undirected.

We use word forms or "types" as vertices of the network, since not all treebanks are lemmatized. Two vertices (i.e. types) of a GSDN are linked if they appear at least once in a modifier-head relation in the treebank. The procedure of creating a GSDN is illustrated in Figure 3.

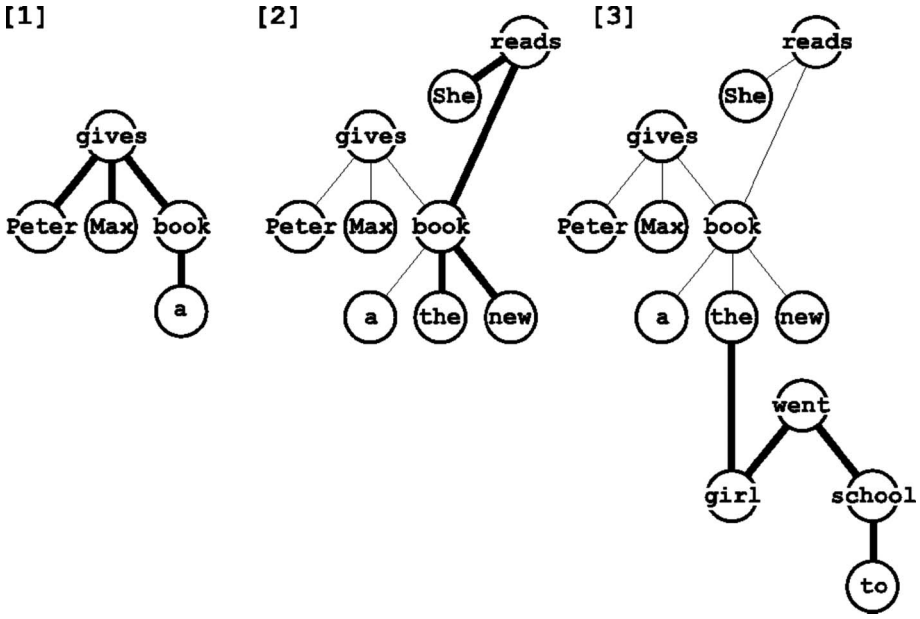

Fig. 3. The figure taken from (Mehler et al., 2010) exemplifies how a GSDN is created after parsing the 1, 2, 3 sentences.

[4]Ferrer i Cancho et al. (2004, p. 2).

The treebank is parsed sentence by sentence and new words are added to the network. Words are linked according to the dependency relations they constitute. When a word is already present in the network (e.g. *book* in Figure 3), more links are added to it. Finally, we get a network containing all words and all dependency relations of a particular treebank. The degree of a word gives the number of different dependency relations to other words.[5] As mentioned in the previous section, punctuation was not included in the network.

4. A RANGE OF TYPOLOGICAL NETWORK INDICES AS INPUT TO QNA

In this section, we look more closely at some selected network features and try to relate them to the properties of language. Starting from a subset of topological indices studied in Mehler (2008) – see Table 2 – we calculate each of these indices for the GSDNs in our network corpus to get input to QNA. Before discussing the coefficients, here are some basic definitions we operate with. A graph $G = (V, E)$ is a GSDN (see Section 3.2). A degree of a vertex v_i is denoted with $d(v_i)$. The number $\delta(v_i)$ of triangles, to which the vertex v_i is connected, is defined as (Schank & Wagner, 2005):

$$\delta(v_i) = |\{\{u, w\} \in E\{v_i, u\} \in E\{v_i, w) \in E\}|. \quad (1)$$

Finally, the number of candidate triples (a triple is a path of length two centred in a certain vertex) that are rooted in v_i, is denoted by $\tau(v_i)$:

$$\tau(v_i) = \binom{d(v_i)}{2} \quad (2)$$

4.1 Clustering Coefficients

We compute two variants of the clustering coefficient: $C_1(G)$ (Watts & Strogatz, 1998) and $C_2(G)$ (Bollobás & Riordan, 2003) (See Features F_1 and F_2 in Table 2). For each graph we compute the clustering coefficient

[5]Note that weights of edges are not considered by this model; that is, if two words occur more than once in a modifier-head relation it does not result in an increase of degrees of these words.

Table 2. The list of composite features (taken from Mehler, 2008) considered in the present study. They fall into three groups, that is, features of complex network theory (1), social network analysis (2) and hypertext structure analysis (3) (as indicated in the last column).

Index	Feature	Short description	Area
F_1	$C_{ws}(G)$	The cluster coefficient of G	1
F_2	$C_{br}(G)$	The cluster coefficient of G	1
F_3	$L(G)$	The average geodesic distance of G	1
F_4	$D(G)$	The diameter of G	1
F_5	$r(G)$	The degree of assortative mixing of G	1
F_6	$\varepsilon(G)$	The average degree of G	1
F_7	$lcc(G)$	The fraction of the largest connected component of G	1
F_8	$\gamma(G)$	The γ of the power law of type $Ck^{-}\gamma$ which best fits to the degree distribution of G	1
F_9	$^{-}R^2\gamma(G)$	The corresponding adjusted coefficient of determination	1
F_{10}	$\gamma S(G)$	The γ of $Cn^{-}\gamma$ which best fits to the size distribution of connected components of G	1
F_{11}	$^{-}R^2\gamma S(G)$	The corresponding adjusted coefficient of determination	1
F_{12}	$\gamma knn(k)(G)$	The γ of the power law of type $Ck^{-}\gamma$ which best fits to the distribution of k_{nn} values of G	1
F_{13}	$^{-}S^2\gamma knn(k)(G)$	The corresponding adjusted coefficient of determination	1
F_{14}	$\gamma C(k)(G)$	The γ of the power law of type $Ck^{-}\gamma$ which best fits to the distribution of $C(k)$ values of G	1
F_{15}	$^{-}R^2\gamma knn(k)(G)$	The corresponding adjusted coefficient of determination	1
F_{16}	$GC(G)$	The graph centrality of G	2
F_{17}	$CC(G)$	The standard deviation of the closeness centrality of G	2
F_{18}	$DC(G)$	The degree centrality of G	2
F_{19}	$Cp(G)$	The compactness of G	3
F_{20}	$Ch(G)$	The cohesion of G	3
F_{21}	C_A	The relative graph connectivity (Mehler et al., 2011)	3

by averaging the cluster values of its vertices. It is the probability of two vertices w, u being linked to one another when they are both linked to a common neighbour v_i. The clustering coefficient of a vertex v_i from (Watts & Strogatz, 1998) is computed as $c(v_i) = \delta(v_i)/\tau(v_i)$, that is the proportion of triangles with respect to triples. The clustering coefficient of the graph is an average value calculated as $C_1(G) = \frac{1}{|V|}\sum_{i=1}^{|V|} c(v_i)$.

In our case we deal with graphs consisting, for instance, of verbs linked to nouns (see Figure 4), nouns linked to articles, adjectives, etc. Edges occur mostly among different word forms: verbs-nouns, nouns-adjectives, etc. That means the probability of triangle relations going, e.g. from *Peter* to *gave*, from *gave* to *book* and back from *book* to *Peter* is very low (Figure 4). This in turn results in a low clustering coefficient for dependency networks in general. Due to dependency syntax, nouns linked to nouns or verbs to verbs should not occur in simple sentences. However, in the case of sentences like *I know Peter read the book* (Figure 4) ''know'' and ''read'' are linked, and since in another sentence ''know'' and ''Peter'' might be linked too, the three words ''know'', ''read'' and ''Peter'' will form a triangle. The above example explains how triangles can nevertheless appear in language networks. That means, we can expect triangles to be present to some extent in all languages.

The interesting question in this context is whether we can distinguish languages based on the amount of triangles, that is, on the value of the clustering coefficient. Languages like Swedish, for example, are more analytic than Russian; thus Russian has more word forms on average representing different inflectional cases than Swedish. When we transfer this observation onto networks we can expect Russian to have a lower clustering value than Swedish, since as in the example above the word ''Peter'' for instance would be written differently depending on the

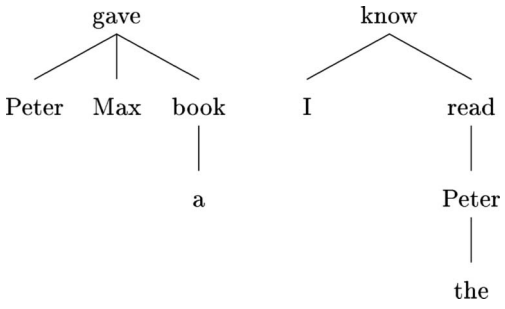

Fig. 4. Two example sentences in dependency notation.

inflectional case (e.g. nominative vs. accusative) which minimizes the probability of a triangle containing, e.g. ''know'' and ''read''. For an analytic language such a connection is more probable due to a sparse morphological variation. Thus, an example like the above relation could frequently occur in English, Swedish etc. but not that frequently in Russian.

The above observations are confirmed by the average cluster values of GSDNs presented in Figure 9 (see Appendix). For both $C_1(G)$ and $C_2(G)$[6], Russian and Czech, for instance, have smaller clustering values than Swedish and Danish. Although, Russian and Swedish GSDNs have comparable order, Russian has the smallest and Swedish the largest cluster values. Clustering plays an important role in classifying languages; when we look at the best-of-feature combinations in Table 14 we see that either C_1 or C_2 as well as the fit of the cluster values to the power law distribution (Features $\gamma_{C(k)}$ and $R^2_{C(k)}$) are informative in distinguishing language families.

4.2 Path Distances

The average geodesic distance $L(G)$ (F_3 in Table 2) of a graph G constitutes together with the clustering coefficient (C_1 or C_2) the small world model (SWM) of Watts and Strogatz. The SWM proved to be an appropriate model for many types of networks, among others: biological, technical and social ones. The clustering coefficient in the SWM is assumed to be high and the geodesic distances to be low on average in comparison to random graphs with the same number of vertices. In the case of dependency networks we can expect distances to be short in general since, e.g. content words or nouns are linked to function words whose number (in natural language) is of limited size. This fact assures the occurrence of short paths in a dependency network. Here again, differences are expected among languages resulting in longer paths for morphologically richer languages like Russian than for analytic languages. The reason is that analytic languages have more function words (like, for example, prepositions) that connect to many different word forms. These kind of vertices serve as short cuts reducing the paths in the network. When we look at Figure 9 (see Appendix), Russian and Slovene have longer average distance values ($L(G)$) than Catalan and Swedish.

[6] C_2 is a variant of C_1 that weights the single $c(v_i)$s by their vertex degrees.

4.2.1 Average Degree

The average degree $\varepsilon(G) = edges/vertices$ of a graph G (F_6 in Table 2) is in principal a very informative feature representing the proportion of edges with respect to the number of vertices. Again, thinking of the synthesis status of a language we can expect an analytic language to have more edges (and proportionally fewer vertices) since the same morphological forms are used more frequently. Thus, the average degree of an *analytic* graph should be higher than the average degree of a *synthetic* graph.

As in the case of $L(G)$, this expectation is confirmed by high values of ε for Catalan and Swedish ($L(G) \sim 5$) in comparison to languages like Slovene ($L(G) \sim 2$) (see Figure 9).

4.3 Connectivity Distribution

This feature is directly related to Zipf's law of word frequencies (Zipf, 1932). Since we study degree distributions in analogy to word frequency distributions, Zipf's law can be directly applied to degrees of our language networks. That is, the probability $P(k)$ of a vertex to have the degree k follows a power law $P(k) \sim k^{-\gamma}$ (Barabási & Albert, 1999). The central question that we tackle by means of this coefficient is: How are frequencies of dependency relations distributed? More specifically, we can ask: Which word classes have the most (least, etc.) dependency relations in a GSDN of a particular language? Such questions can be answered by looking at connectivity distributions.

We check whether our networks fit this model by looking at the exponent $\gamma(G)$ and the adjusted coefficient of determination $R^2_\gamma(G)$ that evaluates the goodness of the fit (Features F8 and F_9 in Table 2). As predicted by Zipf's Law, all values of $\gamma(G)$ are negative, and $R^2_\gamma(G)$ is close to 1. However, the values of these features vary among languages reflecting different frequencies of dependency relations that form the shape of the distribution. For example, in some languages nouns occupy the higher ranks, while in other languages these ranks are occupied by verbs. It is interesting to inspect whether these differences are law-like so that typologically related languages result in similar GSDNs.

4.4 Assortativity

Assortativity is a property of networks that describes connectivity preferences among its vertices. The question thereby is whether vertices of degree k connect to vertices of similar degrees (assortative mixing) or not (disassortative mixing). In terms of GSDNs we ask: Do vertices

(words) of degree k (i.e. k different dependency relations) connect to words of the same or different degree? The correlation coefficient of Newman & Park (2003) (feature F_5 in Table 2) gives the overall tendency for the network – assortative or not. The connectivity correlation (Pastor-Satorras et al., 2001) (features F_{12} and F_{13} in Table 2) allows us to explore instances of which word classes tend to connect to instances of other word classes.

Assume, for example, that nouns and verbs have the same degrees. Then, we can ask whether they are connected to each other or not. If they are, it indicates assortative, if not disassortative mixing. Assortative mixing was observed for, e.g. social networks (Newman & Park, 2003). Disassortative mixing was shown for many real networks (Pastor-Satorras et al., 2001) as well as for Wiki and document networks (Mehler, 2008). All the six GSDNs analysed in Ferrer i Cancho et al. (2007) exhibit disassortative mixing. In this paper we confirm this finding for 11 GSDNs.

A more interesting question in our context is whether assortativity allows to separate genealogically different languages. The correlation coefficient of (Newman & Park, 2003) can be used to compute the assortativity value $r(G)$ of a network considering all possible degrees. A positive value of $r(G)$ indicates assortative mixing of the graph, a negative value the opposite (disassortativity). We calculate $r(G)$ for GSDNs in order to examine whether related languages share these preferences. Apparently, no striking differences among the values of $r(G)$ for related languages can be identified.

Connectivity correlation k_{nn} is another index of assortativity. It is computed as the average degree of the nearest neighbours of vertices with degree k. The value of k_{nn} grows when the network exhibits assortative mixing and shrinks when the network has disassortative mixing. k_{nn} is a better indicator of assortativity than $r(G)$ since the complete distribution of degrees of the nearest neighbours is considered. In the literature (Pastor-Satorras et al., 2001) many networks can be fitted to the distribution of k_{nn} $(k) \sim k^{-\beta}$. Comparing the goodness of fit (i.e. the *adjusted coefficient of determination*) to the distribution of k_{nn} values (Feature F_{13}, see Table 2) with $R^2_{\gamma}(G)$, we observe in general bad results. The fits are better for larger (CZE), and less good for small networks (SLV, RUM, ITA). In general, GSDNs have very similar distributions of nearest neighbor degrees (see Feature F_{12} in Table 2) independent of language family the GSDN is attributed to. This fact was also observed by Ferrer i Cancho et al. (2007). That is, we expect connectivity correlation to be less informative for inter-GSDN comparisons but

allows presumably to distinguish GSDNs from other kinds of (linguistic) networks.

4.5 Centrality

We calculated various centrality measures: the degree centrality (DC) (Feldman & Sanger, 2007), graph centrality (GC) (Hage & Harary, 1995) and (standardized) closeness centrality (CC) (Wasserman & Faust, 1999). All indices rank the GSDNs within the interval [0, 1] with 1 indicating high, and 0 low overall centrality. Centrality measures are vertex related. To get an index of centrality for a graph, we aggregate them by means of several functions allowing it achieve a single value.[7]

An interesting question concerning GSDNs and centrality is the following: Are there words that can have a dependency relation with almost every other word? It is rather unrealistic to assume only one such universally attachable word to appear in a language (i.e. in that case the centrality of a graph would be nearby 1). Further, it is unrealistic to assume that all words have the same probability of being connected to all other words of the same GSDN. This would also contradict the Zipfian distribution of words in language. Thus, realistic centrality values of languages lie within a particular interval. However, how large is the variation in this interval? And does this variation tell us something about the typological properties of a particular language? Keeping these questions in mind, in this section, we look more closely at the centrality of GSDNs exemplified by degree and closeness centrality. We have selected these measures, since they represent two classes of centrality measures – degree based (i.e. DC) and distance based (i.e. GC, CC).

4.5.1 Degree Centrality (DC)

The degree centrality (DC) (Feldman & Sanger, 2007) relates vertex degrees to each other. If there are many vertices of a low degree and one vertex of a high degree connected to all the others, the centrality of the graph will be high (DC $\sim$ 1).[8]

$$\mathrm{DC}(G) = \frac{\sum_{v \in V} d_{\max}(V) - d(v)}{(|V| - 1)(|V| - 2)} \in [0, 1] \tag{3}$$

[7]Note, that all indices are computed only for the largest connected component (LCC). This is, of course, an abstraction and some information might become lost.

[8]The DC is 1 for a "star graph" (i.e. all vertices have degree $d = 1$, and one vertex has degree $d = |V| - 1$).

In the above equation each vertex is compared to $d_{\max}(V)$, i.e. the vertex with the highest degree in the graph. The fewer the vertices equal to $d_{\max}(V)$, the higher is the DC. The DC value of 1 corresponds to a graph of a form like a "star graph" with one vertex of the maximal degree. A graph has a DC value of 0 if each vertex has the same degree.

The first thing that becomes apparent when we look at the DC values in Figure 9 (see Appendix) is that all GSDNs have DC values not higher than 0.3 and not less than 0.1. That is, no more than 30% of the words are central, and yet, at least 10% of central words are needed to form a GSDN. This observation is in line with the power law distribution of vertex degrees (cf. γ), whereby few words are connected to almost all other words, and many words are sparsely connected. With respect to vertex degrees this means that Danish is more centralized than Spanish (i.e. there are more central words in Spanish than in Danish). In our classification task, DC is one of the most informative features appearing in 18 (of 20) best-of combinations (see Table 14).

4.5.2 Closeness Centrality (CC)

(Standardized) Closeness Centrality $CC(v)$ (Wasserman & Faust, 1999) is another vertex related index of centrality, that is a function of geodesic distances rather than the degree of a vertex. Mehler (2008) proposes to compute the $CC(v)$ as follows:

$$CC(v) = \frac{|V| - 1}{\sum_{w \in V} gd(v, w)} \in [0, 1] \tag{4}$$

The closeness centrality of a graph $CC(G)$ is then computed as follows:

$$CC(G) = \left\{ \begin{array}{ll} \frac{\sum_{v \in V} \overset{\wedge}{CC}(V) - \overset{\wedge}{CC}}{|V|-1} : & \min_{v \in V} CC(v) < 1 \\ 0 : & \min_{v \in V} CC(v) = 1 \end{array} \right\} \in [0, 1] \tag{5}$$

with $\underset{\max}{\overset{\wedge}{CC}}(V) = \max_{v \in V} \overset{\wedge}{CC}(v)$, $\overset{\wedge}{CC}(v) = 1 - \frac{1 - CC(v)}{1 - \min_{v \in V} CC(v)}$, and $|V| > 1$. In case of $\min_{v \in V} CC(v) = 1$ all vertices have the same $CC(v)$, and thus, there are no central vertices. Otherwise, the deviation of $\overset{\wedge}{CC}(v)$ from the maximum $\underset{\max}{\overset{\wedge}{CC}}(V)$ is summed up and normalized by $|V| - 1$. The more vertices with the minimal sum of geodesic distances to all other the

smaller is the $CC(G)$ value. Conversely, if there is only one vertex with a short distance to all the other, and others are maximally distant to each other, than $CC(G) = 1$.

With respect to dependency networks high CC of a word (type) means that this word is easily reached from other words (types) in the network. Consequently, a word with a high CC occurs in many different dependency relations serving as a mediator in a dependency sentence. High closeness centrality words are with a high probability function words. The overall $CC(G)$ of the graph indicates how closely connected are the words (types) in a GSDN.

4.5.3 Discussion on Centrality

Though almost all three centrality measures, used in combination with other features, turned out to be informative for classification, when we look at the single values of a particular centrality index, there are apparently no striking differences between the three language families. That is, centrality does not reflect genealogical relationships. Nevertheless, there might by typological differences characterizing individual languages, which are captured by centrality. Typological interpretations of the centrality values are certainly possible at this point. However, similarity assumptions about languages should be made with caution since factors like different dependency theories, punctuation deleted, the size of the network, etc., could also influence the centrality index to some extent and bias the result.

What holds for all indices is the fact that centrality does not exceed the interval between 10% and 30% for all GSDNs. The last observation is in line with the skewed distribution of vertex degrees, whereby each language has a small number of highly connecting words (e.g. function words) and a large tail of low frequency words.

4.6 Compactness

Compactness (Cp) is a coefficient from classical hypertext theory introduced by Botafogo et al. (1992) that measures the interconnectedness among the vertices in a network. High compactness ($Cp = 1$) means that each vertex is connected to all other vertices in the graph resulting in a completely connected graph. A fully disconnected graph, in turn, has a compactness of 0. The benefit of this measure, as stated by Botafogo et al. (1992), is its independence from the size of the network, allowing it to compare networks of different or equal size structurally.

The central question associated with compactness is: How interrelated are the vertices of the network? Centrality measures can be computed only for connected components, whereas *Cp* integrates also disconnected parts of the graph. Thus, we expect additional information about the overall structure of GSDNs by considering *Cp*.

Compactness of Botafogo et al. (1992) reformulated by Mehler (2008) can be computed for a graph as follows:

$$Cp(G) = \frac{(Max(G) - (\sum_{v \in V} \sum_{w \in V} gd(v,w) + D_{\max}(G) \sum_{G' \in Com(G)} |G'||V| - |G'|^2))}{Max(G) - Min(G)} \quad (6)$$

with $Max(G) = D_{\max}(G)*(|V|^2 - |V|)$, $Min(G) = (|V|2 - |V|)$ and $Com(G)$ as the set of connected components of G. Further, $D_{\max}$ (G) is the maximal value a diameter of a linear graph of order G summed to 1 (this is done in order to set a distance for disconnected nodes which is larger than the maximally possible distance by one).

As noted in Mehler (2008), $Cp(G)$ is related to $\gamma(S)$ and *lcc*[9] though being more informative about graph structure than only about graph order and size. Even in cases of a single connected component not all vertices need to be connected. *Cp* captures the internal structure of the graph, that is, for the same *lcc* value *Cp* may differ reflecting the degree of connectivity of all vertices within a graph. Transferred to GSDNs, $Cp = 1$ means that all pairs of words are equally likely to be connected. This is a strong presupposition that contradicts the dependency principal, of hierarchical syntactic organization resulting in selective attachment of vertices. Of course, in a language like English, where the same word forms can function as different parts-of-speech the compactness could approximate 1 when the size of the treebank is sufficiently large. However, the *Cp* of 1 (i.e. complete connectivity of vertices) is highly improbable due to selective connectivity in GSDNs. Though, in general we expect the *Cp* values to be high in hierarchical networks that are known for short distances among vertices, that is, are compact (Alava & Dorogovtsev, 2004).

[9]The features $\gamma(S)$ and $R^2_{\gamma S}(G)$ represent the power-law fit of the distribution of connected components of G, and *lcc* is the fraction of the largest connected component of G (see Features F_{10} and F_{11} in Table 2).

4.7 Cohesion

The measure of cohesion *Ch* is used as an alternative measure of compactness (see Mehler, 2008). It is defined for an undirected graph as the fraction of all edges in the graph to the number of edges in a completely connected graph:

$$Ch(G) = \frac{\sum_{v \in V} d(v)}{|V|^2 - |V|} \in [0, 1] \tag{7}$$

Since the GSDNs are far from being completely connected (see the Zipfian distribution of degrees, Features F_8 and F_9) we expect rather small values of *Ch* for all languages. However, there can be typological and genealogical differences between languages as already outlined for the average degree.

5. TWO ALTERNATIVE APPROACHES TO AUTOMATIC LANGUAGE CLASSIFICATIONS

In this section we discuss approaches to automatically determine and classify languages as two alternatives to QNA. We evaluate their performance with respect to genealogical classes.

5.1 Language recognition: the NG-approach

Language classification is somehow related to the field of language recognition (LR). LR applies several techniques to guess the language of an input text (or speech) in order to solve an information retrieval task. Methods in LR use common words (Grefenstette, 1995), closed word classes (Lins & Gonçalves, 2004), single characters (Churcher et al., 1994; Takci & Sogukpinar, 2004) or *n*-grams (Cavnar & Trenkle, 1994; Dunning, 1994; Combrinck & Botha, 1995; Ahmed et al., 2004; McNamee, 2005) as features to distinguish among languages. The important thing concerning these approaches is the fact that they are all supervised. That means, firstly, that the classifier is trained on some data, and then new data pieces are categorized in comparison to the learned classes. That way, languages can be recognized. We implement the *n*-gram based approach (Cavnar & Trenkle, 1994) (henceforth abbreviated by NG) and apply it to our data in order to compare the outcomes with QNA. We check the genealogical performance of this method on the 11 languages. In the following, we describe the NG-approach.

The leading assumption of the NG-approach is that a language uses some character sequences or *n*-grams more frequently than others (Cavnar & Trenkle, 1994). Thus, these sequences can be used as indicators to classify languages. The overall classification procedure implemented here is the following:

1. Collect all *n*-grams present in a treebank.
2. Rank them according to their frequency of occurrence.
3. Build a vector of length *k* containing the most frequent *n*-grams[10] for each treebank.
4. Classify the treebanks by means of a classification algorithm (e.g. the one described in Section 6.1).

5.2 Quantitative Typology: the QT-approach

Quantitative typology starts from the view on language as a system with interrelated components, that is, from a holistic view (See Section 1). Consequently, approaches in this field aim to find correlations among different components or typological characteristics in order to get an insight into language structure (Greenberg, 1966).

Altmann and Lehfeldt (1973) propose a range of quantitative indices that allow processing of different linguistic levels (morphology, syntax, etc.) quantitatively. Altmann and Lehfeldt (1973) argue that the implementation of all the features should provide an adequate picture of a language. However, the lack of fully annotated language resources (e.g. with inflectional segmentation) needed to calculate the indices, complicates this task. We calculate some of the features proposed by Altmann and Lehfeldt (1973) (see Table 3 for an overview of these features) that are applicable to dependency treebank data. We try to classify languages by means of these features. The goal is the same as in the case of NG: to check whether this approach (henceforth abbreviated by QT – quantitative typology) in comparison to QNA.

The indices from Altmann and Lehfeldt (1973) are shown in Table 3. These selected indices possess expressive potential in the areas of morphology and (dependency) syntax.[11]

[10]We selected 300 *n*-grams as suggested by Cavnar and Trenkle (1994) for $n = \{1, \ldots, 6\}$.
[11]We selected these indices since they could be applied to our sort of data, i.e. dependency treebanks.

Table 3. Typological features from Altmann and Lehfeldt (1973). The last column lists the aggregation functions applied to the corresponding feature: μ, arithmetic mean; σ, standard deviation; H, entropy.

Feature	Short description	Linguistic area	Aggregation functions
Si	Synthesis index (Skalička, 1935)	Morphology	–
Dm	Dependency measure (Altmann and Lehfeldt, 1973)	Dep. syntax	μ, σ, H
Cn	Centrality (Andreev, 1967)	Dep. syntax	μ, σ, H
Sd	Sentence depth (Altmann and Lehfeldt, 1973)	Dep. syntax	μ, σ, H
Sw	Sentence width (Altmann and Lehfeldt, 1973)	Dep. syntax	μ, σ, H

The synthesis index is a quantitative feature applied to the whole treebank. The other features are tree-related, that is, they are calculated for each observation of a dependency tree. Aggregation functions are applied to these observations in order to get a single value of each index characterizing a treebank. Here, we use three aggregation functions to aggregate single observations of each value of an index in a treebank: arithmetic mean (μ), standard deviation (σ) and the entropy (H) (Figure 3, last column).

5.2.1 Synthesis Index

The synthesis index of Skalička (1935) is attributed to the morphological complexity of a language. For a sample (dependency treebank) the synthesis index is calculated as the number of sentences $|S|$ divided by the number of words $|W|$:

$$S_i = \frac{|S|}{|W|} \tag{8}$$

Skalička (1935)'s index is the simplest one since it does not require a morphological analysis of treebanks (Altmann & Altmann, 2005). This index allows the assignment of a value to the language on the analytic vs. synthetic scale of morphological complexity. It ranges between [0, 1] if $|S| < |W|$. The higher is the index, the more synthetic is the corresponding language. However, some authors (Altmann & Altmann, 2005) claim that the index is inappropriate for typological studies due to its high variability. In fact, *Si* is the reciprocal of the sentence length, which can

be influenced not only by a language type but also by other textual factors like author style, genre, etc. Unfortunately, factors like genres, authors, text types, etc., are not equally balanced in every treebank. However, we have extracted samples of equal size from each treebank and consider the mean and standard deviation values. From every treebank we select 29 text samples 1000 words each.[12]

Table 4 lists the *Si* values for the 11 languages. The variability of the index within the same language is rather small (see STD) for all treebanks, other than predicted by Altmann and Altmann (2005). However, it might be the case that our randomly selected samples cover only a small range of the language internal variation, so that in fact the variability of the index is higher. As expected, Russian as a highly synthetic language exhibits the highest $Si = 1$. Other languages have rather small values. Typologically the results obtained by the index are realistic, though, the differences between the single languages cannot be seen as valuable if they differ on the second or third decimal place.

5.2.2 Dependency Measure[13]

This index considers the question of how many dependent elements are subordinated to the root of a dependency tree. This concerns a) the number of elements directly or indirectly depending from the root as well as b) indirect dependents on deeper levels. The dependency measure

$$Dm = \frac{\sum_{j=1}^{m} j * x_j}{Dm_{Max}} \in [0, 1] \tag{9}$$

calculates this information for a single dependency tree: j is an index of levels starting from the root, that is, root has level 1, direct children have level 2 etc; m is the maximal level of a particular dependency tree (e.g. in Figure 5 $m = 4$ for graph (d)); j is used as a weighting factor multiplied by the number x_j of vertices on the corresponding level. Thus, the deeper the

[12]29 is the maximal number of samples with 1000 tokens that can be taken from the smallest treebank (i.e. from Slovene). That is, we select 29 as the least common number of samples for each treebank.

[13]This and the following indices were calculated on a sample of 1499 sentences from each treebank. This number is the smallest common number of sentences obtainable from each treebank.

Table 4. Mean and standard deviation values of *Si* for each treebank averaged over 29 text samples each of which contains 1000 words.

Language	Mean	Standard deviation
RUS	1.000	0.022
RUM	0.112	0.015
BUL	0.100	0.038
DAN	0.073	0.012
DUT	0.073	0.037
CZE	0.068	0.009
SLV	0.066	0.012
SWE	0.061	0.007
ITA	0.044	0.006
CAT	0.040	0.007
SPA	0.038	0.011

Table 5. *Dm* mean, entropy and standard deviation (STD) values. The languages are arranged in decreasing order of mean. Maximal entropy and STD values are underlined.

Language	Mean	Entropy	STD
DUT	0.792	6.109	0.134
RUM	0.708	6.556	0.178
BUL	0.600	7.082	0.140
RUS	0.538	7.985	0.158
CZE	0.512	8.721	0.177
DAN	0.488	8.378	0.161
SWE	0.470	8.896	0.146
SLV	0.461	8.673	0.132
SPA	0.402	9.622	0.163
CAT	0.382	9.732	0.149
ITA	0.370	9.726	0.202

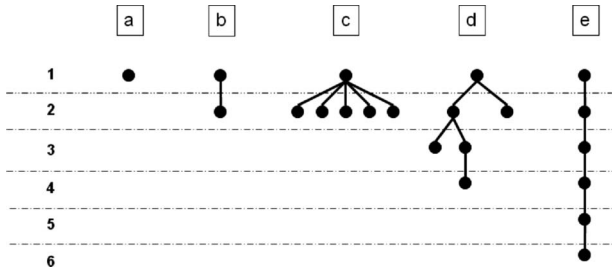

Fig. 5. Examples of different dependency trees.

tree, the more weighted are the vertices on deeper levels, the higher the value of *Dm*.

$$Dm_{Max} = \sum_{j=1}^{m} j = \frac{n(n(+1)}{2}$$

is the *Dm* of a linear graph (e.g. graph (e) in Figure 5). Thereby $m = n$ in this particular case, since each level of this graph has only 1 vertex.

Obviously, languages that preferably use short and flat dependency structures will have smaller *Dm* values than languages having complex dependency hierarchies. Typologically, this index can be highly informative if two similar languages, for example, prefer particular types of trees. However, the complexity of a dep. tree can also be an artefact resulting from the particular dependency theory rather than reflecting a property of language.

Table 5 lists the results for 11 treebanks in decreasing order of their mean values. Obviously, there are languages that are similar to graphs like (d), and others which are similar to graphs like (c) (in Figure 5). Dutch and Romanian are examples of more complex tree graphs that are closer to a linear graph than the other. Czech and Russian have presumably rather flat dependency structures (according to *Dm*), and Bulgarian lies somewhere in the middle. Romance languages Catalan, Spanish and Italian occupy the lower boundary of the *Dm* spectrum, having short and simple dependencies on average. Romance languages exhibit also the highest entropy, which assumes a higher redundancy of similar structures. The standard deviation is only high for Italian; presumably, there is a larger variation among the dependency structures compared with other Romance languages.

All in all, *Dm* allows for interesting insights into the organization of dependency relations in language. However, in order to make any serious claims various factors like the uniformity of dependency annotations, genre and style variations should be controlled.

5.2.3 Centrality

The coefficient of *Cn* originates from Andreev (1967). Here we use its modified version from Altmann and Lehfeldt (1973). *Cn* expresses the centrality of a predicate according to the linear order of the sentence. The sentence is represented as a sequence $a_l \ldots a_3\ a_2\ a_1\ Pa_1\ a_2 \ldots a_k$ with the predicate *P* as its central element; *k* is the index running from the left-most word after *P* and *l* is the index running backwards from the right-

Table 6. *Cn* mean, entropy and standard deviation (STD) values. The languages are arranged in decreasing order of mean. Maximal entropy and STD values are underlined.

Language	Mean	Entropy	STD
CZE	0.344	5.138	0.325
DUT	0.290	3.197	0.311
SLV	0.221	4.660	0.289
DAN	0.206	3.489	0.310
RUS	0.204	3.005	0.367
BUL	0.159	2.592	0.265
CAT	0.129	4.554	0.200
SWE	0.126	2.835	0.250
SPA	0.116	4.329	0.196
RUM	0.116	1.409	0.309
ITA	0.108	3.933	0.190

most word before *P*. Thus, *k* and *l* represent the number of words on the right or left side of the predicate. *Cn* is computed as follows:

$$Cn = 1 - \frac{|k - l| - \delta}{k + l} \in [0, 1] \tag{10}$$

where δ is used to avoid impreciseness in the case of an uneven number of words in a sentence:

$$\delta = \left\{ \begin{array}{l} 0 \; if k + l \; is \; even \\ 1 \; if \; k + 1 \; is \; even \end{array} \right\}.$$

A sentence $a_2\ a_1\ Pa_1\ a_2\ a_3$, for example,[14] is perfectly central, but if we omit δ we get a *Cn*-value less than 1:

$$Cn' = 1 - \frac{|3 - 2|}{3 + 2} = 1 - \frac{1}{5} = \frac{4}{5}.$$

Yielding δ we get the maximal centrality of 1:

$$Cn = 1 - \frac{|3 - 2| - 1}{3 + 2} = 1 - \frac{0}{5} = 1.$$

Typologically, there are languages with a centralized vs. polarized syntax (e.g. Latin, Hindi, Japanese etc.) (Altmann & Lehfeldt, 1973). If a language has mostly sentences of *Cn* nearby 1, then this language has a centralized syntax. If the predicate is likely to occur on the left or right hand side of the

[14]The example is taken from Altmann and Lehfeldt (1973).

sentence a language exhibits polarized syntax. Unfortunately, the index does not distinguish between right- and left-polarized types.

In our case, neither of the languages is centralized. Small differences occur between them, thereby, Spanish, Romanian and Italian have mostly polarized syntax, and Czech, for example, is rather centralized. Remarkably, Czech and Russian exhibit the highest standard deviations, which could reflect free word order, and the variation in the position of the predicate. In addition high entropy for Czech and low entropy values for Romanian undermine the above observation.

5.2.4 Sentence Depth

Sd is the proportion of the number of levels m[15] to the number of words in a dependency sentence.

$$Sd = \frac{j_{Max}}{n} = \frac{m}{n} \in [0, 1] \tag{11}$$

Sentence depth arranges languages similar to *Dm*. Obviously, these two measures are related since both take the depth of the dependency tree into account. *Sd* has nearly the same standard deviation values for all treebanks, which shows that the index is relatively stable. Italian, Spanish and Catalan group together again with lowest mean and entropy values.

5.2.5 Sentence Width

The *Sw* represents the rate of the maximal width of a single level ($W_{\max}$) to the number of elements in a dependency tree except for its nucleus (i.e. $n - 1$).

$$Sw = \frac{W_{\max}}{n-1} \in [0, 1] \tag{12}$$

This measure is relative to the number of elements in a sentence, since a small number of elements results in a higher *Sw*. *Sw* also correlates negatively with *Sd*, since deep sentences have mostly sparse levels (i.e. a small number of words on each level). This can be illustrated with graphs (c) and (d) in Table 9, whereby (c) results in small *Sd* and high *Sw* values and (d) the other way round.

[15] m is the maximal value of the index j (see *Dm*), i.e. the deepest level in the dependency tree.

Comparing *Sd* and *Sw* of the 11 languages, we see that the negative correlation holds, for example, for Dutch or Russian (high *Sd* and low *Sw*). However, Spanish and Catalan have both low *Sd* and *Sw* values on

Table 7. *Sd* mean, entropy and standard deviation (STD) values. The languages are arranged in decreasing order of mean. Maximal entropy and STD values are underlined.

Language	Mean	Entropy	STD
DUT	0.710	4.551	0.152
RUM	0.643	5.034	0.186
RUS	0.505	5.848	0.143
BUL	0.504	5.061	0.128
DAN	0.465	6.031	0.157
CZE	0.458	6.452	0.176
SWE	0.411	6.507	0.139
SLV	0.403	6.337	0.123
ITA	0.386	7.485	0.162
SPA	0.352	7.276	0.141
CAT	0.341	7.305	0.134

Table 8. *Sw* mean, entropy and standard deviation (STD) values. The languages are arranged in decreasing order of mean. Maximal entropy and STD values are underlined.

Language	Mean	Entropy	STD
ITA	0.461	7.487	0.173
BUL	0.458	5.019	0.142
RUM	0.434	4.741	0.172
DAN	0.415	6.076	0.224
DUT	0.385	4.039	0.128
SLV	0.371	6.547	0.160
RUS	0.359	5.449	0.180
SWE	0.351	6.340	0.128
CZE	0.349	5.718	0.154
SPA	0.314	7.213	0.141
CAT	0.305	7.196	0.120

Table 9. Values of *Dm*, *Sd* and *Sw* for the trees in Figure 5.

Index	a	b	c	d	e
Dm	1	1	0.52	0.71	1
Sd	1	1	0.33	0.66	1
Sw	1	1	1	0.40	0.2

average. This indicates that the two languages have both flat and small sentences. This could be, of course, a typological peculiarity of the two languages or simply the influence of text genre (mostly newswire texts).

We calculate the above measures for all treebanks and test their combined performance in a language classification task.

6. EXPERIMENTATION

6.1 Classification Scenario

To evaluate the performance of QNA, NG and QT regarding the genealogical gold standards, we classify languages by means of feature vectors consisting of feature values from one of the three approaches. The classification procedure instantiates the QNA algorithm of Mehler (2008) that can be summarized as follows:

1. Initially, each input network is represented by a vector of topological indices.
2. In the next step, a genetic search is performed to find salient features within the vectors that best separate the networks according to the underlying gold standard. However, this process may stop at a local maximum so that it does not necessarily find an optimal feature subset.
3. Based on the appropriately projected feature vectors, a hierarchical agglomerative clustering is performed together with a subsequent partitioning that is informed about the number of target classes.

In summary, QNA takes the set of input GSDNs together with the parameter space of linkage methods and distance measures to find out the feature subset that best separates the data according to the underlying classification.

The application of QNA can be illustrated by using the topological indices displayed in Table 2 as follows:

1. In the first step we extract a set N of GSDNs from the treebanks (see Section 3.2).
2. Then, we select network features $M = F_1, \ldots, F_n$ (see Section 4) and compute them for every graph G_i of the set N (Table 2).

3. Thirdly, we build a feature vector $v_i = (F_1(G_i), \ldots, F_n(G_i))$ consisting of composite feature values for every instance of F in graph G.
4. Finally, we cluster the networks by means of these feature vectors.

In analogy to QNA, we compute NG- and QT-related features and make them input to the classification algorithm included into QNA.

6.2 Evaluation

We evaluate our classifications by means of *F*-measure statistics. These statistics are usually applied in machine learning to evaluate, for example, the goodness of a classification of documents to predefined categories. They are based on two measures, *precision* and *recall*, which show for each category (language family) the amount of correctly-classified languages.

$$\text{Precision} : \frac{\#\{\text{correct} \cap \text{classified}\}}{\#\{\text{classified}\}} \notin [0, 1]$$

$$\text{Recall} : \frac{\#\{\text{correct} \cap \text{classified}\}}{\#\{\text{correct}\}} \in [0, 1]$$

$$\text{F-score} : \frac{2}{1/\mathit{precision} + 1/\mathit{recall}} \notin [0, 1]$$

Precision relates the number of correctly classified languages to the total number of languages classified to this group. Recall relates the correctly classified languages to the number of languages which are known to belong to this group. Both measures are in the range of [0, 1], where 1 indicates that languages are classified perfectly and 0 that the classification failed. The *F*-score mediates between the two measures evaluating the overall performance of the classification. That means, if precision and recall are high, the *F*-score of a category is also high (i.e. close to 1). The *F*-measure is a weighted harmonic mean that considers

the *F*-scores of all the categories. Given a known partition of languages M (e.g. known genealogical classes), the set of language classes $c_i \in C$ and the total number of languages L, the *F*-measure is computed as follows (Hotho et al. 2005):

$$F - Measure(M) = \sum_{i=1}^{|C|} \frac{|c_i|}{L} F_{score}(c_i) \in [0, 1] \tag{13}$$

With the *F*-measure we get a single value between 0 and 1 which characterizes the overall success of the classification.

The results are tested by comparison with two kinds of baseline, one with a known-partition (i.e. the algorithm "knows" how many languages should be in each group), and the other assuming an equi-partition of languages. Using the two scenarios, languages are randomly assigned to the target categories, and the probability (in terms of *F*-measure) of assigning languages correctly that way is computed. This random assignment does not necessarily result in an *F*-measure value of 0. Its value can be understood as an expected value of assigning languages completely by chance. Thus, the classification k succeeds, if random clustering is outperformed such that the *F*-measure F_k of the classification is $F_k \gg F_{rand}$.

6.3 Genealogical Classification

To evaluate the three competing approaches NG, QT and QNA we check whether we can successfully classify languages into 3 genetic groups: Slavic, Germanic and Romance (Section 7) using QT, NG and QNA. We determine the number of clusters to be equal 3 and check whether languages are classified correctly.

A problem with the *n*-gram based approach (NG) concerns Russian and Bulgarian, which both use the Cyrillic writing system. This means that the *n*-grams of both languages cannot be directly compared with the other nine languages using the Latin alphabet. Rather some transliteration effort is needed. In order to avoid biases due to different writing systems, we tested an additional variant of the NG-approach excluding the two Slavic languages (abbreviated by NG-RB). This addition results in four different procedures that are compared with each other. The best performing combination of NG is visualized by means of a dendrogram.

7. RESULTS AND DISCUSSION

The results of the semi-supervised clustering experiment[16] on 11 languages are presented in Tables 10 to 13. The tables are structured as follows. The first column explains the clustering procedure used. The second column presents the corresponding *F*-measure values, and the third how many features were used from the total number of features in the setting (e.g. 11/13). The best results for the three approaches and the best random baselines are listed in Figure 9. Figures 8 (i.e. QNA), 6 (i.e. QT) and 7 (i.e. NG) display the best performing results in terms of a dendrogram. The height of a bar connecting two languages or clusters of languages shows the degree of dissimilarity. Thus, the lower the degree of agglomeration of two clusters, the higher the similarity of the languages (e.g. Figure 8, Italian and Romanian) within these clusters.

7.1 QT-experiment

From Table 10 and Figure 6 we can see that a subset of 11 from 14 features accounts for the best possible classification of languages into 3 groups (*F*-measure: .76389). That is, about 70% of languages are classified correctly. Figure 6 illustrates the within-cluster similarities resulting from applying the best-off-feature combination (11 features). Remarkably, the three Slavic languages Slovene, Russian and Bulgarian are classified within the same block. Romance languages, Spanisch and Catalan as well as Italian and Romanian are also pairwise similar, though not within the same Romance cluster. Swedish is grouped together with Catalan and Spanish, and Danish attaches to Swedish, though, with a greater dissimilarity (i.e. see the height of the bar). The total outliers are Czech and Dutch.

The overall similarities of languages reveal that the QT approach is able to recognize genealogical relationships of most languages used here. However, for the rest of the languages the classification fails, which can be due to several reasons; for example, the size of the treebank (large size of Czech), style and register variations, different annotations could have biased the result. Further, it is still possible that the observed similarities within the clusters reflect typological similarities between languages that

[16]All the computations of the cluster analysis are made using MATLAB version 7.11.0.584 7. (R2010b) including the Statistics and Curve Fitting Toolboxes (www.mathworks.de).

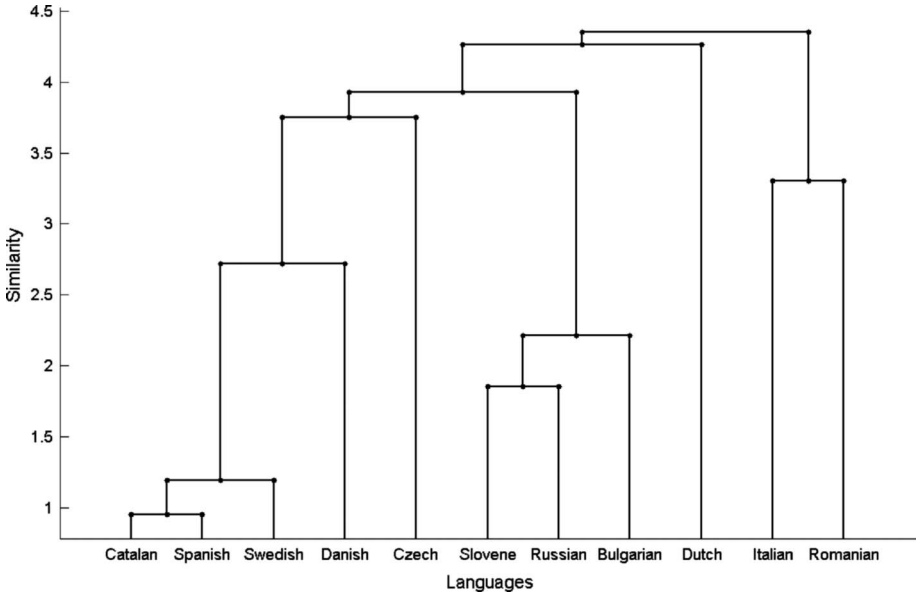

Fig. 6. The similarity tree of languages generated by the best feature combination of QT.

Table 10. *F*-measures of classifying 11 languages into three genetic groups by means of QT.

Procedure	*F*-measure	Features
QT [mahalanobis, complete]	0.76389	11/14
QT [mahalanobis, weighted]	0.76389	11/14
QT [mahalanobis, complete]	0.76111	5/14
QT [mahalanobis, complete]	0.65972	14/14
AVG over non-random approaches	0.73720	
Random baseline II	0.56286	Known partition
Random baseline I	0.54869	Equi-partition

come to the fore when considering lengths, depths, widths and centralities of sentences. A closer look at the distributions of single features (see Section 5.2 for a discussion) should lead to a better understanding of the typological values of the features.

We repeated the genetic search for best feature combinations 10 times to see which features remain in each of the combinations. However, only

one best-off combination of 11 features was able to produce the highest *F*-measure value.

7.2 NG-experiment

The NG-based classification of all the 11 languages (including Russian and Bulgarian) achieves an *F*-measure value of 0.81061.

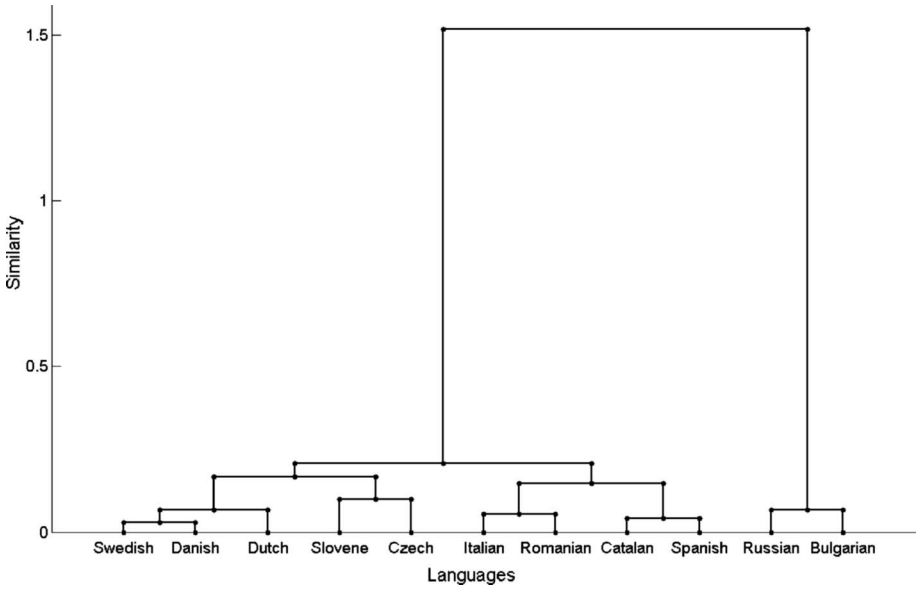

Fig. 7. The similarity tree of languages generated by the best feature combination of the NG-experiment.

Table 11. *F*-measures of classifying 11 languages into three genetic groups by means of NG.

Procedure	*F*-measure	Features
NG [correlation, complete]	0.81061	28/61
NG [correlation, average]	0.81061	28/61
NG [correlation, single]	0.80606	28/61
NG [correlation, average]	0.80606	61/61
AVG over non-random approaches	0.80830	
Random baseline II	0.58995	Known partition
Random baseline I	0.57790	Equi-partition

The NG-experiment shows that we are able to classify languages perfectly excluding the both Slavic languages. However, adding these languages results in a drop of the *F*-measure. On the one hand, this is certainly a loss, since we achieve an *F*-measure of 1 (see Table 12) only when dealing with nine instead of 11 languages. On the other hand, the approach would presumably perform perfectly if Russian and Bulgarian were transliterated into the Latin writing system. So all in all, the NG-experiment shows good performance using about 30 *n*-gram features. Each language family seems to have particular characters in common that are not shared (or not commonly shared) within other families.

In the context of the present study, this simple but well performing approach tells us the correct language family; however, it can also be misleading typologically, since orthographic standards change in languages and distance between languages based solely on graphemes might be biased by orthographic peculiarities. In French, for example,

Table 12. *F*-measures of classifying nine languages into three genetic groups by means of NG-RB (*n*-gram based classification, Russian and Bulgarian excluded).

Procedure	*F*-measure	Features
NG-RB [correlation, complete]	1.0	36/61
NG-RB [correlation, average]	1.0	36/61
NG-RB [correlation, weighted]	1.0	36/61
NG-RB [correlation, complete]	0.89206	61/61
AVG over non-random approaches	0.97300	
Random baseline II	0.58426	Known partition
Random baseline I	0.57725	Equi-partition

Table 13. *F*-measures of classifying 11 languages into three genetic groups by means of QNA.

Procedure	*F*-measure	Features
QNA [mahalanobis, complete]	1.0	7/22
QNA [mahalanobis, ward]	1.0	8/22
QNA [mahalanobis, complete]	1.0	10/22
QNA [mahalanobis, weighted]	0.90909	13/22
AVG over non-random approaches	0.97730	
Random baseline II	0.56297	Known partition
Random baseline I	0.55442	Equi-partition

Table 14. Best feature combinations of QNA resulting in an *F*-measure of 1 found by the genetic search (the algorithm was run 20 times, seven best-off combinations were found). The most frequently selected features are marked with black triangles.

Feature		Combinations							Sum
	ε							✓	1
	C_{br}			✓	✓		✓	✓	4
	C_{ws}		✓			✓			2
	lcc	✓				✓		✓	3
	L		✓						1
	r	✓							1
	γ	✓		✓				✓	3
▶	$R^2\gamma$	✓		✓	✓	✓	✓	✓	6
	$\gamma knn(k)$		✓						1
▶	$R^2knn(k)$	✓	✓	✓	✓	✓	✓	✓	7
	$\gamma C(k)$	✓							1
	$R^2C(k)$	✓	✓						2
	Cp	✓	✓		✓	✓	✓		5
	CC			✓	✓	✓	✓	✓	5
	GC					✓			1
	DC		✓	✓	✓		✓		4
	γS		✓						1
	$R^2\gamma S$	✓							1
	diam					✓			1
	Coh	✓				✓			2
	C_A	✓	✓						2

the grapheme-to-phoneme correspondence is not trivial; many characters are written but not pronounced (e.g. *manquent* vs. [mãk]). This fact should influence the distance of French to other Romance languages when considering character based distances. This is just one example. However, we should be aware of such factors when we aim to go beyond identification of language families.

7.3 QNA experiment

QNA also achieves the best possible *F*-measure of 1.0 using at least six features. That means six network characteristics suffice to separate languages perfectly with respect to genealogical relationships.

At this point, we can conclude that the genealogical classification succeeds but in order to get more fine-grained knowledge about the typological information gain of the approach we have to examine single network characteristics (as done in Section 4). In order to see which

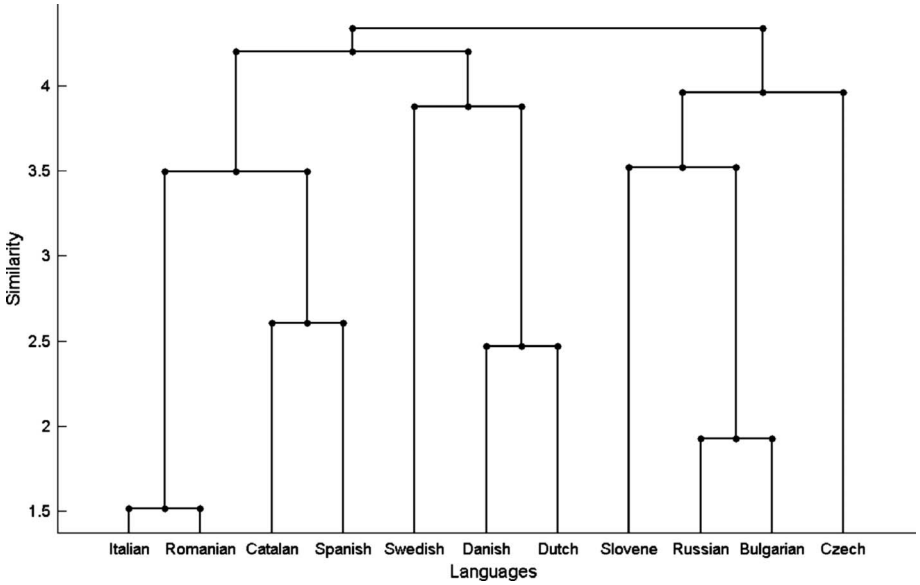

Fig. 8. The similarity tree of languages generated by the feature combination no. 6 (cf. Appendix, Fig. 9) used to perform QNA.

network characteristics perform best, we have run a genetic search for best feature combination twenty times and got seven best-off combinations producing an *F*-measure of 1.0 (see Table 14). The best performing features are definitely the two adjusted coefficients of determination of the distributions of nodes' degrees and degrees of nearest neighbours. This result is surprising, since we did not expect a good separability of GSDNs by means of these features (i.e. due to expecting a homogeneous impact of Zipf's Law). Further, centrality, clustering and compactness seem to be important building blocks that are present in each of the combinations.

The success of clustering can be explained by the loss of inflection in the particular languages. As discussed in Section 4, the probability of a word form to appear in many different dependency relations is higher for analytic, than for synthetic languages. This increases the probability of clusters in a GSDN. Taking C_{ws} as and example, we see from Figure 9 that Germanic languages have clearly higher C_{ws} values, than the Slavic languages. The Romance group is less homogeneous: Spanish and Catalan are close to Germanic, and Romanian and Italian to the Slavic group according to their C_{ws} values. So, this feature alone is not sufficient

in order to separate the languages perfectly, other features are needed to complete the picture. But C_{ws} can be nicely used to examine the typological properties of languages by means of networks. Note also that in cases where C_{ws} is not selected, C_{br} or $\gamma C(k)$ with $R^2_{C(k)}$ are present instead.

Further, C_{ws} and L are strongly negatively correlated ($corr = -0.8997$), which nicely points to the small world property of GSDNs that have short geodesic distances and high cluster values. Russian, for instance, has the smallest C_{ws} value and the highest value of L. Swedish and Catalan, in turn, exhibit high clustering and the shortest geodesic distances among the 11 GSDNs.

Features of centrality tell us something about the amount of central vertices in a network. A network of few central and many peripheral vertices is centralized. Conversely, the centrality values becomes the smaller, the more equally distributed the degrees (distances) of the vertices. In the case of GSDNs, this means to have many highly connected word forms in terms of their dependency relations. Germanic languages have lower centrality values, than Slavic, and partially Romance languages. This means that there are much more central word forms in Germanic networks than in Slavic ones. It is plausible to assume that this relates to prepositions that play a greater role in Germanic languages.

At least, compactness expresses the overall connectedness of a GSDN. Compactness is negatively correlated, though not significantly ($corr = -0.3015$, $p = 0.3676$), with L, since the larger L the less compact the graph.

In fact, many features that we used are correlated[17] (e.g. the correlation between DC and CC is positive $corr = 0.7511$, $p = 0.0077$, between Cp and lcc is $corr = 1$, $p = 0$ (perfect correlation), and between C_{ws} and L is negative $corr = -0.8997$, $p = 0.0002$).[18] The positively correlated features can be easily exchanged without the loss in F-measure. If, for example, feature DC is selected, then feature CC is not needed to improve the result. The same holds for both clustering coefficients. Negative correlations mean that large values of feature X result in small values of feature Y (or vice versa); however, both

[17]We have computed the pairwise correlations among all features. Here and in the following paragraphs we show only results that are statistically significant.
[18]Note that all correlations exemplified here are significant with a p-value < 0.05.

Table 15. The overall *F*-measures of the genealogical classification. The best non-random *F*-measures represent the best classification results from Tables 11 to 14. The average non-random are the average *F*-measures over different combinations and (average) random *F*-measure values.

Method	QT	NG	NG-RB	QNA
Best non-random	0.76389	0.81061	1.00000	1.00000
Average non-random	0.73720	0.80830	0.97300	0.97730
Random	0.56286		0.58995	0.56286

features can be informative in classification (like in case of C_{ws} and *L*). *Cp* and *lcc* exhibit a perfect correlation, and it becomes obvious from Table 14 that in three of two cases where *lcc* is used, *Cp* is used too. This is not what we would expect; rather one of the two features should suffice for the best-off selection. However, the mahalanobis distance used for clustering ensures that the features used get statistically independent, that is, correlated features become uncorrelated in the final representation.

8. CONCLUSION

In this article, we experimented with three different approaches to automatic language classification. We tested their performance in the task of genealogical language classification.

In case of applying QNA to classify languages, a network of a language was created by taking a dependency treebank of a particular language as input and mapping words of the treebank to vertices and dependency relations to edges. In this way, a network of a language was constructed. Since we dealt not only with networks of different treebanks but with networks of treebanks from different languages, the leading assumption was that particular characteristics of the language might have left their fingerprint on the structure of the network. To account for this assumption, we calculated 21 topological characteristics on the language networks and classified languages by means of them in order to see whether the similarities in network structure reflect some ''real'' similarities among languages.

Indeed, we found out that some network indices perfectly reproduce the genealogical relations between the languages. The network structure

seems to cover several linguistic levels (i.e. morphology, syntax, lexis) and to provide a more abstract, general (holistic) view on language. Further, our results revealed four classes of features. Features in a class are positively correlated. They are selected by the genetic search depending on which features from other classes are selected. So, we can speak of a network of features. In future work, we aim at a systematic examination of this feature network.

The main advantage of QNA lies in the integrated view on language that enlarges the range of possibilities to examine the language as a whole system. Of course, a disadvantage of the approach is the lack of transparency with respect to the role of single features, which should be examined together with other, and in isolation. Further, we still do not know about all possible sources of bias (i.e. influence of genres, dependency theories, etc.). Though, the results are highly encouraging by producing a perfect classification, future work should systematically identify and eliminate possible error factors before we will be able to make judgements about the overall potential of QNA in the area of language classification.

Quantitative typological indices were computed either for single sentences or the whole treebank (sample), and used for classification. This approach is the least efficient in terms of *F*-measure. Typologically, though, QT might be interesting; why do, for instance, languages like Czech and Dutch differ from their language family members? And how can these deviations be explained with respect to features of dependency structure as explored here? However, we expect QT features to be biased even more than the network based classification. QT features like sentence width, depth, etc., can depend on the type of texts (or genre) in a sample and vary to a larger extent even within a single language. Here, we aim to examine the role of such indices within and among languages in order to be able to better interpret the results.

At least the expressiveness of *n*-gram based classifications was again confirmed in this article. NG is typologically presumably not that interesting. From the point of view of application, however, NG is more easily implemented than the other two. Neither annotated treebanks nor the need to calculate indices is required for NG. Of course, transliteration might be a problem though, but all in all, NG is a robust means when it comes to determining genetic relationships. When we aim to look at typological relations between linguistic levels, QNA and QT might be a better choice.

In summary, the three approaches to automatic language classification show good performance. Typologically promising are QNA and QT, though further studies should follow in order to learn more about the possibilities and limitations of these approaches.

ACKNOWLEDGEMENTS

This work was supported by the Linguistic Networks project (http://www.linguistic-networks.net/) funded by the German Federal Ministry of Education and Research (BMBF), and by the German Research Foundation Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication".

We are grateful to Ramon Ferrer i Cancho, Barbara Job, Tatiana Lokot and the anonymous reviewers for their useful comments.

REFERENCES

Ahmed, B., Cha, S.-H., & Tappert, C. (2004). Language identification from text using n-gram based cumulative frequency addition. *Proceedings of the CSIS Research Day*. New York: Pace University.

Alava, M., & Dorogovtsev, S. (2004). Preferential compactness of networks. *Condensed Matter*, page 0407643, arXiv:cond-mat/0407643v1.

Altmann, G., & Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Wilhelm Fink.

Altmann, G., & Altmann, V. (2005). Erlkönig und Mathematik. Retrieved 3.09.2011, from http://ubt.opus.hbz-nrw.de/volltexte/2005/325/

Andreev, N. D. (1967). *Statistiko-kombinatornye metody v teoretičeskom i prikladnom jazykovedenii*. Leningrad: Nauka.

Anttila, R. (1972). *An introduction to historical and comparative linguistics*. New York: The Macmillan Company.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Batagelj, V., Kerzic, D., & Pisanski, T. (1992). Automatic clustering of languages. *Computational Linguistics*, *18*(3), 339–352.

Blanchard, P., Petroni, F., Serva, M., & Volchenkov, D. (2009). Networking phylogeny for indo-European and Austronesian languages. *Nature Proceedings*. Retrieved 3.09.2011, from http://precedings.nature.com/oai2 (United States).

Boguslavsky, I., Chardin, I., Grigorieva, S., Grigoriev, N., Iomdin, L., Kreidlin, L., & Frid, N. (2002). Development of a dependency treebank for Russian and its possible applications in NLP. *Proceedings of LREC 2002*, Las Palmas, 825–856.

Bollobás, B., & Riordan, O. M. (2003). Mathematical results on scale-free random graphs. In S. Bornholdt & H. G. Schuster (Eds), *Handbook of Graphs and Networks. From the Genome to the Internet* (pp. 1–34). Weinheim: Wiley-VCH.

Bosco, C., Lombardo, V., Vassallo, D., & Lesmo, L. (2000). Building a treebank for Italian: a data-driven annotation schema. *Proceedings of LREC 2000*, Athens, 99–105.

Botafogo, R. A., Rivlin, E., & Shneiderman, B. (1992). Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, *10*(2), 142–180.

Bryant, D., Filimon, F., & Gray, R. (2005). Untangling our past: Languages, trees, splits and networks. In R. Mace, C. Holden & S. Shennan (Eds), *The Evolution of Cultural Diversity* (pp. 67–84). London: UCL Press.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, US, 161–175.

Churcher, G., Hayes, J., Johnson, S., & Souter, C. (1994). Bigraph and trigraph models for language identification and character recognition. *Proceedings of 1994 AISB Workshop on Computational Linguistics for Speech and Handwriting Recognition*, Leeds.

Civit, M., Bufí, N., & Valverde, M. P. (2004). CAT3LB: a Treebank for Catalan with Word Sense Annotation. *TLT2004*, Tübingen University, 27–38.

Civit, M., & Martí, M. (2005). Building Cast3LB: A Spanish Treebank. *Research on Language and Computation*, *4*, 549–574.

Combrinck, H., & Botha, E. (1995). Text-based automatic language identification. *Proceedings of the Sixth Annual South African Workshop on Pattern Recognition*, Rand Afrikaans University, Gauteng, South Africa.

Daumé III, H. (2009). *Non-parametric Bayesian model areal linguistics*. North American Chapter of the Association for Computational Linguistics (NAACL). Boulder, CO: ACM.

Dunning, T. (1994). Statistical identification of language. *Technical Report MCCS-94-273*. Computing Research Lab (CRL), New Mexico State University.

Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtský, Z., & Žele, A. (2006). Towards a Slovene dependency treebank. *Proceedings of LREC 2006*, Genoa, 1388–1391.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.

Ferrer i Cancho, R., Solé, R. V., & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E*, *69*, 051915, 69:5.

Ferrer i Cancho, R., Mehler, A., Pustylnikov, O., & Díaz-Guilera, A. (2007). Correlations in the organization of large-scale syntactic dependency networks. *TextGraphs-2: Graph-Based Algorithms for Natural Language Processing* (pp. 65–72). Rochester, NY: HLT/NAACL.

Franks, S. (2005). The Slavic languages. In G. Cinque & R. Kayne (Eds.), *Handbook of Comparative Syntax* (pp. 373–419). Oxford: Oxford University Press.

Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of Language* (pp. 73–113). Cambridge: MIT Press.

Grefenstette, G. (1995). Comparing two language identification schemes. *3rd International Conference on the Statistical Analysis of Textual Data*, CISU, Rome, 263–268.

Hage, P., & Harary, F. (1995). Eccentricity and centrality in networks. *Social Networks*, *17*, 57–63.

Hajič, J. (1998). Building a syntactically annotated corpus: The Prague Dependency Treebank. In E. Hajičová (Ed.), *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová* (pp. 106–132). Prague: Karolinum, Charles University Press.

Haspelmath, M., Dryer, M., Gil, D., & Comrie, B. (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Holman, E. W., Wichmann, S., Brown, C. H., Vilupillai, V., Müller, A., Brown, P., & Bakker, D. (2008). Explorations in automated language classification. In T. Fanego (Ed.), *Folia Linguistica* (pp. 331–354). Berlin: De Gruyter.

Holt, R. C., Schürr, A., Elliott Sim, S., & Winter, A. (2006). GXL: A graph-based standard exchange format for reengineering. *Science of Computer Programming*, *60*(2), 149–170.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *Journal for Language Technology and Computational Linguistics* (JLCL), *20*(1), 19–62.

Hristea, F., & Popescu, M. (2003). A dependency grammar approach to syntactic analysis with special reference to Romanian. In F. Hristea & M. Popescu (Eds), *Building Awareness in Language Technology* (pp. 65–76). Bucharest: University of Bucharest Publishing House.

Kondrak, G. (2002). *Algorithms for language reconstruction*. PhD thesis, University Toronto.

Kromann, M. T. (2003). The Danish dependency treebank and the underlying linguistic theory. In J. Nivre & E. Hinrichs (Eds), *Proceedings of TLT 2003* (pp. 217–220). Sweden: Växjö University Press.

Kruskal, J. B., Black, P., & Dyen, I. (1992). *An Indo-European Classification. A Lexicostatistical Experiment (Transactions of the American Philosophical Society)*. American Philosophical Society.

Lins, R. D., & Gonçalves, P. (2004). Automatic language identification of written texts. *Proceedings of the ACM SAC '04*, ACM, New York, NY, USA, 1128–1133.

Liu, H. (2008). The complexity of Chinese syntactic dependency networks. *Physica A*, *387*, 3048–3058.

Liu, H., Zhao, Y., & Huang, W. (2010). How do local syntactic structures influence global properties in language networks? *Glottometrics*, *20*, 38–58.

Masayoshi, S., & Bynon, T. (1995). Approaches to language typology. A conspectus. In S. Masayoshi (Ed.), *Approaches to Language Typology* (pp. 1–25). Oxford: Oxford University Press.

McNamee, P. (2005). Language identification: a solved problem suitable for undergraduate instruction. *Journal of Computing in Small Colleges*, *20*(3), 94–101.

Mehler, A. (2008). Structural similarities of complex networks: A computational model by example of Wiki graphs. *Applied Artificial Intelligence*, *22*, 619–683.

Mehler, A., Geibel, P., & Pustylnikov, O. (2007). Structural classifiers of text types: Towards a novel model of text representation. *LDV Forum*, *22*(2), 51–66.

Mehler, A., Pustylnikov, O., & Diewald, N. (2010). The geography of social ontologies: The Sapir-Whorf hypothesis revised. *Computer, Speech and Language. Special Issue on Network models of social and cognitive dynamics of language*. London: Academic Press Ltd.

Mehler, A., & Lokot, T. (2012). Towards an adequate measure of compactness of graphs. In preparation.

Minkov, E., & Cohen, W. W. (2008). Learning graph walk based similarity measures for parsed text. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, 907–916 (Association for Computational Linguistics).

Newman, M. E. J., & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, *68*, 036122, arXiv:cond-mat/0305612v1.

Nivre, J., Nilsson, J., & Hall, J. (2006). Talbanken05: A Swedish treebank with phrase structure and dependency annotation. *Proceedings of LREC 2006*, Genoa, 1392–1395

Osenova, P., & Simov, K. (2004). *BTB-TR05: BulTreeBank Stylebook. BulTreeBank Project Technical Report Nr. 05*. Linguistic Modelling Laboratory, Bulgarian Academy of Sciences.

Pastor-Satorras, R., Vázquez, A., & Vespignani, A. (2001). Dynamical and correlation properties of the internet. *Physical Review Letters*, *87*, 258701.

Pustylnikov, O., & Mehler, A. (2008). Towards a uniform representation of treebanks: Providing interoperability for dependency tree data. *Proceedings of First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, Hong Kong SAR, January 9–11.

Pustylnikov, O., Mehler, A., & Gleim, R. (2008). A unified database of dependency treebanks. Integrating, quantifying & evaluating dependency data. *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco, 3359–3365.

Sapir, E. (1921). *Language*. New York: Harcourt, Brace and World.

Schank, T., & Wagner, D. (2005). Approximating clustering coefficient and transitivity. *Journal of Graph Algorithms and Applications*, *9*(2), 265–275.

Skalička, V. (1935). *Zur ungarischen Grammatik*. Prague: Filosofická fakulta University Karlovy.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings American Philosophical Society*, Philadelphia, 453–463.

Takci, H., & Sogukpinar, I. (2004). Centroid-based language identification using letter feature set. In A. Gelbukh (Ed.), *LNCS 2945* (pp. 640–648). Berlin/Heidelberg: Springer-Verlag.

Tesnière, L. (1959). *Eléments de syntaxe structurale*. Paris: Klincksieck.

van der Beek, L., Bouma, G., Malouf, R., & van Noord, G. (2002). The Alpino dependency treebank. In T. Gaustad (Ed.), *Computational Linguistics in the Netherlands CLIN* (pp. 1686–1691). Amsterdam: Radopi.

Warnow, T., Ringe, D., & Taylor, A. (1996). Reconstructing the evolutionary history of natural language. *Proceedings of ACM-SIAM Symposium on Discrete Algorithms* (SODA), Society of Industrial and Applied Mathematics, 314–322.

Wasserman, S., & Faust, K. (1999). *Social Network Analysis. Methods and Applications*. Cambridge: Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440–442.

Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press.

APPENDIX

Network	$\|V\|$	$\|E\|$	ϵ	C_{hr}	C_{ms}	lcc	L	r	γ	$\bar{R}^2$	$\gamma_{\bar{k}_{nn}(k)}$	$\bar{R}^2$	$\gamma_{C(k)}$	$\bar{R}^2$	Cp	CC	GC	DC	St	γ_S	$\bar{R}^2$	D	Ch	C_A
CAT	38882	215308	5.5374	0.0098	0.231	0.9978	3.0579	-0.171	-1.4616	0.9988	-0.5971	0.9726	-0.5355	0.9388	0.9957	0.2629	0.0565	0.2679	0	-2.1158	0.9839	9	0.0002	0.3382
ITA	7984	24269	3.0397	0.0122	0.1414	0.9898	3.412	-0.2265	-1.6307	0.9954	-0.5435	0.8412	-0.5699	0.7919	0.9795	0.2026	0.037	0.1901	0.0002	-1.9923	0.9956	11	0.0007	0.3039
RUM	8867	23901	2.6955	0.0053	0.0932	0.9981	3.4466	-0.1862	-1.801	0.9947	-0.5563	0.81	-0.6024	0.7949	0.9961	0.2081	0.0451	0.2297	0.0002	-1.8073	0.9784	12	0.0006	0.2861
SPA	17101	56911	3.3279	0.0069	0.1788	0.9804	3.1749	-0.1815	-1.6785	0.9966	-0.6263	0.9431	-0.5659	0.9011	0.961	0.2679	0.0622	0.2787	0.0001	-5.1055	0.9998	10	0.0003	0.3052
BUL	32421	95698	2.9517	0.0055	0.1693	0.995	3.3111	-0.2025	-1.5476	0.9976	-0.5932	0.9479	-0.5921	0.9313	0.99	0.2335	0.0376	0.2057	0	-2.0809	0.9986	12	0.0001	0.2731
CZE	146504	696379	4.7533	0.0038	0.1342	0.9714	3.3809	-0.0817	-1.1678	0.9992	-0.6178	0.9818	-0.566	0.9454	0.9436	0.2593	0.0314	0.2376	0	-3.5597	0.9998	16	0	0.1992
RUS	58283	177942	3.053	0.0045	0.0883	0.9927	3.7141	-0.0975	-1.463	0.9981	-0.4888	0.9391	-0.4422	0.8688	0.9854	0.2157	0.0094	0.1565	0	-2.8336	0.9999	21	0.0001	0.1742
SLV	8342	20453	2.4518	0.0097	0.0946	0.9647	3.6044	-0.1879	-1.8486	0.9948	-0.554	0.8737	-0.6013	0.8263	0.9304	0.234	0.0538	0.1641	0.0003	-3.3107	0.9999	9	0.0005	0.3727
DAN	19133	50858	2.6581	0.0127	0.1867	0.9876	3.3257	-0.2677	-1.4443	0.9971	-0.4866	0.9439	-0.5817	0.9178	0.9753	0.2096	0.0327	0.1429	0.0001	-1.8512	0.9939	16	0.0002	0.2027
DUT	32599	112613	3.4544	0.0061	0.1367	0.9934	3.4044	-0.1956	-1.2783	0.9979	-0.6119	0.9288	-0.5486	0.8383	0.9868	0.2066	0.0371	0.1814	0	-2.777	0.9973	11	0.0002	0.3054
SWE	25097	126526	5.0414	0.0309	0.2629	0.9943	3.1386	-0.2345	-1.153	0.9987	-0.408	0.9183	-0.3787	0.7891	0.9885	0.2167	0.0379	0.1496	0	-3.6197	0.9998	11	0.0004	0.2821
				▲					▲	▲		▲				▲		▲						

Fig. 9. The feature vectors used for classification. SLV, Slovene; SPA, Spanish; SWE, Swedish; ITA, Italian; RUM, Romanian; DUT, Dutch; BUL, Bulgarian; CAT, Catalan; DAN, Danish; RUS, Russian; CZE, Czech. Black triangles in the lower row point at features from one of the six-best-of-combinations.