Morpheme networks reveal language dynamics

Daniela Barbara Keller[1], Jörg Schultz[1]

[1]Dept. of Bioinformatics – Biozentrum, Am Hubland – University of Würzburg
D-97074 Wuerzburg – Germany

Abstract

Morphemes are the smallest meaningful parts of words. As one word can be composed of multiple morphemes, one morpheme can be present in more than one word. Therefore, morphemes represent a natural unit to study the evolution of words. Here, we analyzed three different Indo-European languages and compared variants of these languages differing in time and region. For this purpose, we adopted a network based approach from bioinformatics. It is frequently used for analyzing domains, the structural, functional and evolutionary units of proteins. Despite the global similarity of the morpheme networks, we identified characteristics associated with fundamental differences in word formation. Comparisons of the networks revealed that the fate of a morpheme is highly influenced by its connectivity. As a morpheme contains meaning, differences between the networks revealed cultural changes over time and between regions. Therefore, morphemes represent not only meaningful but also evolutionary parts of words.

Keywords: Language, evolution, cultural change.

1. Introduction

Caused by the growing amount of electronically available text data, the importance of computational methods for the study of language structure and language evolution has increased over the last years. This trend fulminated in the Google books project which analyzed more than 5 million books indicating a paradigm shift from query driven to data driven research (Michel *et al.*, 2011). A similar shift has happened in molecular biology with the emergence of genome sequencing. In analogy to the `-omics' sciences in biology Michel *et al.* (2011) coined the term `culturomics' for the analysis of cultural trends using text data. Obviously, the word is not the only unit to analyze the evolution of a language. It happens only rarely that a so far meaningless string becomes associated with a meaning. An example is the association of the string `smurf' with little blue creatures with white hats. More frequently, new words arise by the fusion of two so far not related words or meanings. As an example, the word of the year 2010 in Germany was `Wutbürger' (anger-citizen) which was generated by fusing two words (`Wut'—anger and `Bürger'—citizen) (Gesellschaft für deutsche Sprache, 2010). This new word denotes middle class citizens who are increasingly unsatisfied with political decisions. Thus, to understand the evolution of words, one also has to look at the parts which compose a word. So called morphemes represent the smallest meaningful parts of a word (Haspelmath et Sims, 2010). For example the word `unchained' can be decomposed into three morphemes `un', `chain' and `ed'.

Looking at the recombination of morphemes could provide an insight into the concept of word formation and the creation of new words. As morphemes are the smallest meaningful parts, a comparative analysis should also reveal changes in culture.

To model the evolution of morphemes in words, we built networks of morphemes following the concept of protein domain networks in biology (Wuchty, 2001; Wuchty et Almaas, 2005). Domains are the smallest structural, evolutionary and functional units of proteins (Copley *et al.*, 2002; Janin et Chothia, 1985). Therefore they can be seen as analogous to morphemes in words. We used headwords from eight dictionaries and lemmata from two corpora differing in time and region (Figure 1).

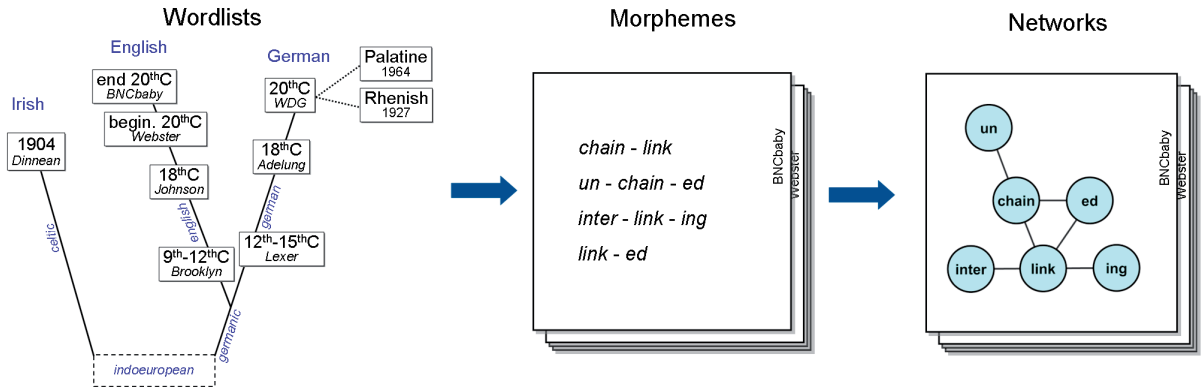

Figure 1: Generating morpheme networks. Left: Dictionaries and corpora data represented as a genealogical tree of languages. Middle: Wordlists with decomposed words. Right: Resulting network from decomposed wordlist.

In the first step, each word was broken down into its morphemes (Creutz et Lagus, 2005). Next, we created networks for each wordlist with the morphemes as nodes and an undirected edge between two morphemes occurring next to each other in a word (Figure 1). We investigated the global network properties, the properties of the morphemes and used pairwise comparisons to discover differences in time and space. Our results describe the underlying structure of word formation in terms of topology, give insights into the dynamics of language change, creation of new words and highlight cultural changes.

2. Material and Methods

2.1. Data sources

Headwords of eight dictionaries and lemmata of two corpora served as data for the network analyses.

The time aspect was investigated for the German by looking at the dictionaries of Lexer (Wörterbuchnetz, 2011), Adelung (Wörterbuchnetz, 2011) and WDG (DWDS-Projekt, 2011) representing the German language in the 12th to 15th, 18th and 20th century, respectively. The dialect dictionaries of the regions Pfalz (Wörterbuchnetz, 2011) and Rheinland (Wörterbuchnetz, 2011) originating both from the 20th century and were used for investigation of the regional

component. For the diachronic aspect of the English we took the Brooklyn corpus (Burnard, 2004a), the dictionaries of Johnson (McDermott, 1996) and Webster (Project Gutenberg, 2010) and the BNCbaby corpus (Burnard, 2004b) representing the 5th to 12th, the 18th and the beginning and end of 20th century, respectively. Dinneen's Irish dictionary (Nyhan *et al.*, 2009) originating from 1904 served as additional Celtic language.

Morphemes were identified by Morfessor 1.0 (Creutz et Lagus, 2005) with the default settings. The decomposition into morphemes was evaluated for Adelung and WDG by comparing the results of Morfessor 1.0 to a 1% sample of manually decomposed words in both dictionaries. 84.37% of the decompositions in WDG were correctly identified with a false positive rate of 15.63% and a false negative rate of 36.15%. In Adelung 85.64% of decompositions were correct with false positive rate 14.36% and false negative rate 27.44%. Morfessor 1.0 found 83% of the morphemes in WDG and 86% of the morphemes in Adelung.

2.2. Network analyses

For each dictionary a network was built out of the decomposed wordlist, where the morphemes are the nodes and two nodes are linked in the network if the morphemes occur side by side in a word. Multiple and loop edges were skipped and the edges were considered as undirected and un-weighted. Network analysis, calculations and graphics were performed in R 2.10.1 (R Development Core Team, 2009).

To describe the characteristics of the network, different topological measures were calculated based on the topological properties of the nodes. Interesting node properties were the degree (number of edges at a node), the shortest path length between two nodes, the clustering coefficient (fraction of interlinked neighbors of a node) and the assortativity value (mean degree of the neighbors of a node). As overall measures for the network the size, the mean degree, the mean path length and the mean clustering coefficient were calculated (Table 1). The mean path length and the mean clustering coefficient of an Erdős-Rényi graph of the same size served as comparison to a random network to check the small world property (Table 1). Looking at the degree distribution P(k) revealed the scale free property, if $P(k) \sim k^{-\lambda}$ and thus followed a power law. Another feature of the network is the hierarchical organization which was investigated by looking at the mean clustering coefficient depending on the nodes degree C(k). The dependency between assortativity value and degree shows assortative or disassortative mixing of the network, which was proved by calculation of the spearman correlation (Table 1).

	English				Irish	German			German Region	
	5th-12th century	18th century	begin 20th century	end 20th century	1904	12th-15th century	18th century	20th century	Palatine 1964	Rhenish 1927
	Brooklyn	Johnson	Webster	BNCbaby	Irish	Lexer	Adelung	WDG	PfWB	RhWB
n	6893	37588	45236	63077	28025	75540	54663	86129	80595	120419
N	1666	6547	7683	9544	3486	8493	7049	11256	9069	10716
E	6806	33410	42932	55910	24063	69437	50675	77817	76156	115662
k	8.17	10.21	11.18	11.72	13.81	16.35	14.38	13.83	16.79	21.59
L	3.05 (3.76)	2.99 (4.03)	3.01 (3.96)	3.00 (3.98)	2.74 (3.40)	2.73 (3.56)	3.04 (3.62)	3.11 (3.82)	2.88 (3.55)	2.73 (3.34)
C	0.22 (0.0041)	0.21 (0.0017)	0.22 (0.0015)	0.24 (0.0013)	0.31 (0.0042)	0.33 (0.0020)	0.18 (0.0020)	0.15 (0.0013)	0.21 (0.0018)	0.24 (0.0021)
R	-0.3508	-0.4403	-0.3785	-0.3531	-0.3278	-0.4706	-0.3494	-0.2866	-0.3485	-0.3844

Table 1: Key values of the networks. n number of entries in the wordlist, N and E number of vertices and edges in the network respectively, mean degree k, mean path length L and mean clustering coefficient C with random value (Erdös-renyi-model) in parentheses respectively and assortativity value R calculated as the spearman-correlation.

3. Results

3.1. Morpheme networks reveal word formation concepts

Considering the global architecture, all networks showed the same topological features, i.e. they were small world (Watts et Strogatz, 1998), scale free (Barabási et Albert, 1999), hierarchical (Ravasz *et al.*, 2002) and disassortativ (Newman, 2002). Despite this global similarity, there were profound differences in the details. The mean clustering coefficient and the mean degree (Figure 2) are sufficient to clearly separate between languages that exhibit low complexity in their word creating process (Irish, English and German of the 12th—15th century) and those with a higher word creating complexity (German of the 18th and 20th century together with the recent German dialects Palatine and Rhenish). This increase in complexity refers to a change of word formation in the evolution of German languages to more complex morpheme combinations. This evolution was observed for the fraction of German compounds among nouns shifting from 6.8% in the first half of the thirteenth century to 25.2% in modern German (Gardt, Hauss-Zumkehr, & Roelcke, 1999; Wellmann, Reindl, & Fahrmaier, 1974). Indeed, the complexity of word formation of present-day German is without structural analogy among the European neighbor languages (Erben, 2006). By analyzing network characteristics, we dated this fundamental step in the evolution of the German language between the 15th and the 18th century.

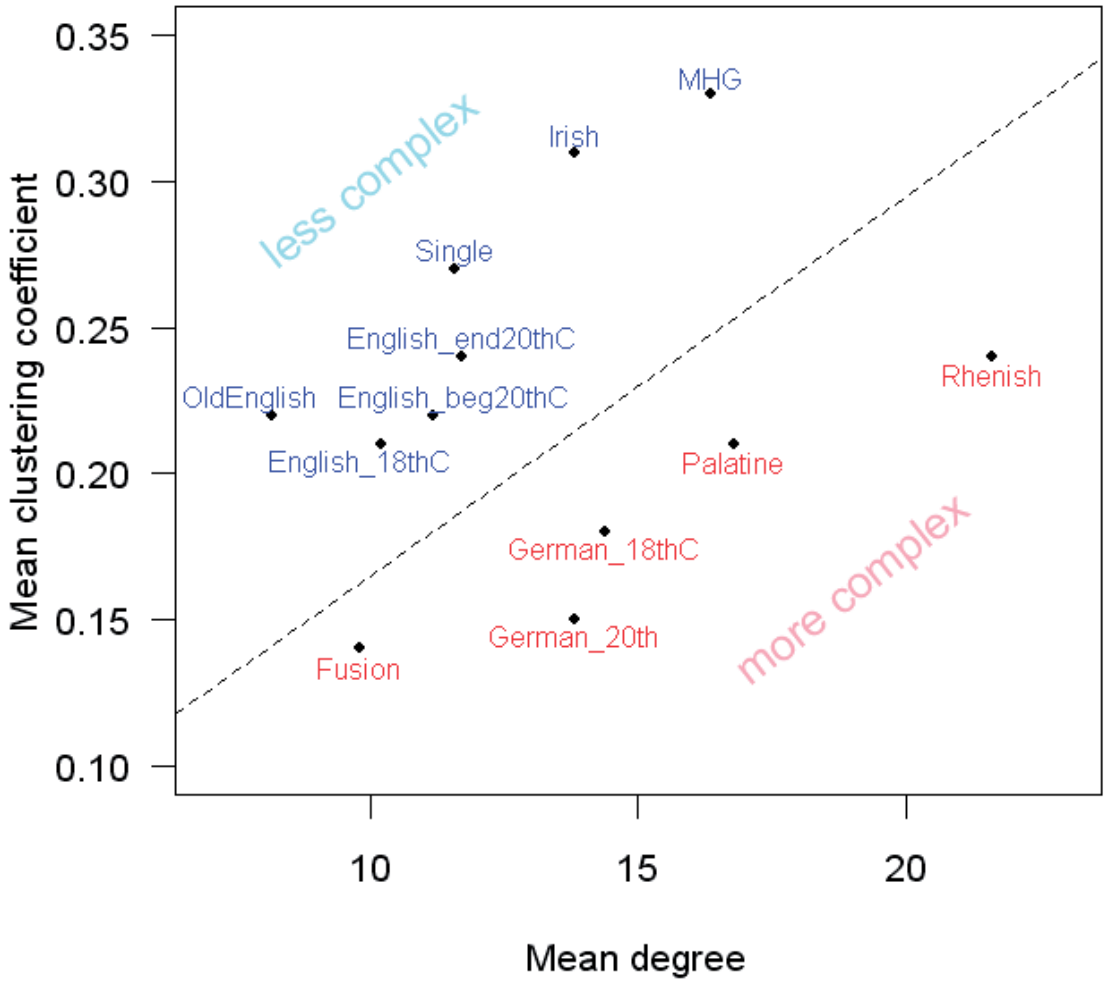

Figure 2: Complexity of word formation visualized by mean clustering coefficient dependent on mean degree. To strengthen this observation we investigated two versions of the English data: Fusion and Single denoted different encodings of the end 20th century English. Whereas in the Single dataset all hyphens in the entries were replaced by blanks and the resulting part of words were considered as separate entries in the word list, Fusion means that all hyphens and blanks are deleted. Thus, Fusion contains more complex words and is grouped with the newer German languages.

3.2. Hubs are markers for cultural change

A key feature of scale free networks is the existence of a small number of highly connected nodes, called hubs. These hub-morphemes are present in many different words. Hubs which are free morphemes should represent concepts important for the specific time or region (Table 2).

Hub-morphemes present in all dictionaries were `house / home', `water', `stone', `wood' and `man' indicating a common cultural background of the analyzed Germanic languages. In the Old English dictionary terms concerning politics were crucial (`law', `judgment', `noble') possibly indicating a bias in the written texts at that time. From the 18th to the 20th century there was a shift from morphemes describing nature to work-related morphemes happening independently in the English and German languages, highlighting cultural changes at that time. The German

dialect dictionaries are enriched with terms in the scope of nature and agriculture. Until today, these dialects are spoken more dominantly in rural areas and small villages.

	Time →			→			Region	
	German			English			German	
connectivity ↑	12th–15th century	18th century	20th century	9th–12th century	18th century	end 20th century	Palatine	Rhenish
	hûs (house)	stein (stone)	haus (house)	riht (law)	man	man	kopf (head)	kopf (head)
	man (man)	wasser (water)	wasser (water)	ræd (advise)	age	land	sau (sow)	sack (bag)
	stein (stone)	holz (wood)	zeit (time)	weard (mind)	sea	age	holz (wood)	stein (stone)
	meister (master)	baum (tree)	bau (building)	land (land)	stone	field	stein (stone)	holz (wood)
	waȝȝer (water)	berg (mountain)	arbeit (work)	æþel (noble)	wort	wood	haus (house)	mann (man)
	boum (tree)	kraut (herb)	holz (wood)	mod (spirit)	house	car	wasser (water)	wasser (water)
	wîn (wine)	eisen (iron)	dienst (service)	dom (judgement)	head	work	sack (bag)	kraut (herb)
	lant (land)	haus (house)	land (land)	dæg (day)	horse	head	kraut (herb)	apfel (apple)
	stat (state)	feld (field)	spiel (game)	leas (mistake)	work	time	acker (acre)	maul (vap)
	rëht (law)	geld (money)	werk (work)	aþ (oath)	wood	house	berg (mountain)	blume (flower)
	gëlt (money)	land (land)	tisch (table)	tun (estate)	corn	water	baum (tree)	arsch (ass)
	holz (wood)	meister (master)	welt (world)	mann (man)	master	stone	loch (hole)	loch (hole)
	îsen (iron)	fisch (fish)	haupt (head)	ende (end)	car	hand	wein (wine)	baum (tree)
	mer (sea)	feuer (fire)	hand (hand)	east (east)	hand	rate	bach (brook)	kuh (cow)
	wurz (herb)	hof (court)	stein (stone)	burg (city)	land	day	kuh (cow)	haus (house)

*Table 2: Most connected free morphemes comprising nouns— associated with nature, **associated with work.***

3.3. High variation in time and space

To identify evolutionary trends in the emergence and loss of morpheme complexes and morphemes, we mapped different networks onto each other using identical morphemes as anchors (Shou *et al.*, 2011).

Next, we analyzed how new words arise by fusion of already existing morphemes and how words get lost from the language although the composing morphemes stay part of it. An English example visualizing the rewiring process is shown in Figure 3.

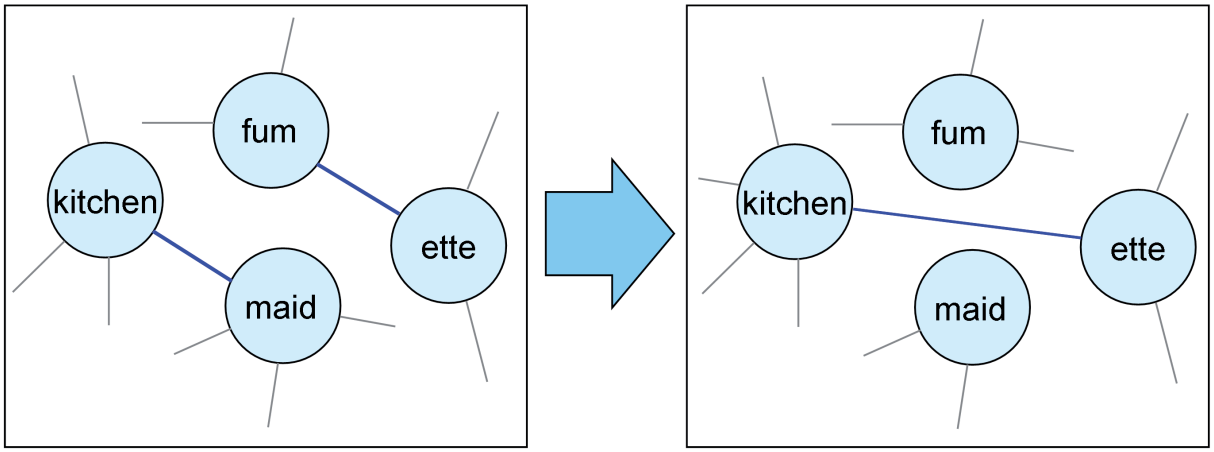

Figure 3: Typical rewiring example for English 18th to 20th century.

In total, from 54% to 75% of the edges were changed when comparing the networks of the same language over time. The highly connected morphemes maintained their high degree (Figure 4) even though switching usually 40% or more of their links. Thus, while hubs stay hubs, they interchange a lot of their edges.

Examples for the emergence of new words out of common morphemes in present German were `ab'-`gas' (waste gas), `druck'-`luft' (compressed air), `spott'-`billig' (dirt cheap) and `zig'-`mal' (dozens of times). Examples where the connection between still present morphemes was lost from the 18th to 20th century in German are `drossel'-`beere' (rowanberry), `butter'-`brühe' (butter—stock), `buß'-`stück' (millinery term). A cultural change, often associated with technical innovations, becomes obvious when looking at the rewiring of one morpheme over time: the morpheme `mond' (moon) forms in 18th-century-German the words `mondmilch' (mineralogy term) and `mondblind' (disease of horses) which do not appear in present German. Instead in present German there are the new words `mondflug' (moonshot) and `mondrakete' (moon—rocket). Although both dictionaries contain the morphemes `milch' (milk), `blind' (blind), `flug' (flight) and `rakete' (rocket). The same holds for `spülmagd' (dishwashing—maidservant) (18th) and `spülmaschine' (dishwasher) (20th).

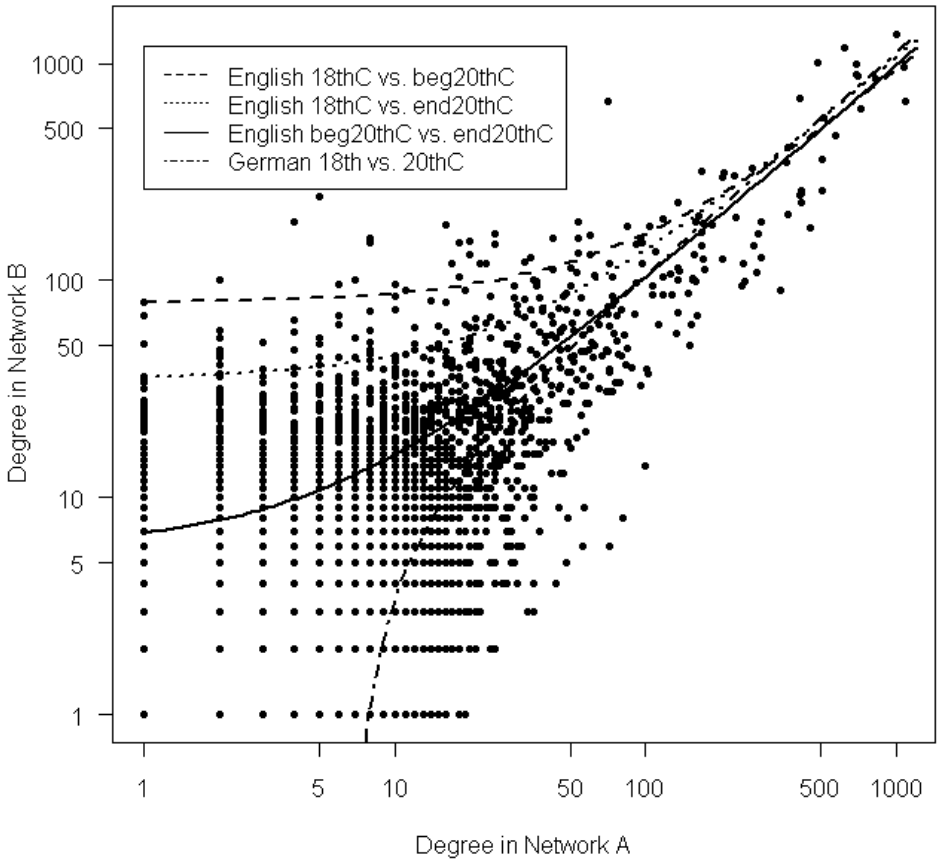

Figure 4: Plot of degree values in two compared networks. Dots show the data points of English beg. 20th vs. end 20th century. Lines correspond to the fitted linear models on hub-values for each comparison.

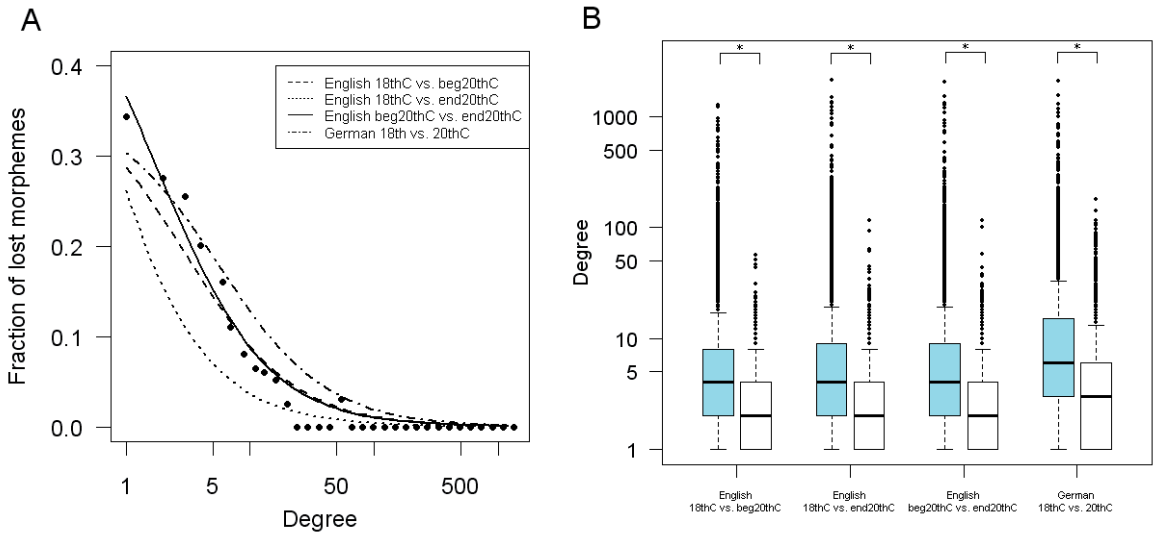

*Figure 5: Connectivity of lost and gained morphemes. **A** Dependence of loss of morpheme on degree. Fit of the function y=a/(x+b) with least squares. **B** Comparison of degree of common (grey) and gained (white) morphemes. * show highly significant difference.*

When comparing the networks of the same language over time between 14% and 45% of the nodes interchanged. The probability for a morpheme to get extinct depends highly on its degree. Typically, gained morphemes like `turbo' (turbo), `akku' (rechargeable battery) and `video' (video) were sparsely linked (Figure 5B). Highly connected ones have a lower probability of dying (Figure 5A). No differences between lost, gained and common morphemes were found for clustering coefficient and assortativity.

Thus, new morphemes enter the language first in a few different words. These morphemes are the most probable candidates to get extinct from the language. When looking for exceptions from these rules, we found gained morphemes with a high degree and lost morphemes in spite of a high degree. On the one hand these resulted from orthography change (i.e. `frey'=`frei' (free), `theil'=`teil' (part)). On the other hand morphemes which strongly mark a cultural change turn up (i.e. `film' (movie), `auto' (car), `sport' (sports), `kultur' (culture)) or get lost (i.e. `seiger' (mining term), `druse' (mineralogy term)).

4. Discussion

If one wants to analyze how a language changes over time, one obvious unit of observation is the word. But frequently a word can contain more than one meaningful part. Therefore, a word can die, but the parts it is composed of are still present in the language. Complementary, new words can arise by combining two morphemes which have already been present in the language. Thus, the analysis of morphemes, the smallest meaningful parts of words, can reveal on the one hand how words evolve and on the other how a meaning is integrated in a language.

Not surprisingly, we found that the morpheme networks for the analyzed Indo-European languages show the same global features. As many other biological networks they were small world, scale free and hierarchical. Despite these global similarities, the morpheme networks and thereby the languages differed in detail. We were able to separate the less complex languages (according to their underlying word forming mechanisms) from the more complex modern German language. Thus morpheme networks revealed an objective quantitative measure for the classification of word building mechanisms.

In more detail, we found that the re-wiring of existing morphemes is a major mechanism in the evolution of words. Interestingly, being a hub seems to be a consistent feature of a morpheme. Although strongly re-wired, a hub morpheme stays a hub morpheme over time. Additionally, these hub morphemes are the most resistant against loss. Most of the emergence and loss of morphemes happens at the border of the network. Thus, morpheme networks and their dynamics revealed principles underlying word formation and evolution in Germanic languages.

As we focused on the meaningful parts of the words, we were also able to uncover cultural changes buried in the language. When looking statically at a language, hub morphemes reveal concepts important for people at the given time and region. We re-covered a shift from nature- to work-related morphemes over time. More detailed, the emergence and loss of highly connected morphemes revealed important concepts getting extinct from or invading a language. The latter are mainly linked to new innovations important to today's culture like cars, radio or movies.

The existence of analogous structures between language and biology has been fiercely debated (and still is). Anyhow, when looking at morphemes and how they are combined to build words,

the analogy to domains and proteins suggests itself. In both cases, meaningful / functional building blocks are combined to create a larger meaningful / functional unit. In some cases, the meaning / function of the larger part might be completely explained by its building blocks. In many other cases, the whole can become more than the sum of its parts. Although knowing the words `eye' and `glance', the meaning of the word `Augenblick' (eye-glance) might not be obvious to non-German speakers (it denotes a very short time span, maybe translated as a `snatch'). Just as biologists might know that there is more to a protein than its domains, there is more to a word than its morphemes. Still the analysis of domains gave fundamental insights into the evolution of proteins. We therefore think that the analysis of morphemes will be fundamental in the understanding of the evolution of words.

Acknowledgements

We would like to thank Prof. Dr. W. Wegstein, Lehrstuhl für Computerphilologie, University Würzburg and Prof. Dr. A. Rapp, Germanistische Computerphilologie, University Darmstadt, for technical support and conceptual advice and Luise Borek and Esther Ratsch for generating the manually decomposed word lists. D. B. K. was financed by the BMBF Grant OIUA0815-C.

References

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509-512. doi:10.1126/science.286.5439.509

Burnard, L. (2004a). The Brooklyn Corpus of Old English on BNC baby CD v 1.0.

Burnard, L. (2004b). BNC baby CD v 1.0.

Copley, R.R., Ponting, C.P., Schultz, J., & Bork, P. (2002). Sequence analysis of multidomain proteins: past perspectives and future directions. *Advances in Protein Chemistry*, *61*, 75-98.

Creutz, M., & Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. *Publications in Computer and Information Science* (Report 81A., Vol. 81). Helsinki University of Technology.

DWDS-Projekt. (2011). Digitales Wörterbuch der deutschen Sprache.

Erben, J. (2006). *Einführung in die deutsche Wortbildungslehre* (5th ed.). Berlin: Schmidt (Erich).

Gardt, A., Hauss-Zumkehr, U., & Roelcke, T. (1999). *Sprachgeschichte als Kulturgeschichte*. Berlin: Walter de Gruyter.

Gesellschaft für deutsche Sprache. (2010). Press release 17.12.10.

Haspelmath, M., & Sims, A. (2010). *Understanding Morphology*. London: Hodder Education.

Janin, J., & Chothia, C. (1985). Domains in proteins: definitions, location, and structural principles. *Methods in Enzymology*, *115*, 420-430. Macmillian Magazines Ltd.

McDermott, A. (1996). Samuel Johnson, A Dictionary of the English Language on CD-ROM. University Press.

Michel, J.B., Shen, Y.K., Aiden, A. P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., *et al.* (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176. American Association for the Advancement of Science. doi:10.1126/science.1199644

Newman, M. (2002). Assortative Mixing in Networks. *Physical Review Letters*, *89*(20), 1-4. doi:10.1103/PhysRevLett.89.208701

Nyhan, J., Purcell, E., Hazard, B., McCarty, E., & O Halloran, M. (2009). Digital Dinneen. CELT: Corpus of Electronic Texts: a project of University College.

Project Gutenberg. (2010). The Project Gutenberg Etext of The 1913 Webster Unabridged Dictionary.

R Development Core Team. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Ravasz, E., Somera, a L., Mongru, D. a, Oltvai, Z.N., & Barabási, a L. (2002). Hierarchical organization of modularity in metabolic networks. *Science (New York, N.Y.)*, *297*(5586), 1551-5. doi:10.1126/science.1073374

Shou, C., Bhardwaj, N., Lam, H.Y.K., Yan, K.-K., Kim, P.M., Snyder, M., & Gerstein, M. B. (2011). Measuring the evolutionary rewiring of biological networks. (B. Snel, Ed.)*PLoS Computational Biology*, *7*(1), e1001050. Public Library of Science.

Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440-2. doi:10.1038/30918

Wellmann, H., Reindl, N., & Fahrmaier, A. (1974). Zur morphologischen Regelung der Substantivkomposition im heutigen Deutsch. *Zeitschrift für deutsche Philologie*.

Wuchty, S. (2001). Scale-free behavior in protein domain networks. *Molecular Biology and Evolution*, *18*(9), 1694-1702. SMBE.

Wuchty, Stefan, & Almaas, E. (2005). Evolutionary cores of domain co-occurrence networks. *BMC Evolutionary Biology*, *5*(1), 24. BioMed Central.

Wörterbuchnetz. (2011). Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier.