Complexity of European Union Languages: A comparative approach.

ARTICLE *in* JOURNAL OF QUANTITATIVE LINGUISTICS · MAY 2008

Impact Factor: 0.33 · DOI: 10.1080/09296170801961843 · Source: DBLP

Journal of Quantitative Linguistics
2008, Volume 15, Number 2, pp. 185–211
DOI: 10.1080/09296170801961843

Routledge
Taylor & Francis Group

Complexity of European Union Languages: A Comparative Approach*

Markus Sadeniemi[1], Kimmo Kettunen[2], Tiina Lindh-Knuutila[1] and Timo Honkela[1],
[1]Helsinki University of Technology, Finland; [2]University of Tampere, Finland

ABSTRACT

In this article, we are studying the differences between the European Union languages using statistical and unsupervised methods. The analysis is conducted in the different levels of language: the lexical, morphological and syntactic. Our premise is that the difficulty of the translation could be perceived as differences or similarities in different levels of language. The results are compared to linguistic groupings. Two approaches are selected for the analysis. A Kolmogorov complexity-based approach is used to compare the language structure in syntactic and morphological levels. A morpheme-level comparison is conducted based on an automated segmentation of the languages into morpheme-like units. The way the languages convey information in these levels is taken as a measure of similarity or dissimilarity between languages and the results are compared to classical linguistic classifications. The results have a significant impact on the design of (statistical) machine translation systems. If the source language conveys information in the morphological level and the target language in the syntactic level, it is clear that the machine translation system must be able to transfer the information from one level to another.

INTRODUCTION

The European Union is culturally and linguistically a much more diverse area than economically comparable areas, in particular the USA and

*Address correspondence to: Markus Sadeniemi, Helsinki University of Technology, E-mail: Markus.Sadeniemi@tkk.fi; Kimmo Kettunen, University of Tampere, E-mail: Kimmo.Kettunen@uta.fi; Tiina Lindh-Knuutila, Helsinki University of Technology, E-mail: Tiina.Lindh-Knuutila@tkk.fi; Timo Honkela, Helsinki University of Technology, E-mail: Timo.Honkela@tkk.fi

Japan. Over the years there have been many attempts to enhance the communication between European businesses and citizens through development of machine translation tools. A particularly important attempt was the Eurotra project that lasted from 1982 to 1993 and was funded by the European Commission (Durand et al., 1991). The first aim of Eurotra was to develop machine translation systems between the nine EC languages of that time (Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish). The second aim was to stimulate research in computational linguistics within the EC. Eurotra never delivered a working machine translation (MT) system but had a stimulating effect on research and development efforts. Eurotra was, however, heavily based on human-made linguistic descriptions. Even though high accuracy may be achieved through handcrafting the rules, the coverage and efficiency were limited. Even more importantly, the development time needed for an EU-level system became insurmountable. Since the Eurotra project, the EU has grown to include 21 official languages from the beginning of 2007. It has become clear that much more efficient means for MT system development are needed than the ones used during the Eurotra project.

In order to deal with the knowledge acquisition bottleneck of traditional MT development, many MT systems are nowadays being developed by applying some statistical machine learning techniques. This approach has proved to be useful and many current systems are hybrid systems in which the handcrafting and learning approaches are used in different combinations. It seems, though, that increased efficiency could still be achieved through more systematic use of statistical machine learning methods. This can be considered to be a hypothesis that also has strong opponents (consider, for example, Nirenburg & Wilks, 2000). We are, however, convinced that substantial developments are to be expected.

This article aims to provide some foundational information that could be used in the development of ''next generation'' learning machine translation systems. The basic idea is that one should be able to cover, for instance, 420 pairs of EU languages in the not-too-distant future.[1] This objective cannot be achieved unless the process of developing MT systems is substantially automated. In this article, we do not consider MT itself but rather analyse the complexity of EU languages. The analysis

[1]More information on this objective and related research at Helsinki University of Technology can be found at http://www.cis.hut.fi/research/compcogsys/

aims to support choosing the design principles and learning paradigms for the MT system. The basic insight behind the analysis is the following. Two languages that have a similar level of complexity when corresponding linguistic characteristics are considered are relatively easier to translate to each other than two languages that differ a lot. Moreover, the nature of the differences can also provide useful information for the MT system design. In the end, the kind of analysis reported in this article might serve as a preliminary phase in the creation of MT systems, for example, considering their parameterization.

Diversity of Languages and Emergent Features

As we are interested in the official EU languages, it is clear that despite the great variation in the appearances of languages they are essentially equally powerful to express whatever is needed. There are no fundamental problems in writing, for instance, the proposed EU Constitution in all 21 languages.

Of course there are differences in vocabulary and in how the terms divide up the conceptual space in different languages. A hundred years ago many concepts of our modern society were not needed and no corresponding words existed. On the other hand, words tend to be forgotten or acquire new meanings if they are not needed any more. Many new words have been created in EU languages because of EU administration. But this is normal development of language: in Europe, the spread of Christianity and development of science and technology have all made people make or borrow new concepts and words as needed.

Different languages, however, code the message in different ways. Words may be borrowed as they are or modified. In some languages morphology is rich and carries a lot of the information of the text. In many languages word order is fixed (to follow the pattern subject–verb–object, for example) while in others there is more flexibility. Articles, particles, prepositions or postpositions have an important role in some languages; in others, they are less important.

All this is basic material in traditional linguistics. In this research, however, we analyse automatically generated emergent features of the EU languages. The features we are interested in are sought for by unsupervised methods and the results are compared to traditional linguistic classifications of languages. As far as we know, unsupervised methods have not been used in the analysis of such a broad spectrum of languages so far.

Challenges of Multilingualism

To present an exact or even an approximate number of languages in the world is impossible: in some cases it may be difficult to state whether two similar languages are actually separate languages or two dialects of the same language. It is estimated that the number of languages in the world is in the thousands (Katzner, 2002), over 6000 being the usual figure mentioned (Haarmann, 2002; Gordon, 2005). Of these over a hundred are spoken in Europe alone. Currently there are 21 official languages in the European Union (including Irish from 1 January 2007). As much of the material produced by the Union has to be translated into all languages, the practical problems of translation are huge.

The problem gets only worse as new member states bring new languages to the Union. With the current 21 languages there are 410 pairs of languages to translate. It is evident that automatic translation would be of great help in many applications even if the quality were quite low.

Language Families in the EU

The 21 official EU languages are Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish (from 1 January 2007), Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Slovak, Slovenian, Spanish and Swedish.[2] Most of these belong to the Indo-European family of languages. One can divide the Indo-European EU languages into Germanic languages (Danish, Dutch, English, German and Swedish), Romance languages (French, Italian, Portuguese and Spanish), Slavic languages (Czech, Polish, Slovak and Slovene), Hellenic languages (Greek), Celtic languages (Irish) and Baltic languages (Latvian and Lithuanian) (Katzner, 2002). In the EU, only Estonian, Finnish, Hungarian and Maltese do not belong to the Indo-European language family. The first three are Finno-Ugric languages, and Maltese is a Semitic language.

The speakers of Germanic and Romance languages form a majority within the EU. In Table 1, the approximate numbers of native speakers of official EU languages are shown.[3] It can be seen that there are about

[2]The number of languages reflects the time of writing; since then the number has increased.

[3]Numbers of native speakers are from a web page of Intersol Inc, CA, USA, with an obvious error in the number of Greek speakers corrected; http://www.intersolinc.com/newsletters/newsletter_37.htm

Table 1. Approximate number of native speakers of official EU languages in Europe (in millions).

Language	Group	Speakers
German (ge)	Germanic (ge)	86
French (fr)	Romance (ro)	63
English (en)	Germanic (ge)	58
Italian (it)	Romance (ro)	56
Polish (pl)	Slavic (sl)	37
Spanish (es)	Romance (ro)	28
Dutch (nl)	Germanic (ge)	18
Hungarian (hu)	Finno-Ugric (fu)	11
Portuguese (pt)	Romance (ro)	10
Czech (cs)	Slavic (sl)	10
Greek (el)	Hellenic (ie)	10
Swedish (sv)	Germanic (ge)	8
Danish (da)	Germanic (ge)	5
Finnish (fi)	Finno-Ugric (fu)	5
Slovak (sk)	Slavic (sl)	5
Lithuanian (lt)	Baltic (ba)	3
Slovenian (sl)	Slavic (sl)	2
Latvian, Lettish (lv)	Baltic (ba)	1
Estonian (et)	Finno-Ugric (fu)	1
Maltese (mt)	Semitic (se)	0.3
Irish (ga)	Celtic (ce)	0.3

175 million native speakers of Germanic languages and 157 million speakers of Romance languages.

We study corpora in these 21 EU languages using statistical and unsupervised learning methods to describe the similarities and differences of the languages in different levels. Our intention is to compare these results with traditional linguistic categorizations like the division into language groups introduced above.

STATISTICAL METHODS FOR LINGUISTIC ANALYSIS

Methodological Choices

Traditional corpus-based linguistics has been hypothesis-driven. The researcher has made hypotheses on certain linguistic questions and then studied their validity using some available corpus. The problems have thus generally been fixed in advance and data is used only to decide

whether the initial guesses should be confirmed or rejected. More recently several studies have been data-driven. Text mining and statistical analysis are used to get insight on the language. It could be said that the text itself is given the opportunity to suggest new directions for research.

Both approaches have their merits, of course. Traditional methods make it possible to study linguistic phenomena in detail, whereas statistical methods often give a well-justified view on how language is usually used, and rare special cases are forgotten even though they may be interesting from the linguistic point of view.

It is also to be noted that the kind of corpus analysis we are conducting in this study does not provide specific information on, for example, conversational use of language or spoken language, in general. Moreover, some aspects of language and its use cannot be easily approached as they may require access to the spatial, social or cultural context of the linguistic expressions.

In this article we attempt to describe the EU languages by statistical methods and compare them to traditional language classifications. For this purpose two approaches are selected. The first method is to estimate the differences between languages using Kolmogorov complexity as a measure. This method is described in more detail in the section ''Compression method''. Our other approach takes into account information on the morphological level. For this task, we are using an automated way of segmenting words into morpheme-like units. This morpheme analysis is described in detail in the section ''Morpheme analysis''.

Emergence of Linguistic Representations

A lot of effort has recently been put into finding emergent linguistic representations. Many methods have been proposed and used; these depend on what level of language the analysis is focused on.

On the morphological level Creutz et al. (2005) and Creutz & Lagus (2005) have developed an unsupervised method of finding morphemes.

On the word sequence or syntactic level, hidden Markov models using bigrams or trigrams have commonly been used, at least since the development of the Candide system (Berger et al., 1994). A Markovian word sequence is not the only possible language model. Context-free grammars have been used although usually as given, not emergent, grammars. Another approach has been taken by the ADIOS project,

Automatic Distillation of Structure (Horn et al., 2004), where patterns in word sequences are sought.

On the syntactic and semantic level, Honkela et al. (2005) have used latent semantic analysis (LSA) (Deerwester et al., 1990), independent component analysis (ICA) (Hyvärinen et al., 2001) and self-organizing map (SOM) (Kohonen, 2001) to automatically extract linguistic relationships and features of words. Semantic maps of English words have also been created (Ritter & Kohonen, 1989; Honkela et al., 1995; Lagus et al., 2004).

Compression Method

Use of (file) compression as a measure for complexity is based on the concept of Kolmogorov complexity. Informally, for any sequence of symbols, the Kolmogorov complexity of the sequence is the length of the shortest algorithm that will exactly generate the sequence (and then stop). In other words, the more predictable the sequence, the shorter the algorithm needed and thus the Kolmogorov complexity of the sequence is also lower (Li et al., 2004; Bennett et al., 2005).

Kolmogorov complexity is incomputable (Li & Vitányi, 1997) but file compression programs can be used to estimate the Kolmogorov complexity of a given file. A decompression program and a compressed file can be used to (re)generate the original string. A more complex string (in the sense of Kolmogorov complexity) will be less compressible.

Compression has been used for different purposes in many different areas. Juola (1998) introduces comparison of complexity between languages on the morphological level for linguistic purposes:

> By selectively altering the expression of morphological information, one can measure the amount of morphological complexity contributes [*sic*] to a corpus by measuring the change in perceived informativeness.

Juola's method is rather straightforward: after randomization of the morphological level, the size of the original compressed file is divided by the size of the altered compressed file. The resulting ratio is taken as a measure of the morphological complexity of the language in question. With the same procedure of systematic distortion other levels of language can also be analysed (Juola, 2008).

Morpheme Analysis

Instead of analysing morphological information with compression analysis as described above, we can divide words into morphemes and analyse how languages differ in this respect.

Morphemes are the smallest individually meaningful elements in a language. How far languages code information in the morphological level varies: some languages, such as Turkish, Finnish or Estonian, are morphologically rich while others are poorer. For the analysis on the morphological level, we are using the Morfessor method (Creutz & Lagus, 2005) which produces an automatic segmentation of a corpus into elements that often resemble linguistic morpheme segmentation.

Morfessor is designed for languages with concatenative morphology. The number of morphemes per word does not have to be known in advance. The current Morfessor software implements a so-called baseline model, which uses a probabilistic maximum *a posteriori* formulation (MAP). The aim is to find an optimal model of language that produces a segmentation of the corpus. As this segmentation is often similar to the linguistic morpheme segmentation, the authors of the Morfessor have decided to call these segments morphs. A similar approach is taken here.

In the basic case, a model of a language consists only of a morph vocabulary (a lexicon of morphs). The lexicon contains the written form of the segmented morphs and their frequencies, i.e., the number of occurrences of that morph in the corpus. Every word form that is in the corpus can be represented as a sequence of the segmented morphs in the lexicon. In the MAP modelling of the language, the most probable word segmentation of all possible ones is chosen. In the search process, a greedy algorithm is used. Initially, each word in the corpus is a morph of its own. As the segmentation proceeds, different morph segmentations are proposed and the segmentation that yields the highest probability is selected. This is continued until no further improvements are made (Creutz & Lagus, 2005).

Our analysis was based on automatic morph segmentation of words in each language. The number of morphs indicates whether a language has rich morphology, but that is not the only information that can be extracted. Do the words have a long morph followed by short ones or the other way round? Are there two long morphs indicating perhaps compound words?

RESULTS

Data and Pre-processing

In this article, we are using the European Constitution corpus that is available in 21 European Union languages in the Open Source Corpus[4] free of charge. The languages covered in this corpus are Czech (cs), Danish (da), Dutch (nl), English (en), Estonian (et), Finnish (fi), French (fr), German (de), Greek (el), Hungarian (hu), Irish (ga), Italian (it), Latvian (lv), Lithuanian (lt), Maltese (mt), Polish (pl), Portuguese (pt), Slovak (sk) , Slovenian (sl), Spanish (es), Swedish (sv).

All corpus files have been UTF-8 coded. The total number of files is 987, i.e. 47 files in each of the 21 languages. The total number of tokens in the corpus is 3,099,290. The original files have been automatically XML-tagged to include, for example, sentence and word boundary information.

Table 2 lists an example phrase from the corpus in each of the 21 languages. The number of tokens in each language in the original files is given in Table 3.

For the analysis some simple pre-processing was needed. The XML-tags were removed as well as numbers and each token was put on a row of its own. All words were in lowercase characters.

Compression Approach

Information Content Carried by Morphology and Word Order

An experiment was made to test the ideas presented above along the lines described in Juola et al. (1998). The data used was the text of the EU Constitution in 21 EU languages.

The text of each of the 21 languages was pre-processed as described above. Then two modifications were made to the cleaned texts, one on the morpheme level and another on the word order level.

There were between 10,000 and 20,000 different word forms in each of the languages. In the first modification each word was replaced by a random number in the range 10,000 – 30,000. So each occurrence of the word ''competence'' was replaced by the same number in the English text but had no relation to the number representing competences.

[4]http://logos.uio.no/opus/EUconst.html

Table 2. An example phrase from each of the EU languages.

cs	Smlouva o ústavě pro evropu
da	Traktat om en forfatning for europa
de	Vertrag über eine Verfassung für Europa
el	Συνθήκη για τη θέσπιση Συνταγμάτος Ευρώπης
en	Treaty establishing a Constitution for Europe
es	Tratado por el que se establece una constitución para Europa
et	Euroopa põhiseaduse leping
fi	Sopimus euroopan perustuslaista
fr	Traité établissant une Constitution pour l'Europe
ga	Conradh ag bunú Bunreachta don eoraip
hu	Szerződés európai alkotmány létrehozásáról
it	Trattato che adotta una Costituzione per l'Europa
lt	Sutartis dėl Konstitucijos Europai
lv	Līgums par konstitūciju eiropai
mt	Trattat Li Jistabbilixxi kostituzzjoni għall-Ewropa
nl	Verdrag tot vaststelling van een grondwet voor europa
pl	Traktat ustanawiają Konstytucję dla europy
pt	Tratado que estabelece uma Constituição para a Europa
sl	Zmluva o ústave pre Európu
sk	Pogodba o ustavi za evropo
sv	Fördrag om upprättande av en konstitution för europa

Another modification was made to the cleaned text. In this case the words in each sentence were shuffled to a random order. (The ending punctuation was kept in its place.)

For comparison of the complexity of the languages three files were compressed using the *bzip2* program. The sizes of the modified compressed texts were then compared with the original compressed one to get a measure of the change of information, when morphological and word-order information in the texts were destroyed. Table 4 shows figures of the compressed language files.

From these figures we made three different relational analyses in the style of Juola (1998). In Figure 1 we have the morphological complexity of the languages shown as the relation between columns B and C of Table 4 (B/C) sorted in ascending order.

A few comments on Figure 1 are necessary. Most of the results are as expected; the morphologically simple languages, Italian, English, Irish, French, Portuguese and Spanish get low scores and the morphologically more complex languages, Finnish, Hungarian and Polish, are at the other end of the scale. But some of the results are unexpected: Slovenian,

Table 3. Number of tokens in original files.

Language	Files	Tokens
cs	47	127,542
da	47	149,211
de	47	144,908
el	47	162,875
en	47	164,697
es	47	177,322
et	47	114,784
fi	47	113,698
fr	47	177,162
ga	47	170,820
hu	47	146,007
it	47	162,594
lt	47	125,859
lv	47	134,036
mt	47	146,076
nl	47	167,945
pl	47	137,191
pt	47	165,435
sk	47	128,428
sl	47	142,986
sv	47	139,714
Total	987	3,099,290

Slovak, Latvian, Czech and Estonian should be higher on the complexity scale. Dutch, Swedish, Danish and German also get quite high values, German being even at the top of the scale. It is probable that compound words cause this effect. Also the type of texts, legalese, could have a boosting effect on the complexity of German.

In Figure 2 we show the morphosyntactic complexity of the languages by adding columns C and D together and dividing the figure from column B of Table 4 with the result, $B/(C+D)$.

In Figure 3 the syntactic complexity of the languages is shown as a relation of columns B and D from Table 4 (B/D). The two languages listed as syntactically quite complex are Finnish and Estonian, which are characterized by a very flexible word order.

Figure 4 shows the languages plotted on a two-dimensional graph using the variables morphological and syntactic complexity (B/C and B/D). Romance languages are grouped neatly into one corner of the

Table 4. Compression results of the files. A = language, B = original (cleaned) compressed file (bytes), C = words replaced by random numbers, file compressed (bytes), D = words of sentences shuffled to random order and file compressed (bytes).

A	B	C	D
cs	158,606	145,956	206,540
da	156,115	138,097	215,904
de	169,236	145,144	224,822
el	181,890	158,274	249,777
en	149,490	141,982	217,175
es	161,700	152,196	239,311
et	151,050	137,791	193,037
fi	161,067	138,409	203,658
fr	160,846	151,428	243,122
ga	168,550	159,304	245,621
hu	168,831	147,829	228,542
it	160,627	152,720	234,036
lt	157,123	145,381	206,011
lv	151,512	140,713	202,518
mt	165,988	149,947	230,652
nl	169,179	151,200	237,162
pl	168,857	148,408	221,580
pt	157,958	147,963	230,835
sk	166,421	149,307	216,623
sl	153,428	145,154	215,130
sv	156,210	138,832	209,294

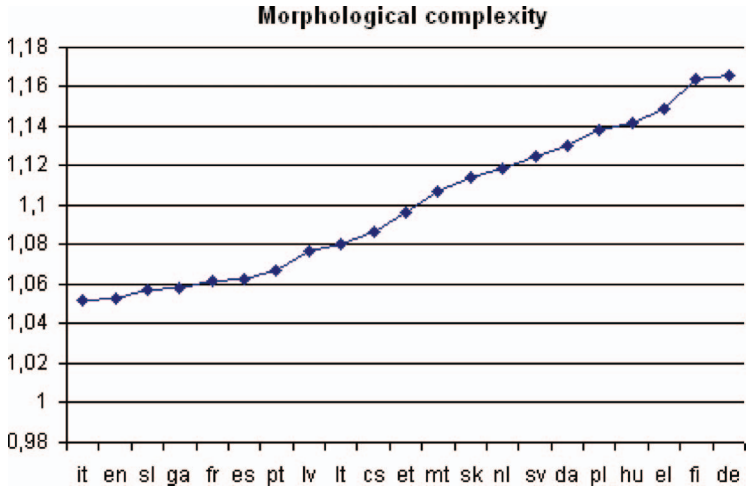

Fig. 1. Morphological complexity of the languages analysed with compression.

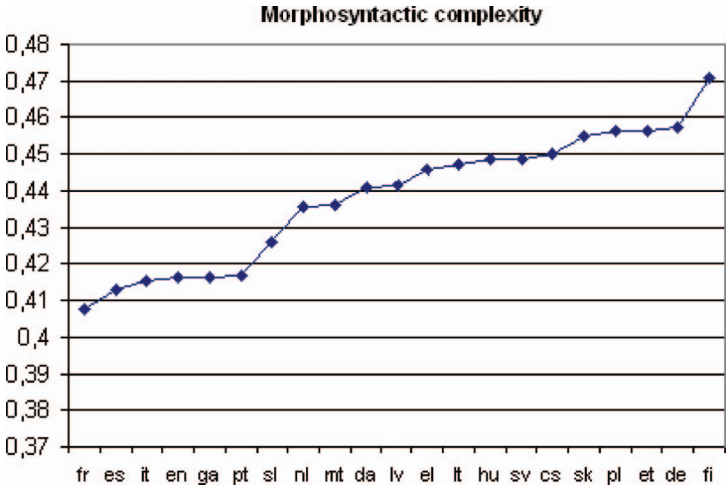

Fig. 2. Morphosyntactic complexity of the languages analysed with compression.

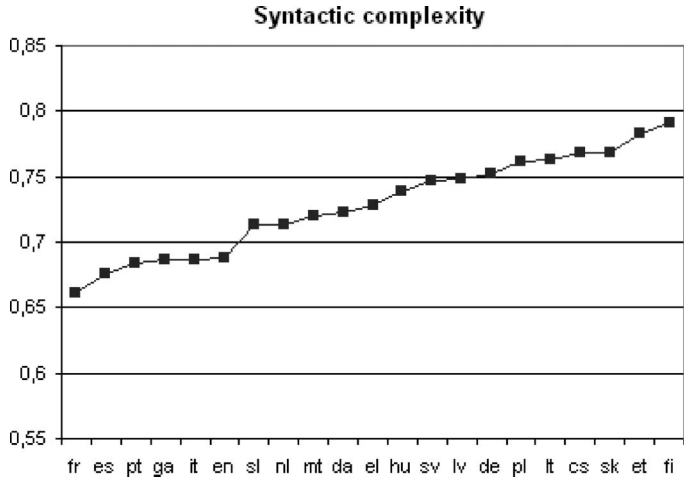

Fig. 3. Syntactic complexity of the languages analysed with compression.

picture and seeing English near them is no surprise. Finnish and Estonian are located on the right of the figure. Baltic and the other Slavic languages are generally more on the lower right side than the Germanic languages although the separation is not very clear.

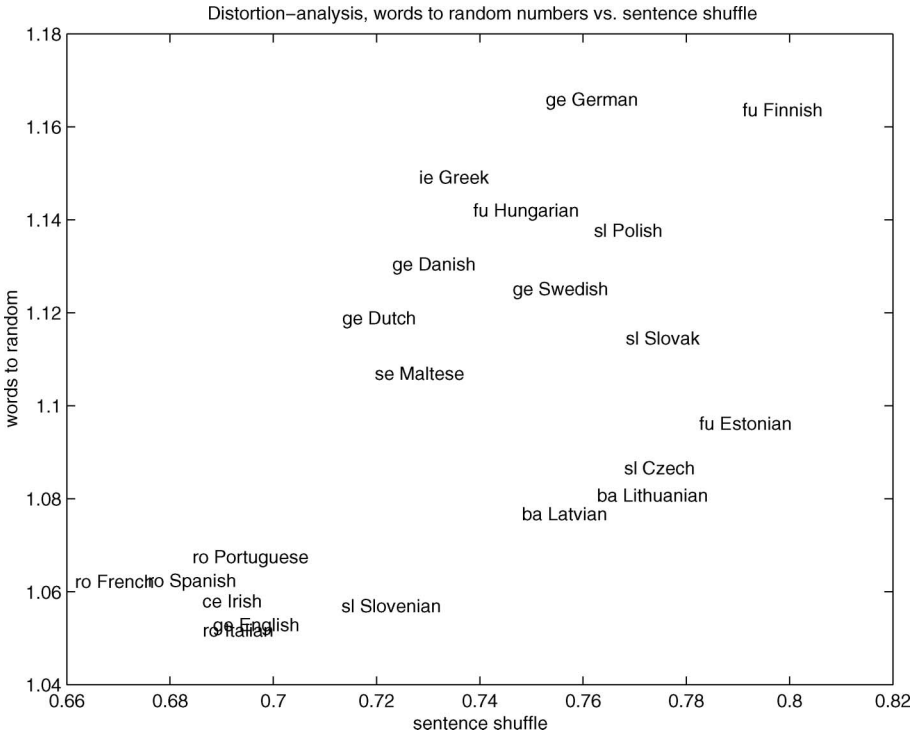

Fig. 4. Language map: morphology vs. word order information.

Overall the results are as expected: Finnish and Estonian have quite free word order, Finnish and German have compound words and a complex morphological structure of lexemes, whereas Romance languages and English are at the other end of scale.

It must be remembered, of course, that when talking of word order we not only mean SVO-like grammatical structures but also such trivial things as particles preceding nouns.

Figure 5 shows languages on a self-organizing map (SOM) (Kohonen, 2001). Input variables in this picture are the three compressed file sizes as such. A SOM is a nonlinear projection from the original feature space to a two-dimensional map. This is done in such a way that observation – here languages – close in original space remains close on the map. Longer distances do not remain proportional, however.

The map in Figure 5 shows Romance languages well clustered again and English near them. Danish and Swedish are close to each other,

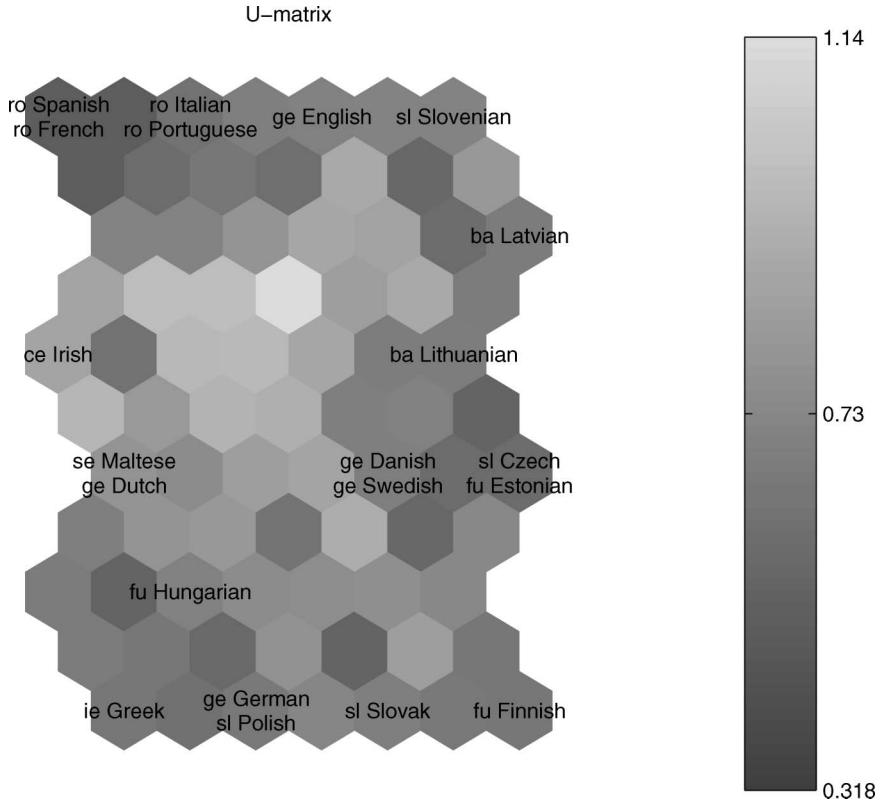

Fig. 5. Languages in SOM: morphology vs. word order information.

as they should be, but Estonian should be near Finnish rather than Czech.

Interpretation of Morphosyntactic and Syntactic Complexities

The morphosyntactic complexity of the languages in Figure 2 is partly as expected and partly not. Most of the languages at the complex end of the scale are as expected; Finnish, German, Estonian, Polish, Slovak, Czech and Hungarian being at the top. Only Swedish seems to be higher on the scale than expected, and Latvian and Slovenian lower than expected.

To get a meaningful interpretation for the order of languages in the word-order complexity counting, linguistic literature was consulted for independent figures.

Bakker (1998, p. 387) introduces the flexibility of a language's word order, which is based on 10 factors such as order of verb and object in the language, order of adjective and its head noun, order of genitive and its head noun, etc. Altogether Bakker has seven constituent-level variables and three clause-level variables in his flexibility counting, and thus constituent-level variables are more important for the result. The flexibility of the language can be given with a numeric value from 0 to 1: if the flexibility figure is close to zero, the language is more inflexible in its word order, if the figure is closer to 1, the language is more flexible in its word order. In the information theoretic framework of the compression approach, flexibility and inflexibility can be interpreted naturally as higher and lower degrees of complexity; that is predictability.

In Table 5 figures based on Bakker's (1998, pp. 417–419) counting of the flexibility values for the individual languages are given together with values given by compression.

If we compare the figures given by Bakker in column 2 to figures given by the compression-based calculation in column 4, we can see that the overall order of the languages based on these independent calculations converge well. There are also some differences in the orders given by the two analyses. The syntactic complexity of Lithuanian seems to be estimated higher by compression than by Bakker's flexibility value (16 versus 8). Slovenian has also a higher flexibility value than its complexity value (14 versus 7). Greek is also higher in Bakker's counting than in complexity analysis (17 versus 11). In our compression calculations Finnish and Estonian are estimated as almost equally complex, but in Bakker's analysis Estonian is less complex than Finnish (18 versus 13). The lower end of the scale is quite analogous in both analyses, consisting of the five languages with differences in the order.

Measuring Morphological Similarities Between Languages by Compression

Another method for comparing the similarity of languages using compression was described by Cilibrasi and Vitányi (2005). Again the size of a compressed text file is used to measure its Kolmogorov complexity as described in Li et al. (2004).

A compression program (here *bzip2*) learns the characteristics of a language as it processes the text. If the language of the text changes

Table 5. Bakker's flexibility values for languages with compression relation complexity of the word order. Czech and Hungarian have been omitted from the table, as Bakker is missing data for them. The compression figures for these languages are 0.74 (Hungarian) and 0.77 (Czech).

Order of the languages based on Bakker's flexibility calculation			Syntactic complexity order of the languages based on compressional relation calculations from Figure 3		
		Bakker's flexibility value			Complexity figure based on compression
1	fr	0.10	1	fr	0.66
2	ga	0.20	2	es	0.68
3	es	0.30	3	pt	0.68
4	pt	0.30	4	ga	0.69
5	it	0.30	5	it	0.69
6	da	0.30	6	en	0.69
7	mt	0.30	7	sl	0.71
8	lt	0.30	8	nl	0.71
9	en	0.40	9	mt	0.72
10	nl	0.40	10	da	0.72
11	de	0.40	11	el	0.73
12	sv	0.40	12	sv	0.75
13	et	0.40	13	lv	0.75
14	sl	0.50	14	de	0.75
15	lv	0.50	15	pl	0.76
16	sk	0.50	16	lt	0.76
17	el	0.60	17	sk	0.77
18	pl	0.60	18	et	0.78
19	fi	0.60	19	fi	0.79

in the middle of processing the compression program has to adapt to the new situation. If the languages are different, it has to unlearn the efficient coding of the first language and learn the characteristics of the new language. On the other hand, if the languages are similar enough, it can use the old coding with perhaps small modifications.

Thus the similarity of languages can be measured by how well the compression manages this transition. In mathematical terms we can mark the size of a compressed text file in language x by $C(x)$ and in y by $C(y)$

and by $C(xy)$ the size of the compressed file for concatenated text xy. The distance measure used here is

$$\frac{C(xy) - C(x)}{C(y)}$$

which measures the change in compressing language y when using x as model. Li et al. (2004) recommends the expression

$$\frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))}$$

instead. The latter expression has the benefit that it is symmetrical: the distance between languages x and y is the same as between y and x. The former expression acknowledges the possibility that the relation can be asymmetrical: perhaps x is better explained by y than vice versa.

Figure 6 shows languages as they appear on a SOM. The results are in many ways problematic. Romance languages are at the lower right corner and English at the upper right, but Hungarian and Maltese being near French is not so logical. At the upper left there are Czech, Slovenian, Latvian and Lithuanian, but Estonian and Greek should not be in the same group.

Morpheme-Level Comparison

Method

For morpheme-level comparison of languages the Morfessor, presented earlier in the section ''Morpheme analysis'', was used to automatically create a morpheme-like morph segmentation for each of the 21 languages. For several of these languages, this is the first time Morfessor has been applied to them. The data used was the European Constitution corpus files slightly pre-processed as before: all words were in lowercase and the XML-tags were removed. Additionally, all numbers were removed.

We calculated various key figures from each of the language data. These included (i) the number of different word forms (types) in the given corpus, (ii) the total number of words in the corpus (tokens), (iii) the average number of morphs per word, (iv) the mean and (v) the variance as well as the (vi) skewness and (vii) kurtosis of the morph length distribution for the given corpus.

The 7×21 data matrix obtained was then normalized based on the variance, and fed to the self-organizing map for further analysis.

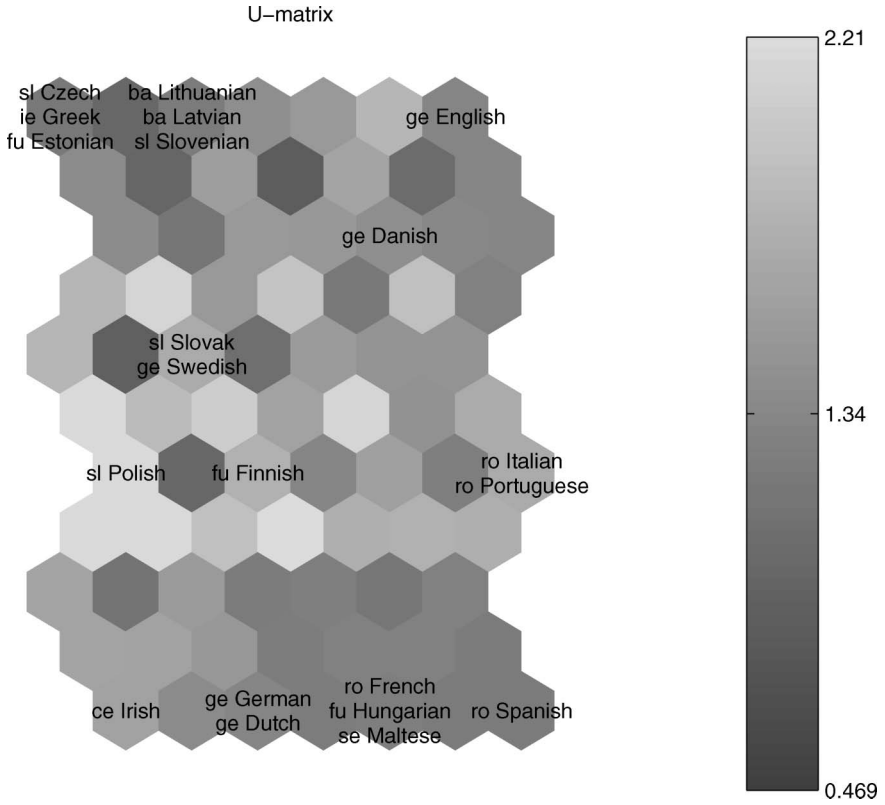

Fig. 6. Language similarities by pairwise compression.

Language Clustering by Morph Features

Figure 7 shows the self-organizing map obtained. Figure 8 illustrates how the values of the various features are distributed in the map. The first variable, *nwrds*, is the number of different word forms (types) in the given data. The second variable, *ntot*, is the total number of words (tokens). The third variable, denoted as *avrparts*, is the average number of morphs per word. The fourth, fifth, sixth and seventh variable, *avrlen*, *avrvar*, *avrskew* and *avrcurt* show the average mean, variance, skewness and kurtosis, respectively, of the morph length distribution of the given data.

Figure 7 shows that the Romance languages form a cluster in the upper right corner of the U-matrix. English is among the Romance languages

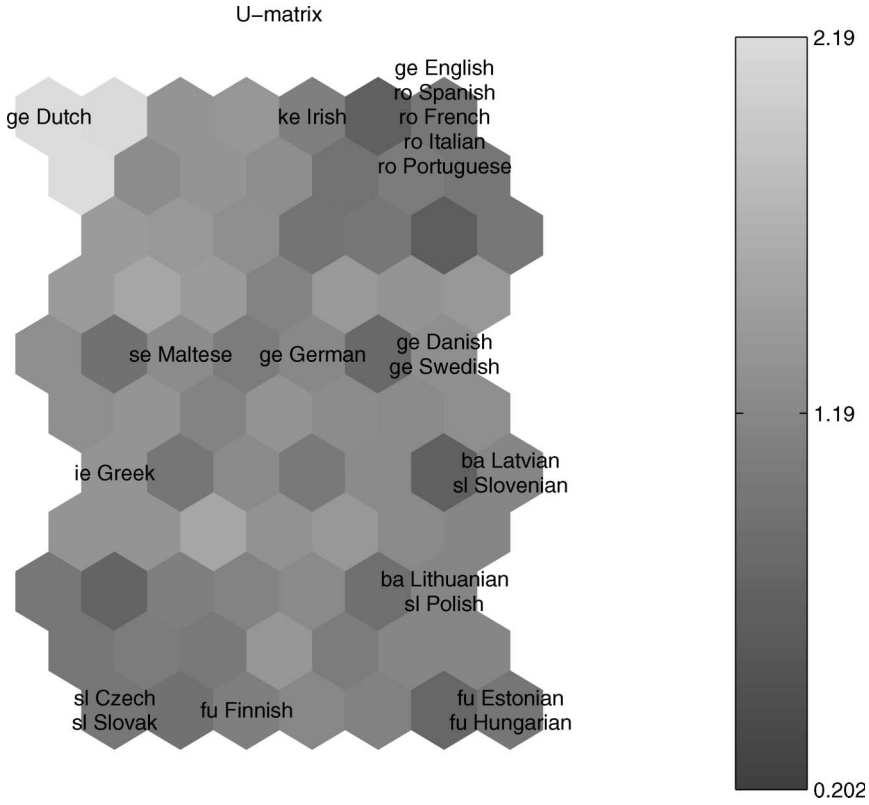

Fig. 7. Official EU languages on the SOM U-matrix.

although it is usually classified as a Germanic language. This is no surprise as English has, since the Middle Ages, derived many words from French and Latin (Stockwell & Minkova, 2001).

Another distinct group is formed by the Finno-Ugric languages: Estonian, Hungarian and Finnish. The Germanic or Slavic languages do not form such clearly distinct groups but, for example, the closeness of Swedish, Danish and German, as well as of Czech and Slovak, is evident.

What more can be inferred by looking at the variable distributions? When we look at Figure 8, it seems that for separating the Romance languages from the Finno-Ugric languages, the most prominent features are *nwrds* and *ntot*; the number of words and the number of different words. For some reason Dutch seems to be separated from others by *avrcurt*.

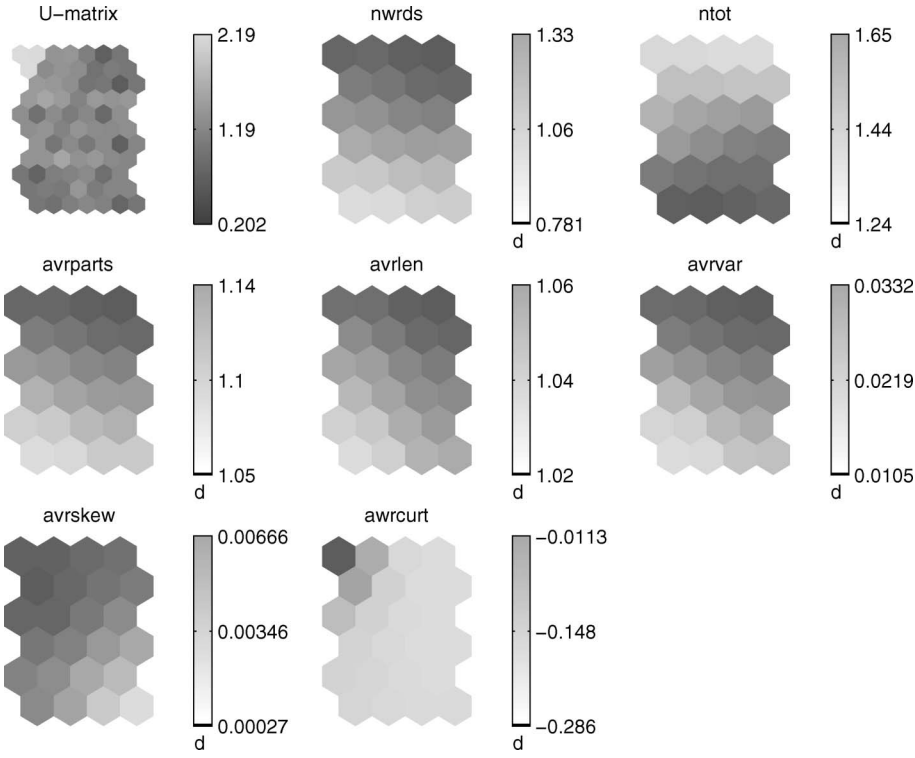

Fig. 8. The distribution of the variables on the map; see text for the explanation of the variables.

Figure 9 shows languages sorted in ascending order of the average number of morphs per word. Although based on different calculations, these results fit well with those of the section ''Compression approach''; rich morphology is naturally also seen as a high number of morphs in a language.

Figure 10 shows languages plotted using two morphological features: ''variance'' is a measure that is small for words with a small number of morphs or when a long morph is surrounded by short ones. Compound words tend to make this feature larger. The other feature, ''skewness'', is big for words with a long morph followed by short ones as is typical for languages with concatenating inflection. Romance languages, again with English, are well-clustered, as well as Finno-Ugric languages. Even Germanic and Slavic languages are separated reasonably well.

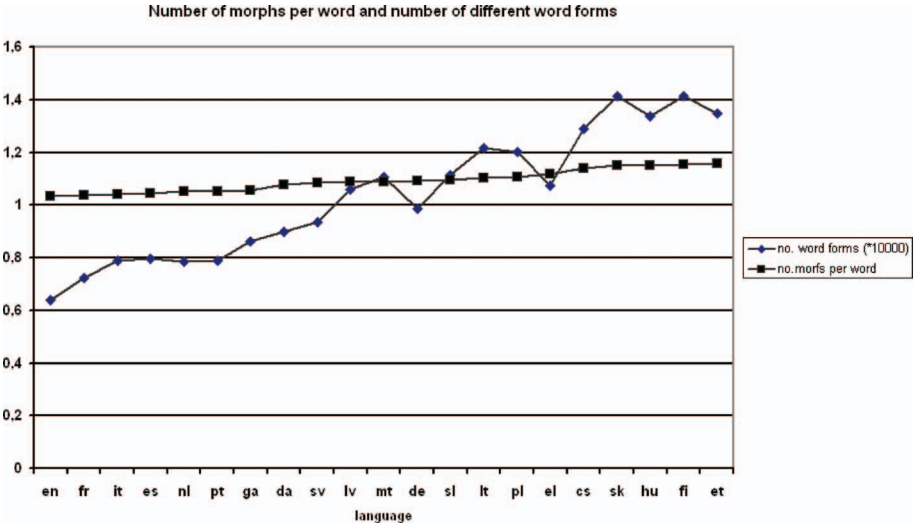

Fig. 9. Average number of morphs per word and number of word forms in each language.

DISCUSSION

The Success of Different Methods

Compression

We used the compression program (*bzip2*) to estimate the Kolmogorov complexity of texts in 21 EU languages modified in several ways. The method of modifying information on different levels of language has been proposed by Li and Vitányi (1997) and Juola (1998), and proved to be useful. We obtained estimates of morphological and syntactic complexity that seemed to group languages in a logical manner.

Another way of using compression was to use compression results of one language to compress another language, as proposed by Cilibrasi and Vitányi (2005). The results we got with this method were less convincing.

It is understandable that the compression method works better when comparing two versions of text in the same language than in the case where two languages are compared. Modifying either the morphological level of a language or shuffling the word order are both well-defined transformations that change the information content in a controllable way. Whether two languages look similar for a compression algorithm is

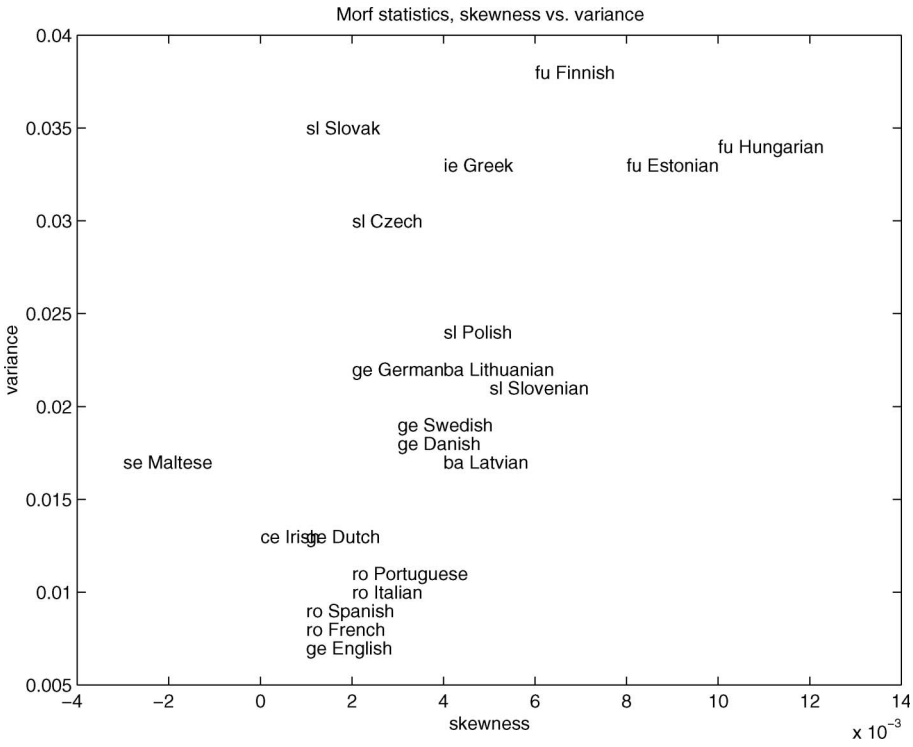

Fig. 10. Average "skewness" and "variance" of morph lengths.

more a matter of chance. The analysis works on the surface level only and even small changes of orthography can change the results drastically.

Morphological Analysis

The Morfessor had not been used on many of the EU languages previously. It is particularly designed for languages with concatenative morphology and had so far been tested for Finnish, English, Turkish and Swedish. We did not make a thorough analysis of how well the segmentation found (the morph set) corresponded to the linguistic morphemes of each language. By inspecting the given segmentation, we could see, however, that in many cases too few morphs were found. Probably the amount of text was not enough for Morfessor. Additionally, the source texts, being mainly legalese, use less of some morphological forms (e.g. verb conjugation) but compound words may correspondingly be longer and more complex in this type of text. Repeating the experiments

using a similar parallel corpus from a more general language domain would be very interesting for gaining insight into this.

In summary, the results gained with the current approach were already interesting and useful. The features that were calculated based on the morphs clearly separated the languages into meaningful groups.

Relation to Traditional Linguistics

The results given by the compression method were well in line with the results given by Bakker (1998), although we used unsupervised statistical methods while Bakker based his analysis on more classical linguistic methods.

Information carried by morphology on the one hand and by word order on the other was observed by the compression method. The method is clearly useful.

Both the compression method and morph analysis could be used to group languages into their language families. Romance languages and English formed a clear group, as did Finno-Ugric languages. Germanic and Slavic languages were somewhat mixed, but morph analysis could separate even these quite well.

Problems in Machine Translation

It is well known that machine translation efforts have had only a limited success. The best results have been achieved in cases where the vocabulary and semantic environment has been very limited, like in weather forecasts (Sigurd et al., 1992).

Reasonably good results have been achieved when languages have been close enough like English and French (Al-Onaizan et al., 1999). Some success has also been reported between English and Chinese, which obviously are not close, but both are poor in morphology (Och et al., 2004; Al-Onaizan et al., 1999).

It can be expected that it would be easier to translate between languages that are linguistically close, but for machine translation it might additionally help to have similarities in vocabulary. Within EU texts there are many words that are similar or at least have only one possible translation. EU jargon-words like ''subsidiarity principle'' are typical examples.

We showed how different languages carry information on different levels of the language. For some languages morphology is important, and for some word order. Division of words into morphemes, or in our case,

automatically extracted morphs, indicates either word inflection or word formation including compound words.

This has implications for machine translation as the information carried in different ways and on different levels of a language has to be translated and carried over to the other language concerned.

CONCLUSION

In this paper we have used a file compression program and subtle text transformation to analyse the complexity of 21 official EU languages. In addition we used the Morfessor system to extract morphs in an unsupervised way and analysed the results.

The results are naturally fairly coarse, but they are reasonably well in line with Bakker's flexibility values for languages. We have also projected the languages on a two-dimensional graph in such a way that languages fall into their natural linguistic groups.

So where could these information-theoretical results be applied? One suggestion is to use them when building statistically oriented machine translation systems. The basic idea is that the translation process can be divided into interrelated tasks following, for example, the classical machine translation triangle model. In this case, however, we foresee that all those tasks can be conducted using statistical methods. For instance, a detailed morphological analysis can be made using the unsupervised learning method (see Creutz et al., 2005, for example) when needed. For some languages, a detailed morphological analysis is not needed. Similarly, for some language pairs one may need to pay special attention to the word order whereas for some other language pairs it may be assumed that the word order in them is rather similar. This assessment influences the complexity of the statistical model needed.

REFERENCES

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F., Purdy, D., Smith, N. A., & Yarowsky, D. (1999). Statistical machine translation. *Technical Report*. John Hopkins University (*Final Report*, JHU Workshop 1999).

Bakker, D. (1998). Flexibility and consistency in word order patterns in the languages of Europe. In A. Siewerska (Ed.), *Constituent Order in the Languages of Europe. Empirical Approaches to Language Typology* (pp. 381–419). Berlin: Walter de Gruyter.

Bennett, C. H., Gacs, P., Ming, Li, Vitanyi, M. B., & Zurek, W. (1998). Information distance. *IEEE Transactions on Information Theory*, *44*, 1407–1423.

Berger, A. L., Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Gillet, J. R., Lafferty, J. D., Mercer, R. L., Printz, H., & Ures, L. (1994). The Candide system for machine translation. *Proceedings of the 1994 ARPA Workshop on Human Language Technology*, Yorktown Heights, NY 10598.

Cilibrasi, R., & Vitnyi, P. M. B. (2005). Clustering by compression. *IEEE Transactions on Information Theory*, *51*(4), 1523–1545.

Creutz, M., & Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora. *Technical Report No. A81*. Helsinki University of Technology, Espoo: Laboratory of Computer and Information Science.

Creutz, M., Lagus, K., Lindén, K., & Virpioja, S. (2005). Morfessor and hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages. *Proceedings of the 2nd Baltic Conference on Human Language Technologies*, Tallinn, 107–112.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, *41*, 391–407.

Durand, J., Bennett, P., Allegranza, V., Eynde, F., Humphreys, L., Schmidt, P., & Steiner, E. (1991). The Eurotra linguistic specifications: An overview. *Machine Translation*, *6*(2), 103–147.

Gordon Jr, R. G. (Ed.). (1995). *Ethnologue: Languages of the World*, 15th edition. Dallas, TX: SIL International.

Haarmann, H. (2002). *Kleines Lexikon der Sprachen. Von Albanisch bis Zulu* (2. überarbeitete Auflage, 2nd edition). München: C. H. Beck.

Honkela, T., Hyvärinen, A., & Väyrynen, J. (2005). Emergence of linguistic features: Independent component analysis of contexts. In A. Cangelosi, G. Bugmann & R. Borisyuk (Eds), *Proceedings of NCPW9, Neural Computation and Psychology Workshop*, Plymouth, England.

Honkela, T., Pulkki, V., & Kohonen, T. (1995). Contextual relations of words in Grimm tales analysed by self-organizing map. In *Proceedings of Icann-95, International Conference on Artificial Neural Networks*, *2*, 3–7. EC2 et Cie.

Horn, D., Solan, Z., Ruppin, E., & Edelman, S. (2004, February). Unsupervised language acquisition: syntax from plain corpus. In *Newcastle Workshop on Human Language*.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent Component Cnalysis*. New York: John Wiley & Sons Inc.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206–213.

Juola, P. (2008). Assessing Linguistic Complexity. In F. Karlsson, M. Miestamo & K. Sinnemaki (Eds), *Language Complexity: Typology, Contact, Change*. Helsinki/ Amsterdam: John Benjamins. Retrieved December 18, 2007 from http:// www.mathcs.duq.edu/~juola/papers.d/assess-comp.pdf

Juola, P., Bailey, T. M., & Pothos, E. M. (1998). Theory-neutral system regularity measurements. *Proceedings of the 20th Annual Conference of the Cognitive Science Society (CogSci-98)*.

Katzner, K. (2002). *The Languages of the World*. London and New York: Routledge.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd extended edition, Volume 30. Berlin/Heidelberg/New York: Springer.

Lagus, K., Kaski, S., & Kohonen, T. (2004). Mining massive document collections by the websom method. *Information Sciences*, *163*(1/3), 135–156.

Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. M. B. (2004). The similarity metric. *IEEE Transactions on Information Theory*, *50*(12), 3250–3264.

Li, M., & Vitányi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications* (2nd edition). New York: Springer Verlag.

Nirenburg, S., & Wilks, Y. (2000). Machine translation at 50. In M. Zelkowitz (Ed.), *Advances in Computers*. New York: Academic Press.

Och, F. J., Gildea, D., Khundapur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., & Radev, D. (2004). A smorgasbord of features for statistical machine translation. *Proceedings of the 2004 Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04)*, Boston, 2004.

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, *61*(4), 241–254.

Sigurd, B., Willners, C., Eeg-Olofsson, M., & Johansson, C. (1992). Deep comprehension, generation and translation of weather forecasts (Weathra). *Proceedings of the 14th Conference of Computational Linguistics, Coling 92*, Nantes.

Stockwell, R. P., & Minkova, D. (2001). *English Words, History and Structure*. Cambridge University Press.