Analysis of EU Languages Through Text Compression

Kimmo Kettunen[1], Markus Sadeniemi[2], Tiina Lindh-Knuutila[2], and Timo Honkela[2]

[1] University of Tampere, Department of Information Studies, Kanslerinrinne 1, FIN-33014, University of Tampere, Finland
Kimmo.kettunen@uta.fi
[2] Helsinki University of Technology, Laboratory of Computer and Information Science
P.O. Box 5400 FI-02015 HUT Finland
Markus.Sadeniemi@iki.fi, tiina.lindh-knuutila@tkk.fi, timo.honkela@tkk.fi

Abstract. In this article, we are studying the differences between the European languages using statistical and unsupervised methods. The analysis is conducted in different levels of language, 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. The analyses of this paper are based on the concept of Kolmogorov complexity, which is used to compare the language structure in syntactic and morphological levels. 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 classification. The results will serve as a tool in developing machine translation system(s), e.g., in the following way: if source language conveys more information in the morphological level and the target language more in the syntactic level, it is clear that the (machine) translator must be able to transfer the information from one level to another.

1 Introduction

The European Union has 21 official languages (including Irish from 1st of January 2007), which have approximately 407 million speakers. In this article we analyze parallel corpora in these 21 languages using statistical, unsupervised learning methods to study the similarities and differences of the languages in different levels. We compare these results with traditional linguistic categorizations like division into language groups, morphological complexity and syntactic complexity. The aim of the study is to evaluate the possibility of using statistical methods in different tasks related to statistical machine translation. For instance, for some language pairs the issues related to morphological analysis may be particularly relevant. For some other language pairs, one may have to pay particular attention to the word order. These kinds of questions can be taken into account when the statistical models to be used are chosen.

Much of the material produced by the Union has to be translated to all languages, and 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

T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 99 – 109, 2006.

there are 410 language pairs to translate. It is evident that even automatic low-quality translation would be of great help.

EU documents are often difficult for a human to read and understand. For automatic processing and translation the situation might not be so problematic. Language used in documents is typically well structured, uses many words with exactly one translation and still embraces only a small part of human life.

This article provides basic information that could be used in the development of "next generation"' learning machine translation (MT) systems. The basic idea is that one should be able to cover, for instance, 420 pairs of EU languages in not too distant future[1]. This objective cannot be achieved unless the process of developing the MT systems is substantially automated. We do not consider MT itself in this paper, but rather analyze the complexity of EU languages. The analysis 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 similar level of complexity when corresponding linguistic characteristics are considered as 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 the MT systems, e.g., considering their parameterization.

1.1 Linguistic Comparison of Languages

It is estimated that the number of languages in the world is in several thousands, over 6000 being a usual figure to be mentioned [1, 2]. Of those, 21 are official EU languages: Czech (cs), Danish (da), Dutch (nl), English (en), Estonian (et), Finnish (fi), French (fr), German (de), Greek (el), Hungarian (hu), Irish (ga) (from 1st of January 2007), Italian (it), Latvian (lv), Lithuanian (lt), Maltese (mt), Polish (po), Portuguese (pt), Slovak (sk), Slovene (sl), Spanish (es) and Swedish (sv). Most of these belong to the Indo-European language family. 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) [1, 2]. In the present EU, only Estonian, Finnish, Hungarian and Maltese do not belong to Indo-European language family. The three first are Finno-Ugric languages, and Maltese is a Semitic language, Arabic written in Latin alphabet.

A working hypothesis is that the automated translation between two languages that belong to the same group, for example Romance languages, is easier than between those that belong to different groups, let alone different language families. In this article, we conduct statistical analyses to assess whether the differences and similarities of the languages could have significance considering the difficulty of translation. A basic assumption is that if two languages share features or have similarity in a particular level of complexity the translation between these languages is relatively easier.

[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/

2 Data and Methods

2.1 Data and Preprocessing

As language material we used parallel texts of EU Constitution in the 21 official languages of the European Union. The texts are smallish but representative, and each text consists of ca. 113 000 – 177 000 word forms and ca. 9100 – 15 000 sentences depending on the language. The character coding of the texts is UTF-8. The total number of files is 987, which means 47 files in each of the 21 languages. The total number of word form tokens in the corpus is 3 099 290. The original files are automatically XML-tagged to include, e.g., sentence, paragraph and word boundary information[2] [3].

The texts of each of the 21 languages were pre-processed by cleaning them from extra tags etc. and making all words lowercased. Then two modifications were made to the cleaned texts, one on the morpheme/word level and another on word order level. 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". In another modification the words in each sentence were shuffled to a random order [cf. 4]. The ending punctuation was kept at its place.

After pre-processing we had three versions of the text in each language: original law texts cleaned from XML tags and slightly normalized, one word per line, and files where word forms were randomized and files with shuffled word order.

2.2 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 is and thus the Kolmogorov complexity of the sequence is also lower [5, 6, 7].

Kolmogorov complexity is uncomputable, 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 [5].

Estimations of complexity using compression has been used for different purposes in many areas. Juola [4] introduces comparison of complexity between languages on morphological level for linguistic purposes. "By selectively altering the expression of morphological information, one can measure the amount of morphological complexity contributes to a corpus by measuring the change in perceived informativeness." Juola's method is simple: after randomization of the morphological level, the size of the original compressed file is divided with 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 random distortion other levels of language can also be analyzed [8].

[2] Materials are available from http://logos.uio.no/opus/index.html

3 Results

3.1 Compression: The Juola Style

For comparison of the complexity of the languages three files were compressed using *bzip2* program[3]. The sizes of modified compressed texts were then compared to the original compressed one to get a measure on the change of information, when morphological and word order information in the texts were destroyed. In Table 1 we have figures of the compressed language files.

Table 1. Compression results of the files. A = original (cleaned) compressed file, B = words replaced by random numbers, file compressed, C = words of sentences shuffled to random order and file compressed, D = language.

A	B	C	D
158606	145956	206540	cs
156115	138097	215904	da
169236	145144	224822	de
181890	158274	249777	el
149490	141982	217175	en
161700	152196	239311	es
151050	137791	193037	et
161067	138409	203658	fi
160846	151428	243122	fr
168550	159304	245621	ga
168831	147829	228542	hu
160627	152720	234036	it
157123	145381	206011	lt
151512	140713	202518	lv
165988	149947	230652	mt
169179	151200	237162	nl
168857	148408	221580	pl
157958	147963	230835	pt
166421	149307	216623	sk
153428	145154	215130	sl
156210	138832	209294	sv

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

[3] Available online at http://www.bzip.org

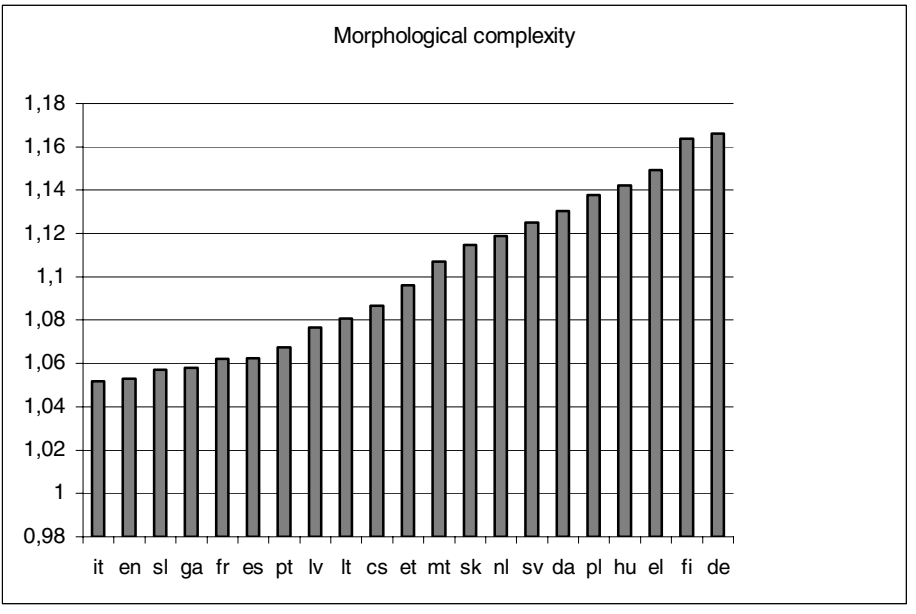

Fig. 1. Morphological complexity of the languages analyzed with compression

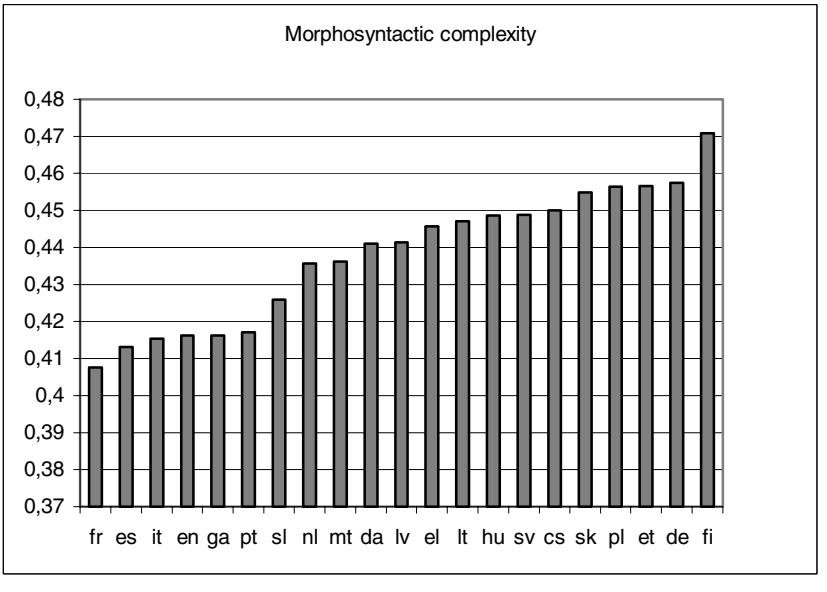

Fig. 2. Morphosyntactic complexity of the languages analyzed with compression

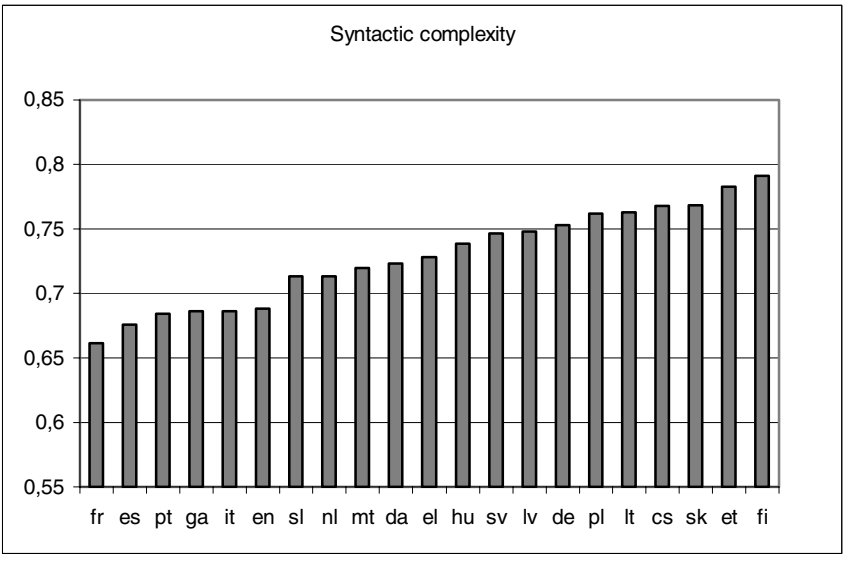

Fig. 3. Syntactic complexity of the languages analyzed with compression

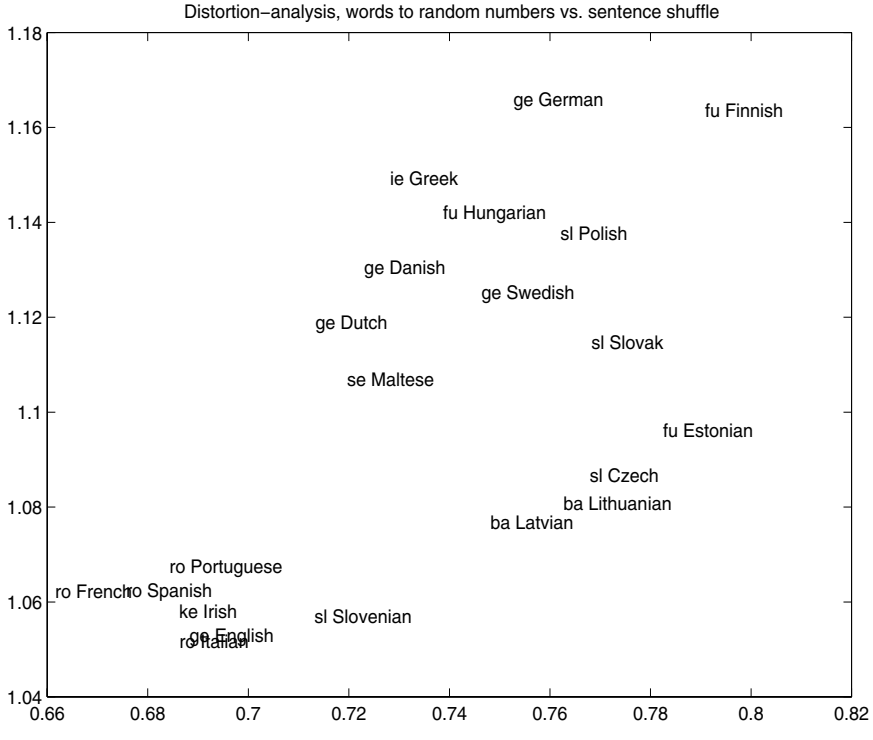

Fig. 4. Morphological and syntactic complexity of the languages in a two-dimensional graph

A few comments of Figure 1 are necessary. Mostly the results are as expected: morphologically simple languages, Italian, English, Irish, French, Portuguese and Spanish are getting low scores and morphologically more complex languages, Finnish, Hungarian and Polish, are in the other end of the scale. But some of the results are not very expected: Slovene, Slovak, Latvian, Czech and Estonian should be higher on the complexity scale. Dutch, Swedish, Danish and German seem to get quite high values, German being even on the top of the scale. It is possible, that compound words cause this effect. Also the type of texts, legalese, could have a boosting effect on the complexity of German and other Germanic languages.

In Figure 2 we show the morphosyntactic complexity of the languages by adding columns B and C together and dividing figure from column A of Table 1 with the result, A/(B + C).

In Figure 3 the syntactic complexity of the languages is shown as a relation of columns A and C from Table 1 (A/C)..

In Figure 4 data of figures 1 and 3 are joined as a two-dimensional graph.

Figure 4 shows the languages plotted on a two-dimensional graph using the variables of morphological and syntactic complexity (A/B and A/C). As we can see, Romance languages are grouped neatly into southwest corner of the picture and seeing English near them is no surprise. Finnish and German are located near the top of the figure. Baltic and other Slavic languages are generally more on the southeast side than Germanic languages, although the separation is not very clear.

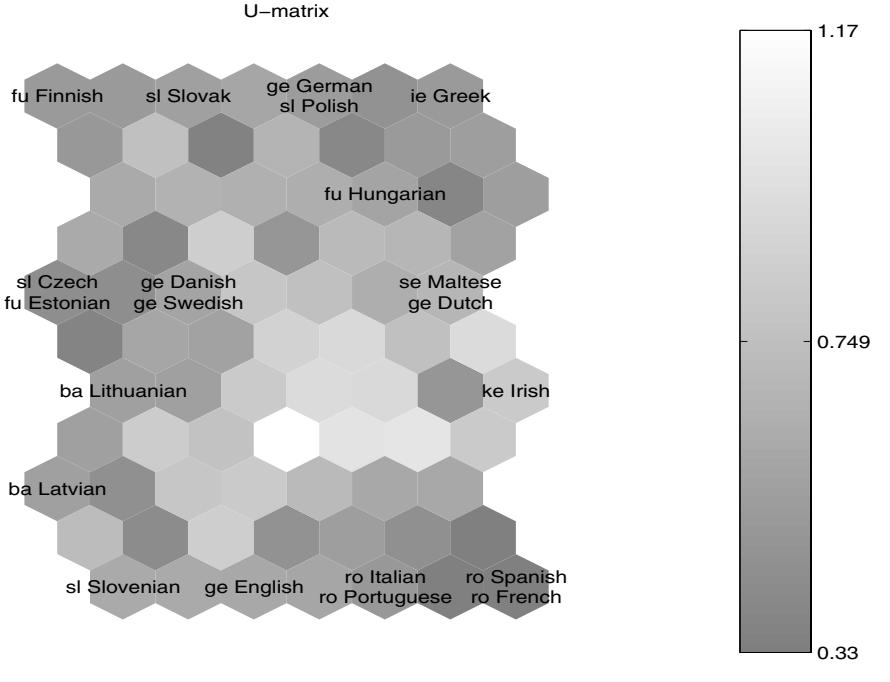

Fig. 5. Languages in a SOM-map: morphology vs. word order information

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 words whereas Romance languages and English are on the other end of the scale. It must be remembered, of course, that when talking about word order, we do not only mean clause level SVO-like grammatical structures but also constituent level things, like nominal heads and their different modifiers.

Figure 5 shows languages on a self-organizing map (SOM-map). Input variables in this picture are the three compressed file sizes as such. A SOM map is a highly nonlinear projection from the original feature space to a two-dimensional map. This is done in a way that observations - here languages - that are close in original space remain close on the map. Longer distances don't remain proportional, however. The map in figure 5 shows Romance languages well clustered again and English near them. Danish and Swedish are close as they should, but Estonian should rather be near to Finnish than to Czech.

3.2 Interpretation of Morphosyntactic and Syntactic Complexities

The morphosyntactic complexity of the languages in Figure 2 is partly as expected, 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 in the top. Only Swedish seems to be higher in the scale than expected and Latvian and Slovene 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 [9, pp. 387–] introduces flexibility of 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 in Bakker's counting can be given with a numeric value from 0 - 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, i.e. predictability.

In Table 2 figures based on Bakker's [9, pp. 417 – 419] counting of the flexibility values for the individual languages are given together with values given by compression analysis.

If we compare the figures given by Bakker in column 2 to figures given by compression based calculation in column 4, we can see, that the overall order of the languages based on these independent calculations converge well. The lower end of the scale is quite analogous in both analyses consisting of same five languages with only minor differences in the order. There are also some bigger 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 vs. 8).

Table 2. 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 compession figures for these languages are 0,74 (Hungarian) and 0,77 (Czech).

Order of the languages based on Bakker's flexibility calculation	Bakker's flexibility value	Syntactic complexity order of the languages based on compressional relation calculations from Figure 3.	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

Slovene has also a higher flexibility value than its complexity value (14 vs. 7). Greek is also higher in Bakker's counting than in complexity analysis (17 vs. 11). In our compression calculations Finnish and Estonian are estimated almost equally complex, but in Bakker's analysis Estonian is less complex than Finnish (18 vs. 13).

3.3 Compression: Cilibrasi and Vitányi Style

Another method for comparing the similarity of languages using compression is described by Cilibrasi and Vitànyi [10]. Again the size of a compressed text file is used to measure its Kolmogorov complexity as described in Li et al. [6].

A compression program (also bzip2 here) learns the characteristics of a language as it processes the text. If the language of the text changes in the middle of processing the compression program has to adapt to a 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.

So the similarity of languages can be measured by how well the compression manages this transition. In mathematical terms we can mark the size of 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

$$(C(xy) - C(x))/C(y) \tag{1}$$

which measures the change in compressing language y when using x as model. The expression acknowledges the possibility that the relation can be asymmetric: perhaps x is better explained by y than vice versa.

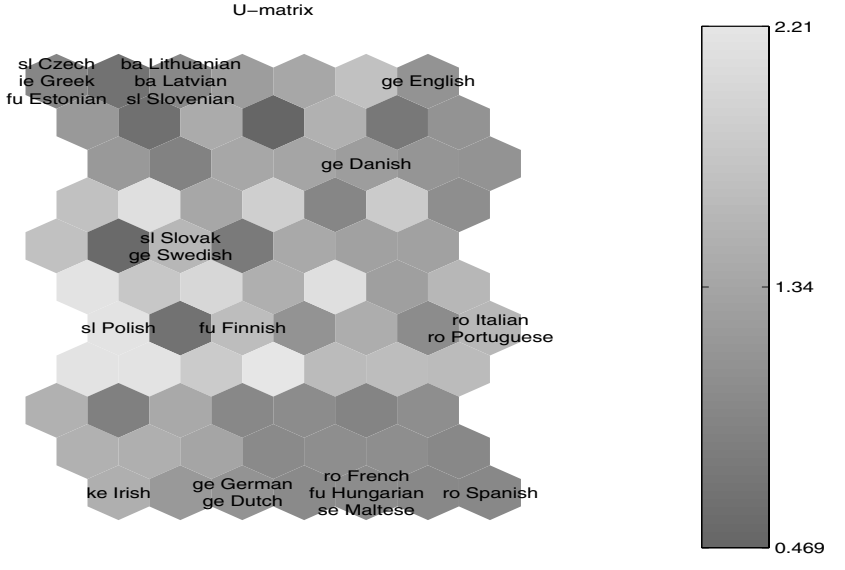

Fig. 6. A SOM-map analysis of languages showing language pair distances

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

4 Discussion and Conclusion

In this paper we have used a file compression program as an analysis tool for complexity of the 21 official EU languages on lexical, morphological and syntactic levels. Our analyses have shown that the approach is capable of showing relations between languages on these levels. The level of analysis is, however, relatively coarse, but

results given are mainly in accordance with linguistic descriptions of the languages; this is most clearly shown with the syntactic complexity analysis, when compression results are related to Bakker's flexibility values for the languages in Table 2.

What, then, could be the use of this type of general level information theoretic analysis? One suggested way to use the analyses would be in development work of a statistical machine translation system. The basic idea is that the translation process can be divided into interrelated tasks following, e.g., 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 unsupervised learning method [11] 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

1. Gordon, R. G., Jr. (ed.): Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL International (2005). Online version**:** http://www.ethnologue.com/
2. Haarman, H.: Kleines Lexikon der Sprachen. Von Albanisch bis Zulu. Verlag C.H. Beck, München, 2.,überarbeitete Auflage (2002)
3. Tiedemann, J., Nygaard, L: The OPUS Corpus - Parallel & Free. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, May 26-28. 2004 http://www.let.rug.nl/~tiedeman/blog/paper/opus_lrec04.pdf. Accessed 30 January 2006.
4. Juola, P.:. Measuring Linguistic Complexity: the Morphological Tier. Journal of Quantitative Linguistics 5 (1998) 206–213
5. Li, M, Vitanyi, P. An Introduction to Kolmogorov Complexity and its Applicatrions. Springer Verlag, New York Berlin Heidelberg (1994)
6. Li, M., Chen, X., Li, X., Ma, B, Vitányi, P.M.B.:The Similarity Metric. IEEE Transactions on Information Theory. 50 .(2004). 3250 - 3264
7. Bennet, C.H., Gács, P., Li, M., Vitányi, P.M.B., Zurek, W.H.:Information Distance. IEEE Transactions on Information Theory. 44 (1998) 1407 - 1423
8. Juola, P.: Compression-Based Analysis of Language Complexity. Approaches to Complexity in Language, abstracts. (2005) http://www.ling.helsinki.fi/sky/tapahtumat/complexity/Abstracts.pdf. Accessed January 15th 2006
9. Bakker, D.: Flexibility and Consistency in Word Order Patterns in the Languages of Europe. In Siewierska, A. (ed.): Constituent Order in the Languages of Europe. Empirical Approaches to Language Typology. Mouton de Gruyter, Berlin New York (1998). 381 – 419
10. Cilibrasi, R., Vitányi, P. M. B.: Clustering by Compression. IEEE Transactions on Information Theory, 51 (2005), 1523–1545
11. Creutz, M., Lagus, K.: Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. Espoo: Publications in Computer and Information Science, Helsinki University of Technology, Report A81 (2005)