# DEPENDENCY PROFILES AS A TOOL FOR BIG DATA ANALYSIS OF LINGUISTIC CONSTRUCTIONS: A CASE STUDY OF EMOTICONS

**Veronika Laippala, Aki-Juhani Kyröläinen, Jenna Kanerva, Juhani Luotolahti, and Filip Ginter**
*University of Turku*

**Abstract**. This study presents a methodological toolbox for big data analysis of linguistic constructions by introducing dependency profiles, i.e., co-occurrences of linguistic elements with syntax information. These were operationalized by reconstructing sentences as delexicalized syntactic biarcs, subtrees of dependency analyses. As a case study, we utilize these dependency profiles to explore usage patterns associated with emoticons, the graphic representations of facial expressions. These are said to be characteristic of Computer-Mediated Communication, but typically studied only in restricted corpora. To analyze the 3.7-billion token Finnish Internet Parsebank we use as data, we apply clustering and support vector machines. The results show that emoticons are associated with three typical usage patterns: stream of the writer's consciousness, narrative constructions and elements guiding the interaction and expressing the writer's reactions by means of interjections and discourse particles. Additionally, the more frequent emoticons, such as:), are used differently than the less frequent ones, such as ^_^.

**Keywords**: dependency profiles, usage-based syntax, computer-mediated communication, emoticons, web corpora, Finnish

**DOI**: https://doi.org/10.12697/jeful.2017.8.2.05

## 1. Introduction

Usage-based corpus studies have shown that multiple cues in the linguistic environment guide speakers' choices, such as dative alternation (*I gave you a book* ~ *I gave a book for you* (Bresnan *et al.* 2007) and various near-synonyms, such as Russian verbs of trying (Divjak and Gries 2006) and Finnish verbs of thinking (Arppe 2008). Indeed, syntactic and semantic information associated with linguistic environments provide a detailed representation of a particular word or expression (see Edmonds and Hirst 2002). With the development of extremely large corpora with syntactic analyses[1], this enables large scale studies on the use of various linguistic constructions, which, however, imposes its own methodological difficulties. As noted by Joachims (1998), linguistic constructions are known to be dense, as many features denote a single construction, but simultaneously sparse, as constructions tend to be described by a small number of high frequency features, while the bulk of unique features is large. In this study, we focus on this methodological challenge by conducting a large-scale analysis on a syntactically annotated corpus, in our case the Finnish Internet Parsebank, a web-crawled corpus with syntactic analyses and 3.7 billion tokens (Luotolahti *et al.* 2015). We rely on usage-based models of language ascribing to the distributional hypothesis (e.g., Harris 1968), where elements with similar environments sharing similar co-occurrence patterns are also likely to share similar semantic/functional properties (see Divjak and Gries 2006, and Gries 2010). We investigate the co-occurrence patterns of an expression with syntactic dependencies, operationalizing these by forming *dependency profiles* for the studied expressions.

There is a long tradition in computational linguistics to use combinations of words, i.e., bigrams, to model linguistic elements (see Jurafsky and Martin 2000: 191–206 for an overview), and dependency relations have been successfully applied to study binary relations between words (Wu and Weld 2010). While word-based analyses are easily scalable for large corpora, they have the disadvantage of typically reflecting topical characteristics (e.g., Scott and Tribble 2006). Therefore, rather than relying on combinations of words, we profit of the syntactic analysis of the Parsebank, and utilize combinations of dependency relations. The lexical realizations of the syntactic structures can, if needed, be included at a later step in the analysis as is demonstrated in Section 7.

As a case study to test the applicability of the method, we present a large-scale analysis of the usage patterns associated with emoticons, as reflected by their dependency profiles. Emoticons are the graphic representations of, e.g., facial expressions and perhaps one of the most conspicuous features of Computer-Mediated Communication (CMC; Herring, Stein, and Virtanen 2013). They are often related to informality, socio-emotional communication and other characteristics of CMC, such as non-standard spelling and telegraphic syntax (see, e.g., Baron 2004, Yus 2011, Herring 2012, and Bieswanger 2013). However, as the studies are often based on small and restricted corpora and as CMC

[1] See <http://wacky.sslmit.unibo.it/doku.php?id=corpora>. Accessed on 19.12.2016.

is considered to vary (e.g., Androutsopoulos 2006, and Dresner and Herring 2010), more research is needed to better understand both the use of emoticons specifically and the variation of CMC more generally. Our research questions are: (1) What are the typical usage patterns associated with emoticons, as reflected by the dependency profiles? (2) Are there differences between the emoticons? (3) What kind of information do the dependency profiles reflect?

Following the findings presented in the previous studies (see Baron 2004, Yus 2011, Herring 2012, Bieswanger 2013, and Vandergriff 2014), our hypothesis for the questions (1) and (2) is that emoticons would be typically co-occurring with elements involving interaction between the participants and informal elements. However, considering the frequency of emoticons in today's communication, we also hypothesize that at least some of them would be spreading to less informal contexts with few other CMC-specific elements. Finally, for the question (3), based on Laippala *et al.* (2015), we presume that dependency profiles would, minimally, enable to extend the analysis of the usage patterns from a topical to a more abstract and functional level, where, instead of describing *what is discussed*, they would reflect *what is being done*.

The article starts with a presentation of previous studies on emoticons, and continues with a description on the Finnish Internet Parsebank and the emoticons in it. Section 4 presents the construction of the dependency profiles. In Sections 5 and 6, we answer the research questions (1) and (2) and analyze the created profiles and the (dis)similarities between the emoticons. This is done by grouping emoticons with similar dependency profiles together with clustering (see Kaufman and Rousseeuw 1990), and by fitting a support vector machine (SVM) to estimate the clusters' typical features, i.e., the dependency syntactic characteristics related to the clusters and to the emoticons in them. Section 7 explores these characteristics and associates them with more general usage patterns, and examines the typical lexical realizations for some of these. This completes the answers for the research questions (1) and (2) and answers the question (3). Finally, the article ends with discussion and conclusions on the functioning of dependency profiles as well as the typical usage patterns of emoticons reflected by them.

## 2. Emoticons

Emoticons are graphic signs formed of ordinary typographical symbols, which have fast become prototypical of CMC, referred to as “constitutive of” (Vandergriff 2014) or “native to” (Dresner and Herring 2010: 13) this mode of communication. Previous studies have associated with emoticons a number of functions. The earlier ones see them as indicators of emotion, as “*visual cues representing feelings*” (Rezabek and Cochenour 1998), while more recent analyses attach pragmatic meanings to them (see Vandergriff 2014, and Dresner and Herring 2010). In addition, emoticons are said to serve similar functions as actual non-verbal behavior in the expression of intimacy and nuances (Derks *et al.* 2007).

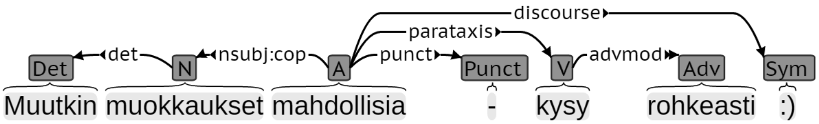
Emoticons are often analyzed together with other characteristics of CMC (see Herring 2012, Dresner and Herring 2010, Bieswanger 2013, and Crystal 2001). These include, among others, non-standard spelling, typography and acronyms, such as *!!!, wassup?* and the classic *lol*. Also telegraphic and fragmented syntax with elided elements is said to be frequent (Herring 2012: 5), to gain efficiency. Many of these features are not typical of standard written texts. Perhaps consequently, CMC is often considered as informal and playful (Derks *et al.* 2007, Baron 2004, Herring 2012, Bieswanger 2013, and Yus 2011).

Recent studies highlight, however, the variation of CMC; not all modes of CMC share the same situational, linguistic or CMC-specific characteristics (see, e.g., Androutsopoulos 2006). As a consequence, also the use of emoticons varies. Among others, Yus (2011: 198) associates emoticons with interpersonal rather than transactional communication, while Park *et al.* (2014) find variation by people’s cultural backgrounds. Dresner and Herring (2010) relate several factors to their use, among others, “*situational factors such as user demographics, topic of discussion, and communication setting*”.

As the use of emoticons varies, the results of the existing studies on emoticons offer, after all, very restricted information, and more research on their use is needed (Dresner and Herring 2010: 13). With the exception of Park *et al.* (2014) on emoticons in Twitter, they are typically based on small samples of writers in specific situations, most often writing in English: Derks *et al.* (2007) on a questionnaire of 158 secondary school students, Baron (2004) on 23 instant message conversations and Yus (2011) on 1700 chat room messages. Our approach taking into account all the emoticons found in a 3.7-billion token corpus enables another perspective. As the corpus is not divided into subcorpora, our approach will not be informative on the use of emoticons across modes of CMC or sociolinguistic parameters. In contrast, we offer a large-scale, generalizable insight to usage patterns of a large number of emoticons with thousands of occurrences.

## 3. Emoticons in the Finnish Internet Parsebank

The data consists of the Finnish Internet Parsebank (Luotolahti *et al.* 2015), a web-crawled corpus on the entire Finnish Internet. The current version includes 3.7 billion tokens and has full morphological and dependency syntax annotations carried out with a state-of-the-art dependency parser by Bohnet (2010), with a *labelled attachment score*[2] of 82.1%. The texts are cleaned of duplicates and lists. The syntax annotations are very detailed with 46 dependency types, ensuring a deep level of description that is useful for linguistic analysis (see Figure 1).



Other modifications possible – ask bravely :)

*Other modifications possible – do not hesitate to ask :)*

**Figure 1**. Dependency annotation of a sentence in the Finnish Internet Parsebank[3].

To create a corpus of emoticons, we first extracted all the tokens with the *Sym* part-of-speech tag used by the parser to label emoticons and other symbols, counted their frequencies and manually extracted the ones corresponding to emoticons. To narrow down the data, we took into account only the emoticons with more than 1,000 occurrences, and extracted the Parsebank sentences with at least one of these, yielding a corpus of 66,389,175 words with 38 different emoticons. Figure 2 presents the analyzed emoticons and their frequencies.[4]

[2] The percentage of tokens for which the system has predicted the correct head and the correct dependency relation.

[3] For details, see <https://universaldependencies.github.io/docs/>. Accessed on 19.12.2016.

[4] Some of the emoticons are most likely erroneously tokenized, such as :P. For technical reasons, however, we preserved the original analyses.

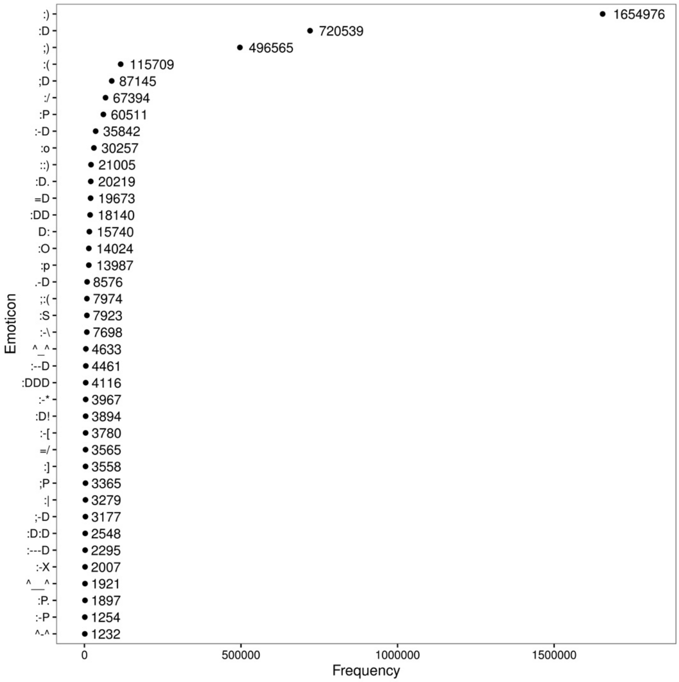


**Figure 2**. The frequency distribution of the emoticons in the Finnish Internet Parsebank.

As shown in Figure 2, the large majority of the found emoticons are horizontal and follow the Western tradition (see Dresner and Herring 2010, and Park *et al.* 2014), which is expected since the data comes from the Finnish Internet. The frequencies of the listed emoticons appear to follow Zipf's law: the three most frequent ones, i.e., :), :D, and ;), comprise more than 80% of the entire data. These findings are very close to those described by Park *et al.* (2014), who reported a similar distribution and found :) to be the most often used in a large corpus of Twitter data.

## 4. Dependency profiles

To investigate the typical usage patterns associated with emoticons, all the sentences with an emoticon were first reconstructed as *syntactic n-grams*, subtrees of dependency syntax analysis (Goldberg and Orwant 2013, and Kanerva *et al.* 2014). We used unlexicalized syntactic biarcs consisting of three tokens and two arcs with all the information except for the dependency relations removed (see Figure 3).

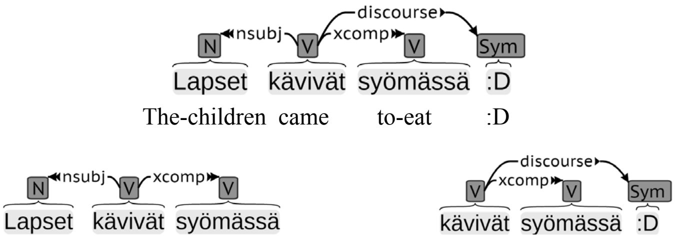


**Figure 3**. Syntactic biarcs of the sentence “The children came to eat :D”.

The use of the unlexicalized biarcs enable us to establish a more abstract, structure-oriented and less topic-dependent representation of the sentence compared to lexicalized n-grams (Scott and Tribble 2006, Laippala *et al.*, 2015, and Ivaska 2015). Additionally, unlexicalized n-grams allow us to alleviate issues related to data sparseness as the inclusion of lexical material would enormously increase the number of bigrams. Given this initial phase, the data set contained 122,809 unique syntactic biarcs.

The second phase of the analysis consisted of forming the co-occurrence patterns of all the unlexicalized syntactic biarcs with each emoticon. Additionally, several measures were taken to remove uninformative or infrequent biarcs. First, the actual emoticons were removed, as they can influence the further automatic analysis. Second, all the biarcs with *dep, name* or *punct* were removed, as these dependencies tend to be linguistically uninformative. Third, two frequency-based cut-off points were imposed to alleviate data sparseness: First, all biarcs with less than eight occurrences were removed, seven corresponding to the median in the frequency distribution across all the biarcs. Second, a type frequency was calculated for the biarcs across the emoticons. The biarcs that did not cover at least 20% of the emoticons were removed from the data. After these data pruning procedures, the final data set contained 10,568 unique biarcs and 38 emoticons along with their co-occurrence frequencies.

The co-occurrence of a particular emoticon with the syntactic biarcs forms a vector which we will refer to as a dependency profile. A snippet of two dependency profiles is illustrated in Table 1.

**Table 1**: A snippet of two dependency profiles. The rows represent the emoticons and the columns contain the unlexicalized syntactic biarcs. The cell values are the relative co-occurrence frequencies of a particular emoticon with a given syntactic biarc. The absolute frequencies are given in parenthesis. In the biarcs, the tokens are numbered by their order from 0 to 2, the number referring to the governor of the token. The dependency types are described in Appendix.

| | ROOT_0-acl:relcl_1-advcl_2 | ROOT_0-acl:relcl_1-advmod_2 |
|---|---:|---:|
| :) | 0.00007813882 (3505) | 0.0001885810 (8459) |
| :D | 0.00009325957 (71) | 0.0001497407 (114) |

As is shown in Table 1, we make use of relative frequencies rather than absolute ones. This is done to factor in the imbalance between the overall frequency of the emoticons in the data (see Evert 2009), as illustrated in Figure 2.

## 5. Grouping emoticons with similar dependency profiles together with clustering

To investigate the (dis)similarities between the dependency profiles of the emoticons, we applied clustering, a widely used method to group similarly behaving elements in data together into clusters. The clustering was carried out in R (R Core Team, 2016), using *k*-means clustering with Euclidian distance in the package flexclust, version 1.3–4 (Leisch 2006). The parameter *k* controls the number of clusters and is user-defined. To evaluate different cluster solutions, we used three methods (see Fraley and Raftery 1998): we evaluated the possible number of clusters, assessed the stability of the different cluster solutions, and visually inspected them.

To evaluate the possible number of clusters, the data were fitted iteratively starting from 2 and stopping at 15 clusters, the sum of within cluster distances being used as the optimization criterion (see Hothorn and Everitt 2014). The results of these different cluster solutions are visualized in Figure 4, suggesting that for these data, the solution with three or four clusters might be the best.

To evaluate the stability of the solutions, we carried out a bootstrap sampling of the original data (see Efron and Tibshirani 1993) with approximately 62.5% of the original data used in a given sample.

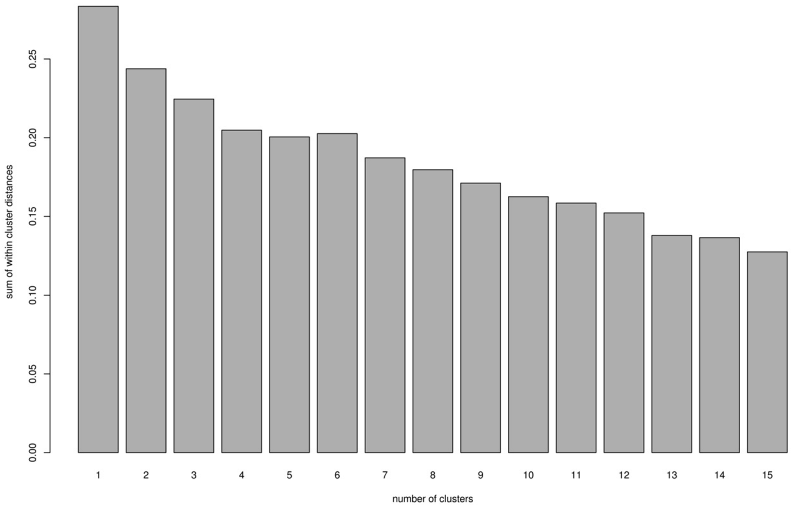


**Figure 4**: Visualization of the sum of within cluster distances (y-axis) for different cluster solutions. The x-axis represent the number of clusters in a given solution.

To assess how often the “same” clustering solution is obtained across the bootstrap samples, we used the Rand index (Hubert and Arabie 1985), defined as the percentage of data point pairs assigned either to the same or different cluster twice. In both cases, the solutions indicate an agreement. A Rand index ranges from 0 to 1, 1 showing a perfect agreement. This sampling procedure was carried out 200 times and the results are visualized in Figure 5.

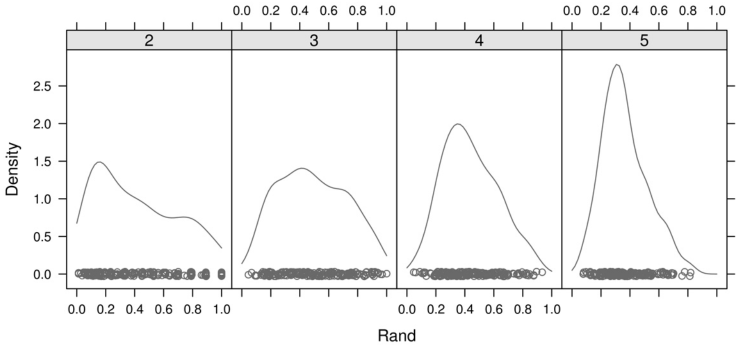


**Figure 5**: Density estimates of 200 bootstrapped adjusted Rand indices for four cluster solutions. The number in the panel indicates the number of clusters in the solution.

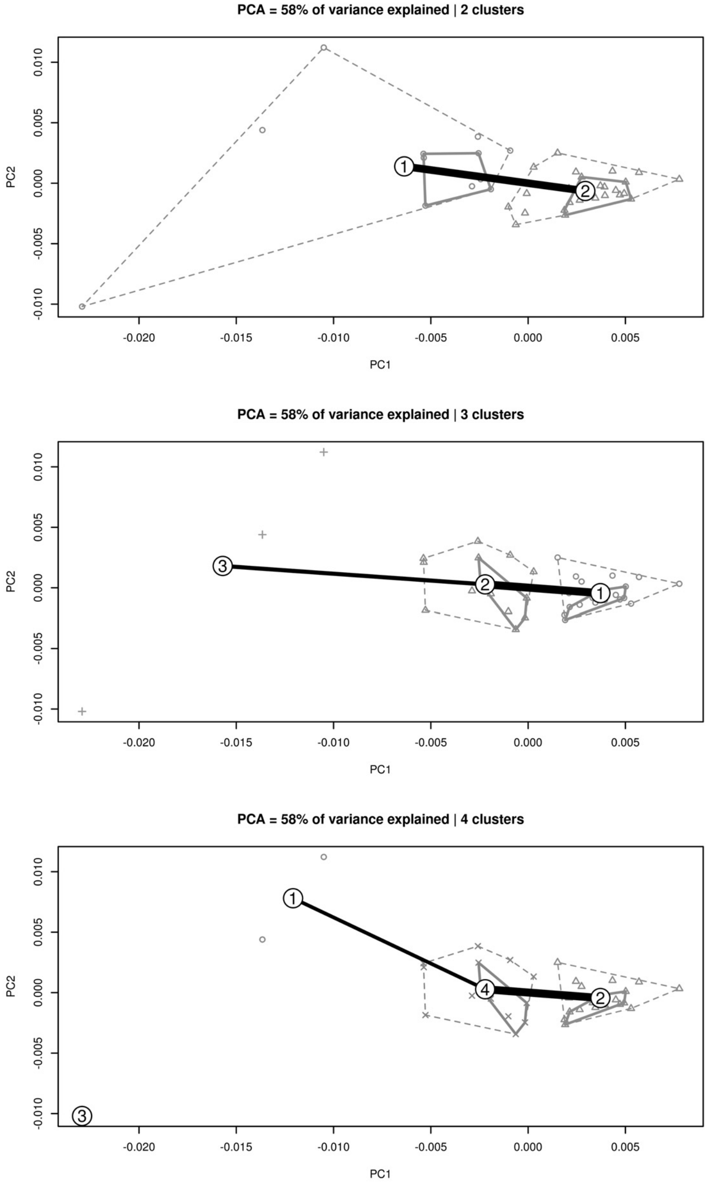


**Figure 6**. Neighborhood graphs for three cluster solutions. The numbers indicate the cluster label and the thickness of the black line represents a degree of separation where thin lines indicate a better separation. The solid colored lines represent 50% of the data and the dotted lines 95% around the clusters.

In Figure 5, the density estimates indicate the degree of instability among the cluster solutions. For these data, the solution with either three or four clusters appears the best, the greatest density being obtained around a Rand index of 0.5. This still indicates a certain instability, perhaps due to some of the emoticons not having a clear cluster membership.

To investigate this instability, we visualized three solutions in Figure 6 using principal component analysis (PCA) to reduce the dimensionality of the dependency profiles (see Pison *et al.* 1999). The results showed that two PCs explained 58% of the variation, indicating a high degree of overlap among the syntactic biarcs, as expected (see Joachims 1998).

The visualizations bring forth that the data appears to contain, in fact, two clusters that remained stable across the different cluster solutions. Therefore, in the analysis, we focus on these two. The distribution of the emoticons in these two clusters is described in the following Section.

## 6. Finding the cluster characteristics by estimating the importance of individual syntactic biarcs with SVMs

Several methods could be used to estimate the most important biarcs of the clusters, i.e., the ones driving the structuring of the emoticons into the two major clusters. These can as well be considered as the typical features of those clusters and analyzed as their linguistic characteristics (see Divjak and Gries 2006). In our case, the estimation of a single feature at a time, for instance with log likelihood, can be problematic (see Guyon and Elisseeff 2003 for discussion on univariate methods). Therefore, we treat the cluster solution as a classification task and use support vector machines (SVMs) to estimate the importance of the biarcs. SVMs are a machine learning algorithm that is highly suitable for learning patterns in high-dimensional data and estimating variable importance (Boser *et al.* 1992, Vapnik 1998, and Guyon *et al.* 2002). For our data, we applied a penalized linear SVM in scikit-learn[5] and used the biarcs of each sentence with an emoticon as predictors for each cluster class. To estimate the most important biarcs for each cluster, we considered only the ones with a positive weight (see Guyon *et al.* 2002 for discussion), as our goal is to find the ones that best characterize a

[5] See <http://scikit-learn.org/stable/>. Accessed on 19.12.2016.

given cluster. In the two following sections, we present the results of this analysis, by first exploring the most important biarcs of the two clusters, and then presenting the frequencies of individual dependency types across the biarcs in Section 6.2.

### 6.1 The role of individual syntactic biarcs in the clusters

Table 2 presents the distribution of emoticons across the analyzed clusters, and their 20 most important unlexicalized syntactic biarcs, estimated by SVMs. These biarcs denote also the syntactic characteristics of the clusters and the emoticons.

First of all, interestingly, an important factor in the grouping of emoticons seems to be their frequency; nearly all of the most frequent emoticons, such as :) and ;), are grouped into a single cluster, while the less frequent ones (e.g., :o and =/) are located in another cluster. This suggests that the most frequent emoticons are used in different syntactic contexts than the less frequent ones. At the same time, this cannot be simply a pure frequency effect because relative frequencies were used in the clustering. We will refer to these two clusters as FREQ and RARE.

Second, it seems that the typical biarcs do not consist only of the core dependency types of a clause, such as direct objects (*dobj*) or sentence roots; the *discourse* relation type denoting interjections and discourse particles is extremely frequent in both clusters, and also appositions (*appos*) and *parataxis*, used to signal loosely connected verbal elements e.g., in brackets, are frequent. Finally, the significant features include various complex constructions composed of clausal complements (e.g., *xcomp:ds, cccomp*), coordinations (*conj*) as well as adverbial and relative clauses (*advcl, acl:relcl*). Lower in the rank but still present are also determiners (*det*), adverbs (*advmod*) and coordinating and adverbial conjunctions (*cc*, *mark*).

**Table 2.** The distribution of 20 syntactic biarcs with the highest positive weight for each cluster. See Appendix for the dependency types.

| Cluster RARE | | | Cluster FREQ | | |
|---|---|---|---|---|---|
| :-\ :-* ^_^ :-[ =/ :] | | | :) :D. :DD =D :O D: ;:( :p :D :o :D! | | |
| ;-D ^__^ :---D :P. ;P | | | :--D ;) :DDD :S :( :D:D :/ :P :-D ;D ::) | | |
| Rank | Syntactic ngram | Weight | Rank | Syntactic ngram | Weight |
| 1 | xcomp:ds/0-conj/1-cc/1 | 0,9054 | 1 | ccomp/0-parataxis/1-acl:relcl/2 | 1.165 |
| 2 | cop/2-nmod:poss/0-cc/2 | 0,8304 | 2 | nsubj/2-xcomp:ds/0-discourse/2 | 1.1363 |
| 3 | nummod/3-cc/1-dobj/0 | 0.7243 | 3 | det/2-conj/0-discourse/2 | 1.1293 |
| 4 | nmod/0-cc/1-appos/1 | 0.685 | 4 | amod/0-conj/1-discourse/1 | 1.072 |
| 5 | det/2-compound:nn/3-discourse/0 | 0.6763 | 5 | discourse/3-det/3-nsubj:cop/0 | 1.0688 |
| 6 | aux/2-acl/0-cc/2 | 0.6693 | 6 | nsubj:cop/3-acl/3-parataxis/0 | 1.0666 |
| 7 | root/0-xcomp:ds/1-parataxis/2 | 0.6638 | 7 | nsubj/0-advcl/1-discourse/1 | 1.0259 |
| 8 | xcomp/0-ccomp/1-discourse/1 | 0.6581 | 8 | parataxis/0-discourse/1-discourse/1 | 1.0119 |
| 9 | neg/3-nmod/3-dobj/0 | 0.6518 | 9 | nmod:own/3-nsubj/3-ccomp/0 | 1.0104 |
| 10 | advmod/2-xcomp:ds/0-discourse/2 | 0.6406 | 10 | conj/0-compound:nn/3-discourse/1 | 1.0006 |
| 11 | nsubj/3-cc/1-xcomp:ds/0 | 0.6241 | 11 | nsubj/3-discourse/1-conj/0 | 0.9831 |
| 12 | dobj/2-acl:relcl/0-discourse/2 | 0.6235 | 12 | advcl/0-advcl/3-parataxis/1 | 0.9814 |
| 13 | nsubj:cop/2-acl:relcl/0-advmod/2 | 0.6218 | 13 | remnant/0-appos/1-conj/2 | 0.9718 |
| 14 | discourse/3-nsubj/3-acl/0 | 0.6141 | 14 | discourse/0-nmod/3-conj/1 | 0.9365 |
| 15 | root/0-nummod/1-vocative/1 | 0.6052 | 15 | root/0-discourse/1-appos/2 | 0.914 |
| 16 | mark/2-nsubj/0-conj/2 | 0.5864 | 16 | ccomp/0-xcomp:ds/1-discourse/1 | 0.9133 |
| 17 | appos/0-advcl/1-nsubj/2 | 0.5795 | 17 | advmod/2-nummod/3-discourse/0 | 0.9036 |
| 18 | appos/0-appos/1-cc/2 | 0.5732 | 18 | nummod/3-conj/1-dobj/0 | 0.9022 |
| 19 | parataxis/0-mark/3-acl:relcl/1 | 0.572 | 19 | conj/0-discourse/1-xcomp/1 | 0.873 |
| 20 | nsubj/2-acl/0-discourse/2 | 0.5694 | 20 | xcomp:ds/0-nmod/1-parataxis/1 | 0.8714 |

### 6.2 Dependency types across the biarcs

To evaluate the contribution of particular dependency types across the top 20 biarcs, we counted their type frequency given in Table 3.

**Table 3**. Individual dependency types across the top 20 biarcs in the two clusters along with their frequency.

| Cluster RARE | | Cluster FREQ | |
|---|---|---|---|
| Dependency type | Frequency | Dependecy type | Frequency |
| cc | 7 | discourse | 14 |
| discourse | 6 | conj | 8 |
| nsubj | 5 | parataxis | 5 |
| xcomp:ds | 4 | nsubj | 4 |
| appos | 4 | xcomp:ds | 3 |
| dobj | 3 | ccomp | 3 |
| acl:relcl | 3 | advc | 3 |
| acl | 3 | nummod | 2 |
| root | 2 | nummod | 2 |
| parataxis | 2 | nsubj:cop | 2 |
| nummod | 2 | nmod | 2 |
| nmod | 2 | nmod | 2 |
| mark | 2 | det | 2 |
| conj | 2 | appos | 2 |
| advmod | 2 | xcomp | 1 |
| xcomp | 1 | root | 1 |
| vocative | 1 | remnant | 1 |
| nsubj:cop | 1 | nmod:own | 1 |
| nmod:poss | 1 | dobj | 1 |
| neg | 1 | compound:nn | 1 |
| det | 1 | amod | 1 |
| cop | 1 | advmod | 1 |
| | | acl:relcl | 1 |
| | | acl | 1 |

These type frequencies indicate that both clusters include frequent dependency types denoting clausal complements, coordination, discourse particles and parataxis. Interestingly, also determiners and adverbs appear as typical in both. However, differences also emerge. The *discourse* and *ccomp* dependencies are more frequent in cluster FREQ, suggesting that discourse particles and clausal complements with *that*-clauses are more typical of this cluster. On the other hand, subordinate and coordinating conjunctions (*mark, cc*) are more frequent in the characteristics of cluster RARE; in fact, coordinating conjunctions do not appear at all on the list of the cluster FREQ. This suggests that conjunctions are more typically used in cluster RARE, while in cluster FREQ coordination and subordination are expressed more implicitly.

## 7. From dependency profiles to typical usage patterns of emoticons

In the previous Section, we analyzed the most important biarcs of the clusters and the emoticons in them. In this Section, we extend the analysis from individual biarcs to the most typical usage patterns they reflect. We also complement this information by counting the most frequent lexicalizations of the tokens governed by certain dependency types. This allows for a multilevel analysis of the usage patterns, applying the structural properties of the sentences as a basis but benefiting also from lexical information.

In the previous section, we showed that various clausal complements, complex constructions and discourse particles are typical of both clusters, although the discourse particles and *that*-clauses denoted by the *ccomp* dependency type are more frequent in cluster FREQ. Our analysis shows that these syntactic characteristics reflect, in fact, three typical usage patterns for emoticons: stream of the writer's consciousness, narratives, and discourse particles guiding interaction and expressing writers' reactions.

### 7.1. Stream of the writer's consciousness

CMC is often associated with telegraphic writing including frequent omitted parts of speech (see Baron 2004, Herring 2012, and Bieswanger 2013). On the other hand, greater sentence lengths and structural complexity have been associated with "more formal" registers, such as news (Biber 1995: 261, and Baldwin *et al.* 2013). Despite this, our findings associate emoticons with long and complex sentences with frequent adverbial clauses, coordination, and clausal complements. However, instead of reflecting formal language variants, the sentences illustrating these are long and clumsy and could rather be seen as manifestations of unedited text where the writer writes as (s)he thinks, adding elements to the sentence while writing. The resulting sentences can be described as stream of the writer's consciousness which, in the literary studies, is applied to denote a narrative method for representing a continuous flow of perceptions and thoughts in the human mind (Baldick 2008: 212).

Example 1 below illustrates a long chain of coordinated elements, where the writer describes his / her actions and feelings during an imaginary situation. The verbs are in first person singular, and the sentence consists of altogether five coordinated elements. The illustrated biarc is given in the caption, and the lexical biarc root is in bold.

Example 1: xcomp:ds/0-conj/1-cc/1, rank 1 in cluster RARE

> Odottaisin sua takkatulen ääressä valmiina ja **ottaisin** ylpeänä vastaan sun naapurista varastamasi nauriit ja tekisin niistä meille sieninauri-soppaa ja kuuntelisin siinä samalla sinun hurjia sankaritekojasi, ja taas ah, olisi niin valloitettu olo että ;-D
> I would wait for you in front of a fire ready and would **accept** proudly the turnips you stole from the neighbour and make us a soup of mushrooms and turnips and at the same time listen to your terrific stories and again, ah, I would feel so charmed ;-D

Each of the clauses of example 1 start with a coordinating conjunction and a conditional. This is in fact typical of cluster RARE, where coordinating and subordinating conjunctions are both frequent. In cluster FREQ, this is not the case suggesting that the explicit use of conjunctions is less frequent, and coordinated elements follow each other implicitly.

In addition to coordination, other elements typically expressing stream of consciousness in our data are clausal complements (*ccomp, xcomp* and *xcomp:ds*), appositions and *parataxis*, a dependency type used for loosely connected phrasal elements. These can be citations in brackets or quotation marks, or, as in the following, a clause following the emoticon used in the middle of the sentence.

Example 2: xcomp:ds/0-nmod/1-parataxis/1, rank 20 in cluster FREQ

> Oksilla oli virtaa alussa niin kovin, että Valtteri käski minun **juosta** koiran kanssa edestakaisin :D idea olikin hyvä, koiran keskittyminen ottamaan kontaktia enemmän minun kuin haahuilemaan toisten luo parani huomattavasti.
>
> Oksi had in the beginning so much energy that Valtteri told me **to run** back and forth with the dog :D the idea was good, the dog's concentration to having contact with me more than wondering around to others got remarkably better.

The first clause of the example describes what is being done, while the second functions as an addition or commentary. The sentence is also an interesting example of unedited text consisting of ideas rather than grammatically correctly bound clauses, as the comparison between *to having contact with me* and *wondering around* is somewhat clumsy and uncoherent.

### 7.2. Narrative constructions

Another typical usage pattern related to emoticons are narrative constructions denoted e.g., by *parataxis* and *ccomp*: the first describes often citations, and the latter *that*-clauses used for reported speech (see Biber 1988: 109). These are typical of the more frequent emoticons placed in cluster FREQ.

Example 3 below illustrates a very typical *that*-clause in our data, a first person narrative, where the reported speech is expressed in the clausal complement.

Example 3: ccomp/0-parataxis/1-acl:relcl/2, rank 1 in cluster FREQ

> Tosin, veljeltä **kuulin**, että jos jättää kaapin vahingossa auki niin joku alikersantti melko varmasti "varastaa" sieltä jotain ihan vain opetukseksi, ettei sitä kaappia pidä jättää lukitsematta :D.
>
> Although, I **heard** from my brother that, if you leave your locker open by accident then some sergeant will most likely "steal" something just to teach a lesson, to never leave the locker unlocked :D.

In fact, a manual analysis of the sentences with *ccomp* and *parataxis* dependencies shows that first person narratives are very typical in our data. This is very natural for many modes of CMC and for emoticons as a way of expressing the writer's emotions and intentions. This does not show, however, in the most important syntactic biarcs of the clusters, as the morphological information was not included.

Another characteristic of the narratives in our data, used together with emoticons, illustrated in example 3 is that instead of forming an independent story, the narrative is embedded in the ongoing discourse (see Becker and Quasthoff 2005). In this example, for instance, it would seem like a written conversation, perhaps in blog comments or other asynchronous mode of CMC.

Second, in addition to describing actual events, the narratives can also depict the writer's mental events, such as thoughts or feelings. This is also supported by the most frequent main verbs used in these constructions: *sanoa* 'to say', *tietää* 'to know', *huomata* 'to notice', *kertoa* 'to tell', *uskoa* 'to believe'. In addition to the typical public verbs used in narratives (see Biber 1998: 103, 134–135) to report on what has actually been done, these include verbs of cognition and perception describing the mental processes of the cognizer.

Finally, example 4 below illustrates another characteristic of narratives co-occurring with emoticons, citations, typically used in quotation marks and marked by the *parataxis* dependency.

Example 4: advcl/0-advcl/3-parataxis/1, rank 12 in cluster FREQ

> Toki jos se on oma mielipide niin jeejee, mutta en itse kyllä menisi sanomaan kenellekään "ÄLÄ VAIN LUE TÄTÄ KIRJAA SE ON IHAN TÄYTTÄ **KURAA**!1" vaikka sitä mieltä olisinkin – joku tykkää, joku ei, ja minusta oikeastaan kaikenlainen kirjojen suosittelu/päinvastoin on täysin turhaa ja aika hyödytöntä :D.
> Of course if it's your own opinion, cool, but personally I wouldn't go and say "MAKE SURE YOU DON'T READ THIS BOOK IT IS COMPLETE **CRAP**!1" to anybody – some like, some don't, and I actually think that recommending / unrecommending books is completely pointless and quite useless :D.

In example 4, the writer is, similarly to the previous examples, writing in first person and describing a particular situation and what (s)he would and would not do there, the citation being in fact a self-citation. With altogether six clauses, this sentence is also a well-suited example of stream of consciousness.

### 7.3. Guiding interaction and expressing reactions

In Section 2, while presenting previous studies on emoticons, we mentioned that they can be seen as "*visual cues representing feelings*" (Rezabek and Cochenour 1998), and that they have been associated with interpersonal communication (Yus 2011: 198). Given this, it is not surprising that our study associates emoticons with interjections and discourse particles signaled by the *discourse* dependency type. These denote typically the writer's reaction and have intersubjective functions as part of dialogues (Hakulinen *et al.* 2004: §797, §856). Also Biber (1988: 131–132) associates them with interactive registers, such as telephone conversations.

As we explained in Section 5, these interjections and discourse particles are extremely typical of the emoticons in cluster FREQ, although they are used in cluster RARE as well. The most frequent lexicalizations of the ones that are part of the typical syntactic biarcs of cluster FREQ are *no*, *noh* 'well, oh', *hei* 'hey', *ai* 'oh', *voi* 'oh', *heh*, *hih* 'ha', *huh* 'phew' and *juu* 'yeah'. Many of these carry similar discourse functions. In particular the most frequent *no* or *noh* 'well, oh' can function, among others, as an interjection marking the writer's reaction, as a dialogue particle guiding the interaction, or as a discourse particle connecting the utterance to the ongoing discourse (Hakulinen *et al.* 2004: §797, §808, §856–857, §1051). Of the others, at least *juu* 'yeah' and *ai* 'oh' would also seem to function as dialogue particles, while others, such as *heh*, *hih* 'ha' and *huh* 'phew', are interjections expressing affective reactions (Hakulinen *et al.* 2004: §856).

Many of the examples in the previous section already illustrated the use of interjections and particles. For example, example 4 includes several, as it begins with *toki* 'of course', a discourse particle linking the sentence to the ongoing discourse, and has also an interjection expressing positive affect: *jeejee* 'cool'. Example 5 below illustrates another case, where, as opposed to the previous examples, *discourse* is part of the most typical biarc.

Example 5: discourse/3-det/3-nsubj:cop/0, rank 5 in FREQ

> Mutta juuh nuo **tohvelit** oli kyllä kivoja, en vain keksinyt että miksi meidän hikivarvas tarvitsisi sellaiset niin jätin kauppaan :D.
> But yeah those **slippers** were nice, I just couldn't figure out why our sweaty toe would need those so I left them to the store :D.

The sentence starts with a combination of two particles, *mutta* (lit. 'but, oh'), *juuh* 'yeah'. The main function of these particles seems to be to keep the dialogue going and the contact between the interlocutors open. Also, *mutta* 'but' in the beginning indicates the addition of the sentence to the ongoing discourse (Hakulinen *et al.* 2004: §801, §812). Another interesting aspect of example 5 is the use of the determiner *nuo* 'those'. As already mentioned in Section 5, the *det* dependency type is part of the typical biarcs for both clusters, and has been noted to be typical of informal registers (Biber 1988: 154).

## 8. Discussion and conclusions

In this article, we presented first steps towards a general methodological toolbox for big data analysis of linguistic constructions. The analysis presented here profits from an automatic syntactic analysis in order to form dependency profiles, i.e., co-occurrence counts of syntactic biarcs with a particular linguistic element. In principle, this method can be used to investigate single words, multiword units or even syntactic configurations. As a case study, we applied the dependency profiles to explore the typical usage patterns associated with emoticons in the 3.7-billion token Finnish Internet Parsebank. Dependency profiles can be further analyzed with different quantitative methods. Here, we applied clustering to investigate the (dis)similarities between the emoticons, and used support vector machines to estimate the most typical biarcs of each cluster.

For the current study of emoticons, the dependency profiles were formed with unlexicalized syntactic biarcs to extend the analysis beyond a topical level, to reach more abstract and functional usage patterns. In addition, the removal substantially reduced the dimensionality of the data. At the same time, it is possible to extend the analysis with lexical dimensions, effectively, transitioning from an abstract, syntactic representation back to the lexical realizations of the dependency relations in the sentences. This further highlights the versatility of this method. In sum, the results indicate that the proposed method is an effective tool allowing us to investigate (dis)similarities among linguistic elements and the contribution of both syntactic relations and lexical realizations.

In terms of the typical usage patterns of emoticons, our study showed that dependency profiles can capture functional elements and register characteristics associated with the analyzed expressions (Biber and

Conrad 2009: 9), the three most typical usage patterns related to emoticons being stream of the writer's consciousness, narrative constructions and elements guiding the interaction and expressing the writer's reactions, such as interjections and discourse particles. On one hand, these confirm our hypothesis on the use of emoticons in interactive contexts. As we noted, interjections and discourse particles are typical of dialogues and spoken language (Hakulinen *et al.* 2004: §797, §856), and also the narrative sequences with emoticons have conversational characteristics, such as the use first person verbs (Biber 1995: 60). On the other hand, some results also challenge previous studies, as many CMC-specific elements, such as telegraphic syntax and acronyms, are absent in the typical usage patterns, meaning that the use of emoticons with these elements is not very distinctive, when they are analyzed across a variety of modes of CMC. Rather, in our data, fast writing shows as unedited text. Additionally, the infrequency of adjectives in the typical usage patterns in somewhat surprising. Finally, concerning the differences between the emoticons, our results suggest that the most frequent and typical emoticons are used differently than the less frequent ones. In particular, the interjections and discourse particles expressing interaction and the writer's reactions as well as narratives tend to co-occur with the more frequent emoticons.

The article opens multiple possibilities for future studies. From a methodological perspective, it is worth pointing out that the method can be enriched even further by incorporating morphological information as part of the dependency profiles. Additionally, the syntactic biarcs also contain information about word order. The incorporation of these sources of information would bring syntax and information structure together simultaneously in a single analysis. We are currently investigating the use of these sources in another project.

Finally, the analysis of the typical usage patterns associated with emoticons and presented in Section 7 leaves many questions open and would merit a more detailed analysis. In addition to further exploring the three patterns we have established, many of the most important biarcs presented in Section 6, such as adverbs, nominal subjects and determiners, could still be further examined. Lastly, also the differences between the clusters and emoticons in them could yet be further studied to form an even more elaborate understanding of the use of emoticons.

## Acknowledgements

We thank the editors and reviewers for their useful comments to improve the manuscript. This research was supported by the Finnish Internet Parsebank project funded by the Kone Foundation.

**Address**
Veronika Laippala
School of languages and translation studies
20014 University of Turku
Finland
E-mail: veronika.laippala@utu.fi

## References

Androutsopoulos, Jannis (2006) "Introduction: Sociolinguistics and computer-mediated communication". *Journal of Sociolinguistics* 10, 4, 419–438.

Arppe, Antti (2008) *Univariate, bivariate, and multivariate methods in corpus-based lexicography: a study of synonymy*. PhD thesis. Helsinki: University of Helsinki.

Baldick, Chris (2008) *The Oxford dictionary of literary terms*. Oxford: Oxford University Press.

Baldwin, Timothy, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang (2013) "How noisy social media text, how different social media sources?". *Proceedings of the 6th international joint conference on natural language processing (IJCNLP 2013)*, Nagoya, Japan, 356–364. Available online at <http://aclweb.org/anthology/I/I13/I13-1041.pdf>. Accessed on 19.12.2016.

Baron, Naomi S. (2004) "See you online. Gender issues in college student use of instant messaging". *Journal of Language and Social Psychology* 23, 4, 397–423.

Becker, Tabea and Uta M. Quasthoff (2005) "Introduction: different dimensions in the field of narrative interaction." In Uta M. Quasthoff and Tabea Becker, eds. *Narrative interaction*, 1– 11. Amsterdam and Philadelphia: John Benjamins.

Biber, Douglas (1988) *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas (1995) *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas and Susan Conrad (2009) *Register, genre, and style*. Cambridge: Cambridge University Press.

Bieswanger, Markus (2013) “Micro-linguistic structural features of computer-mediated communication”. In Susan Herrin, Dieter Stein and Tuija Virtanen, eds. *Pragmatics of computer-mediated communication*, 463–485. Berlin and Boston: De Gruyter Mouton.

Bohnet, Bernd (2010) “Top accuracy and fast dependency parsing is not a contradiction”. *Proceedings of COLING’10*. Available online at <http://www.anthology.aclweb.org/C/C10/C10-1011.pdf>. Accessed on 19.12.2016.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik (1992) “A training algorithm for optimal margin classifiers”. *Proceedings of the fifth annual workshop on computational learning theory*, 144–152. Available online at <http://dl.acm.org/citation.cfm?id=130401>. Accessed on 19.12.2016.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen (2007) “Predicting the dative alternation”. In Gerlof Bouma, Irene Krämer, and Joost Zwarts, eds. *Cognitive foundations of interpretations*, 69–94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Crystal, David (2001/2006) *Language and the Internet*. Cambridge University Press.

Derks, Daantje, Arjan E. R. Bos, and Jasper von Grumbkow (2007) “Emoticons and social interaction on the Internet: the importance of social context”. *Computers in Human Behavior 23*, 1, 842–849.

Divjak, Dagmar and Stefan Th. Gries (2006) “Ways of trying in Russian: clustering behavioral profiles”. *Corpus Linguistics and Linguistic Theory* 2, 1, 23–60.

Dresner, Eli and Susan C. Herring (2010) “Functions of the nonverbal in CMC: emoticons and illocutionary force”. *Communication Theory* 20, 3, 249–268.

Edmonds, Philip and Graeme Hirst (2002) “Near-synonymy and lexical choice”. *Computational Linguistics* 28, 2, 105–144.

Efron, Bradley and Robert J. Tibshirani (1993) *An introduction to the bootstrap*. New York: Chapman & Hall.

Evert, Stefan (2009) “Corpora and collocations”. In Anke Lüdeling and Merja Kytö, eds. *Corpus linguistics: an international handbook*, 233–233. Berlin and New York: Mouton de Gruyter.

Fraley, Chris and Adrian E. Raftery (1998) “How many clusters? Which clustering method? Answers via model-based cluster analysis”. *The Computer Journal* 41, 8, 578–588.

Goldberg, Yoav and Jon Orwant (2013) “A dataset of syntactic-n-grams over time from a very large corpus of English books”. *Second Joint Conference on Lexical and Computational Semantics (*SEM), 1: Proceedings of the Main Conference and the Share d Task: Semantic Textual Similarity*, 241–247. Available online at <https://static.googleusercontent.com/media/research.google.com/fi//pubs/archive/41603.pdf>. Accessed on 19.12.2016.

Gries, Stefan Th. (2010) "Behavioral profiles: a fine-grained and quantitative approach in corpus-based lexical semantics". *The Mental Lexicon* 5, 3, 323–346.

Guyon, Isabelle M., Jason Weston, Stephen Barnhill, and Vladimir N. Vapnik (2002) "Gene selection for cancer classification using support vector machines". *Machine Learning* 46, 1–3, 389–422.

Guyon, Isabelle M. and André Elisseeff (2003) "An introduction to variable and feature selection". *The Journal of Machine Learning Research* 3, 1157–1182.

Hakulinen, Auli, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja R. Heinonen, and Irja Alho (2004) *Iso suomen kielioppi* [The comprehensive grammar of Finnish]. Helsinki: Suomalaisen Kirjallisuuden Seura.

Harris, Zellig (1968) *Mathematical structure of language*. New York: Wiley.

Herring, Susan C. (2012) "Grammar and electronic communication". In Carol A. Chapelle, ed. *Encyclopedia of applied linguistics*. USA: Wiley Blackwell. DOI: 10.1002/9781405198431.wbeal0466.

Herring, Susan C., Dieter Stein and Tuija Virtanen, eds. (2013) *Pragmatics of computer-mediated communication*. Berlin and Boston: de Gruyter Mouton.

Hothorn, Torsten and Brian S. Everitt (2014) *A handbook of statistical analyses using R*. 2nd ed. Boca Raton: CRC press.

Hubert, Lawrence and Phipps Arabie (1985) "Comparing partitions". *Journal of Classification* 21, 193–218.

Ivaska, Ilmari (2015) "Tracing crosslinguistic influences in structural sequences: what does key structure analysis have to offer?" *BeLLS – Bergen Language and Linguistics Studies* 6, 23–44.

Joachims, Thorsten (1998) "Text categorization with support vector machines: learning with many relevant features". *Proceedings of the 10th European Conference on Machine Learning*, 137–142. Available online at <http://web.cs.iastate.edu/~jtian/cs573/Papers/Joachims-ECML-98.pdf>. Accessed on 19.12.2016.

Jurafsky, Daniel and James H. Martin (2000) *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 1st edition. Upper Saddle River, NJ: Prentice Hall PTR.

Kanerva, Jenna, Matti Luotolahti, Veronika Laippala, and Filip Ginter (2014) "Syntactic N-gram collection from a large-scale corpus of Internet Finnish". Proceedings of the sixth international conference Baltic HLT, 184–191. Available online at <https://pdfs.semanticscholar.org/57df/7ef1d75853edf4df97d847f281e571df2b3e.pdf>. Accessed on 19.12.2016.

Kaufman, Leonard and Peter J. Rousseeuw (1990) *Finding groups in data: an introduction to cluster analysis*. New York: John Wiley.

Laippala, Veronika, Jenna Kanerva, and Filip Ginter (2015) "Syntactic N-grams as key structures reflecting typical syntactic patterns of corpora in Finnish".

*Procedia – Social and Behavioral Sciences. Current Work in Corpus Linguistics* 198, 233–241. <doi:10.1016/j.sbspro.2015.07.441>.

Leisch, Friedrich (2006) “A toolbox for K-centroids cluster analysis”. *Computational Statistics and Data Analysis* 51, 2, 526–544.

Luotolahti, Juhani, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter (2015) “Towards universal web parsebanks”. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, 211–220. Available online at <http://www.aclweb.org/anthology/W15-21>. Accessed on 19.12.2016.

Park, Jaram, Young M. Baek and Meeyoung Cha (2014) “Cross-cultural comparison of nonverbal cues in emoticons on Twitter: evidence from big data analysis”. *Journal of Communication* 64, 2, 333–354.

Pison, Greet, Anja Struyf, and Peter J. Rousseeuw (1999) “Displaying a clustering with CLUSPLOT”. *Computational Statistics & Data Analysis* 38, 4, 381–392.

R Core Team (2016) *R: A language and environment for statistical computing. R foundation for statistical computing*. Vienna, Austria. Available online at <https://www.R-project.org>.

Rezabek, Landra L and John J. Cochenour (1998) “Visual cues in computer-mediated communication: supplementing text with emoticons”. *Journal of Visual Literacy* 18, 2, 201–215.

Scott, Mike and Christopher Tribble (2006). *Textual patterns: key words and corpus analysis in language education*. Amsterdam/Philadelphia: John Benjamins.

Vandergriff, Ilona (2014) “A pragmatic investigation of emoticon use in nonnative/native speaker text chat”. *Language@Internet* 11, article 4. Available online at <http://www.languageatinternet.org/articles/2014/vandergriff>. Accessed on 19.12.2016.

Vapnik, Vladimir N. (1998) *Statistical learning theory*. New York: Wiley Interscience.

Wu, Fei and Daniel S. Weld (2010) “Open information extraction using wikipedia”. *Proceedings of ACL*, 118–127. Available online at <http://www.aclweb.org/anthology/P10-1013>. Accessed on 19.12.2016.

Yus, Francisco R. (2011) *Cyberpragmatics. Internet-mediated communication in context*. Amsterdam/Philadelphia: John Benjamins.

## Appendix: Dependency types in the Universal Dependencies scheme, as described in <http://universaldependencies.org/docs/fi/dep/index.html>.

acl: clausal modifier of noun
acl:relcl: relative clause modifier
advcl: adverbial clause modifier
advmod: adverb modifier
amod: adjectival modifier
appos: apposition
aux: auxiliary
auxpass: passive auxiliary
case: case marking
cc: coordinating conjunction
cc:preconj: preconjunct
ccomp: clausal complement
compound: compound
compound:nn: noun compound modifier
compound:prt: phrasal particle
conj: coordinated element
cop: copula
csubj: clausal subject
csubj:cop: clausal copular subject
dep: dependent
det: determiner
discourse: discourse element
dislocated: dislocated elements
dobj: direct object
expl: expletive
foreign: foreign
goeswith: goeswith
list: list
mark: marker
mwe: multi-word expression
name: name
neg: negation modifier
nmod: nominal modifier
nmod:gobj: genitive object
nmod:gsubj: genitive subject
nmod:own: haver
nmod:poss: genitive modifier
nsubj: nominal subject
nsubj:cop: nominal copular subject
nummod: numeric modifier
parataxis: parataxis
punct: punctuation
remnant: remnant in ellipsis
reparandum: overridden disfluency
root: root
vocative: vocative modifier
xcomp: open clausal complement
xcomp:ds: clausal complement with different subject

**Kokkuvõte. Veronika Laippala, Aki-Juhani Kyröläinen, Jenna Kanerva, Juhani Luotolahti ja Filip Ginter: Sõltuvusprofiilid kui vahend suurandmete keeleliste konstruktsioonide analüüsimiseks: uurimus emotikonidest.** Uurimuses esitame metodoloogilise "tööriistakomplekti" keelekonstruktsioonide analüüsimiseks suurandmete põhjal, rakendades sõltuvusprofiile. Sõltuvusprofiil on lingvistiliste elementide koosesinemise esitusviis, kuhu on kaasatud süntaktiline informatsioon. Selleks on laused konstrueeritud sõltuvusanalüüsi alampuudena, kus süntaktiline info on esitatud sõnadevaheliste (kaksik-)kaarte abil. Artiklis rakendame sõltuvusprofiile selleks, et selgitada välja emotikonide kasutusmustrid. Näomiimika graafilised esitused on iseloomulikud arvutisuhtlusele, mida tavaliselt uuritakse piiratud korpuse põhjal, kuid meie kasutame klasterdamist ja tugivektor-masinaid 3,7 miljardi sõna suuruse Soome Interneti Puudepangal. Selgub, et emotikonide kasutus seostub kolme peamise kasutusmustriga: kirjutaja teadvuse vooluga, narratiivsete konstruktsioonidega ning hüüdsõnade ja diskursusepartiklitega, mis juhivad suhtlust ja väljendavad kirjutaja reaktsioone. Lisaks selgub, et sagedastel emotikonidel nagu :), on rohkem erinevaid kasutusi kui harvadel emotikonidel nagu ^_^.

**Võtmesõnad**: sõltuvusprofiilid, kasutuspõhine süntaks, arvutisuhtlus, emotikonid, veebikorpus, soome keel