Linguistic Frontiers • 1(2) • 122-132 • 2018
DOI: 10.2478/lf-2018-0011

Linguistic **Frontiers**

Genetic analysis of cabbages and related cultivated plants using the bag-of-words model

Research Article

Hana Owsianková[1], Dan Faltýnek[1], Ondřej Kučera[2]

[1]The Department of General Linguistics and Theory of Communication, Palacký University, Olomouc, Czech Republic
[2]The Department of Asian Studies, Palacký University, Olomouc, Czech Republic

Received 19 November, 2018; Accepted 22 January, 2019

Abstract: In this study, we aim to introduce the analytical method bag-of-words, which is mainly used as a tool for the analysis (document classification, authorship attribution and so on; e.g. [1, 2]) of natural languages. Quantitative linguistic methods similar to bag-of-words (e.g. Damerau–Levenshtein distance in the paper by Serva and Petroni [3]) have been used for the mapping of language evolution within the field of glottochronology. We attempt to apply this method in the field of biological taxonomy – on the Brassicaceae (Cruciferae) family. The subjects of our interest are well-known cultivated crops, which at first sight are morphologically very different and culturally perceived as objects of different interests (e.g. oil from oilseed rape, turnip as animal feed and cabbage as a side dish). Despite the phenotypic divergence of these crops, they are very closely related, which is not morphologically obvious at first sight. For this reason, we think that Brassicaceae crops are appropriate illustrative examples for introducing the method. For the analysis, we use genetic markers (internal transcribed spacer [ITS] and maturase K [matK]). Until now, the bag-of-words model has not been used for biological taxonomisation purposes; therefore, the results of the bag-of-words analysis are compared with the existing very well-developed Brassica taxonomy. Our goal is to present a method that is suitable for language development reconstruction as well as possibly being usable for biological taxonomy purposes.

Keywords: *bag-of-words • glottochronology • language development • evolution • taxonomy • molecular genetics • Brassicaceae*

Introduction

Language development reconstruction is a domain of historical-comparative linguistics. Historical-comparative linguistics reconstructs the genetic language tree by tracing the language changes in the different stages of language development. This approach deals with the detailed description of the phonetic and grammatical development processes and their interconnections in the structure of particular languages. In the early 1950s, this approach was complemented by glottochronology, pioneered by Morris Swadesh, which deals with relations among languages based on changes occurring in vocabularies.

In the glottochronology approach, a list of words that are resistant to language borrowing is used: the Swadesh list contains personal pronouns, body parts, animals, verbs of basic actions, colours, numerals 'one' and 'two', heavenly bodies and so on. The amount of differences between the words determines the distance between the languages in the tree. The main comparison tool of words from the Swadesh list is the Damerau–Levenshtein distance. The Damerau–Levenshtein distance expresses the minimum number of transformations of one string to the other. The string transformations that are observed are inserting, deleting, replacing parts of a string (substitution) or changing the positions of parts of the string (transposition).

As a convenient example of usage of the Damerau–Levenshtein distance measured for words from the Swadesh list, we consider the phylogenetic tree of languages from the article by Serva and Petroni [3]. For more information about glottochronology and examples of using the Swadesh list, see e.g. [4–6].

Another linguistic method appropriate for classification is the bag-of-words model. This method is primarily used in the field of natural language processing, information retrieval, computer vision and document classification (e.g. [7, 8]); it has also been used for the analysis of genetic texts more recently [9, 10]. The bag-of-words method involves

* *Corresponding author: E-mail: hana.owsiankova01@upol.cz*

representing the text in its words, regardless of their order in the text, and taking into account their frequency. The first mention of this model can be found in the article by Harris [11]. The bag-of-words model allows the evaluation of the similarity of texts based on a comparison of their vocabulary.

Each method works with words differently, but they can still lead to similar results. What the Damerau–Levenshtein distance considers as a transformation from one word to the other, the bag-of-words model regards as two different words whose frequency in the text it counts.

Language classification can also serve as an exemplary case for the utilisation of the bag-of-words model. The following graph (Figure 1) represents a hierarchical clustering tree whose branches divide languages into language families – the severance is based on the vocabulary similarities of the analysed texts. We compared 'Our Father' prayer translated into 24 languages belonging to different language families. We utilised the software QUITA (Quantitative Index Text Analyser) for analysis – the default parameters were retained, the texts were tokenised and transformed into 2-grams of characters.

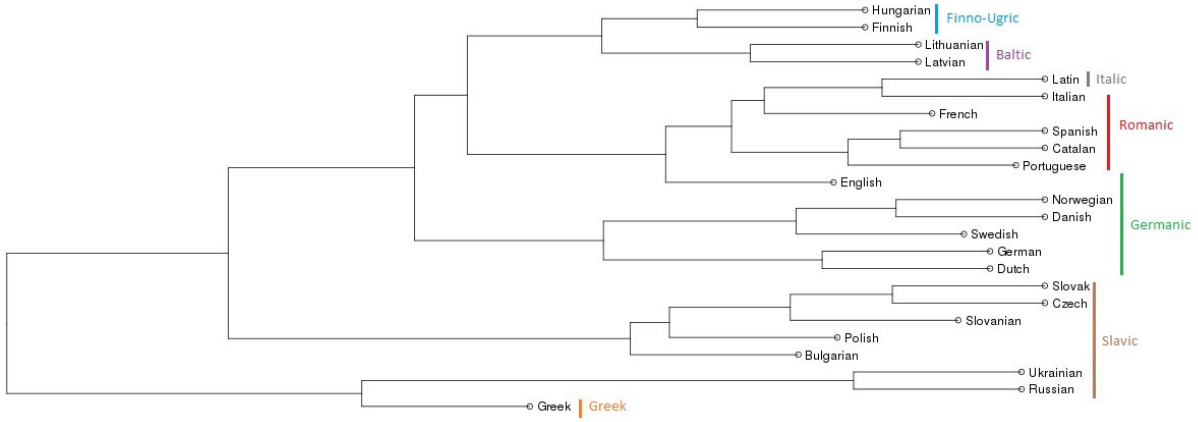

Figure 1: Classification of languages based on the bag-of-words model of 2-grams of characters; the analysed text was 'Our Father' prayer.

As the graph points out, the bag-of-words method is suitable for language classification as well. Our goal in this study is to demonstrate that this method could possibly be usable also for biological taxonomy purposes. Our assumption is that resemblance (and therefore proximity) between natural language texts or between DNA texts (sequences) can be compared on a similar principle, i.e. based on components (words, bases) contained in the unit (sentence/text, sequence).

Materials and methods

For reasons explained earlier (morphological divergence versus genetic proximity), we think that Brassicaceae crops are proper illustrative examples for introducing the bag-of-words method for taxonomic purposes. The taxonomy of the Brassicaceae family has been carried out very systematically on the basis of molecular markers mainly for individual genera and tribes (tribe is a taxonomic category lower than the family and higher than the genus) and also for species (see, e.g. [12–20]). An overview of research in this area is provided by Al-Shehbaz et al. [21]. The oldest work on the taxonomy of the Brassicaceae family dates back to the early 20th century – Hayek [22] introduced it in 1911. The phylogenetic relationships of the species *Brassica oleracea, B. rapa, B. napus, B. carinata, B. juncea* and *B. nigra* were described by the so-called U-model proposed by Nagaharu [23] (see Figure 2). This model represents a basic orientation diagram of the crossing between the named species (see e.g. *Genetics, Genomic and Breeding of Vegetable Brassicas* [24]). The tribes of the Brassicaceae family were described in 1936 by Schulz et al. [25]. Al-Shehbaz et al. [21] and Bailey et al. [26] report on 25 tribes, 338 genera and 3,700 species within the Brassicaceae family. This taxonomy is based on internal transcribed spacer (ITS) marker analysis. Liu et al. [27] confirm this taxonomy by analysing the maturase K (matK) marker and attempt to refine the relations between individual taxa. An overview of other markers (trnL-trnF, trnH-psbA, simple sequence repeats [SSRs] and so on) used in the taxonomy of Cruciferae is reported by Al-Shehbaz et al. [21].

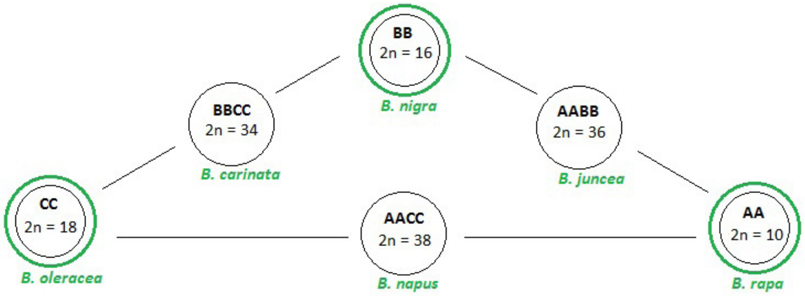

Figure 2: U-model of relations among *Brassica* species; n = number of chromosomes (author: Hana Owsianková).

Our study uses the family Brassicaceae as an illustrative example for introducing the bag-of-words method in order to outline the possibility of using this method as a tool for biological taxonomy. We are interested in a sample of cultivated crops belonging to the family Brassicaceae, which have significant economic use and position in culture. These crops are classified into different genera and species.

Besides the reconstruction of the genetic tree of the *Brassica* species, we are also interested in the biological taxonomy of other family crops – e.g. the crops of the genera *Eruca* (arugula/rocket), *Raphanus* (radish) and *Sinapis* (mustard).

In the analysis, we work with genetic markers of species *B. nigra* (black mustard), *B. juncea* (brown mustard), *B. carinata* (Abyssinian mustard), *B. rapa* (field mustard), *B. oleracea* (wild cabbage), *B. napus* (rapeseed), *Eruca vesicaria* (arugula/rocket), *Raphanus sativus* (radish), *Sinapis alba* (white mustard) and varieties *B. oleracea botrytis* (cauliflower), *B. oleracea italica* (broccoli), *B. oleracea capitata* (cabbage), *B. oleracea alboglabra* (kai-lan), *B. oleracea acephala* (kale), *B. rapa chinensis* (Chinese cabbage), *B. rapa pekinensis* (Pekingese cabbage) and *B. rapa oleifera* (biennial turnip rape).

For analysis, we used the matK chloroplast marker and the ITS nuclear ribosomal marker. Both of these markers have been used in the earlier-mentioned studies for the taxonomy of the Brassicaceae family. Samples were derived from the National Center for Biotechnology Information (NCBI) genetic bank (for samples, see Appendix).

The method used in this study, the bag-of-words model, has been described earlier in this paper. Since this method is based on the delimitation of text into words, it is necessary to define the form of a 'genetic word' for the purposes of biological taxonomy. Therefore, for the representation of the word in a genetic text, we used its *n*-gram analysis: we split the genetic texts into equal parts (sub-strings) – into combinations of two bases (2-grams), three bases (3-grams) and so on. Individual species are represented by a series of values expressing the presence or absence of the given word (*n*-gram, sub-string) in individual sequences and by the difference in the frequency of the words in the individual sequences. The DNA sequences thus represented are then used in a clustering analysis, whose results are shown in graphs of hierarchical clustering.

In this approach, one can see a similarity to the alignment-free bioinformatic method called 'feature frequency profile', which is also based on word frequency, in which case, they are represented by *k*-mers. *K*-mers are usually defined as all possible sub-strings of length *k* that are contained in a string. In different words, *n*-grams and *k*-mers are the same, and the bag-of-words model and the feature frequency profile both work with their frequency.

For better understanding, we present a complete ITS marker of *B. oleracea* and *B. rapa,* as well as parts of their bag-of-words models (Table 1–3).

Brassica oleracea

```
ACTCTCGGTGGGCCGGTATCTTAGCTGATTTCGTGCCTACCGATTCCGTGGTTATGCGTTCGTCACCGGCCC
AGTTTCGGTTGGATTGTACGCATAGCTTCCGGATATCACCAAACCCCGGCACGAAAAGTGTCAAGGAACATT
CAACTAAACAGCCTGCTTTCGCCAACCCGGAGACGGTGTTTGTTCGGAAGCAGTGCTGCAATGTAAAGTCTA
AAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAAGAACGTAGCGAAATGCGATACTTGGTGT
GAATTGCAGAATCCCGTGAACCATCGAGTCTTTGAACGCAAGTTGCGCCCCAAGCCTTCTGGCCGAGGGCA
CGTCTGCCTGGGTGTCACAAATCGTCGTCCCCCAATCCTCTCGAGGATATCGGACGGAAGCTGGTCTCCCG
TGTGTTACCGCACGCGGTTGGCCAAAATCCGAGCTAAGGATGCCAGGAGCGTCTTGACATGCGGTGGTGAA
TTCAATTCTCGTCAAATCGTCAGTCGTTTCGGTCCGAAAGCTCTTGATG
```

(NCBI – *'Brassica oleracea ITS1 5.8S ITS2' – code 'AY722423.1'*)

Brassica rapa

```
TCGTACCCTGGAAACAGAACGACCTGAGAACGATGAAACATCACTCTCGGTAGGCCGGTTTCTTACTGTGC
CTGCTGATTCCGTGGTTATGCGTTCATCCTTGGCCAAGACTTCAGTTTTGGTTGGATCGTACGCATAGCTTC
CGGATATCACCAAACCCCGGCACGAAAAGTGTCAAGGAAAATGCAACTAAACAGCCTGCTTTCGCCAACCC
GGAGACGGTGTTTGTTCGGAAGCAGTGCTGCAATGTAAAGTCTAAAACGACTCTCGGCAACGGATATCTCG
GCTCTCGCATCGATGAAGAACGTAGCGAAATGCGATACTTGGTGTGAATTGCAGAATCCCGTGAACCATCG
AGTCTTTGAACGCAAGTTGCGCCCCAAGCCTTCTGGCCGAGGGCACGTCTGCCTGGGTGTCACAAATCGT
CGTCCCCCCATCCTCTCGAGGATATGGGACGGAAGCTGATCTCCCGTGTGTTACCGCACGCGGTTGGCCA
AAATCCGAGCTAAGGACGTCAGGAGCGTCTTGACATGCGGTGGTGAATTTAATTCTCGTCATATAGTCAGA
CGTTCCGGTCCAAAAGCTCTTGATGACCCAAAGTCCTCAAC
```

(NCBI – ‘*Brassica rapa ITS1 5.8S ITS2’ – code ‘AF531563.1’*)

Table 1: *Brassica oleracea*: ITS marker, the 10 most frequently represented 5-grams and their frequencies

Rank (frequency order)	n-gram (word)	Frequency
1	t c t c g	6
2	t c g t c	5
3	c t c t c	4
4	t t t c g	4
5	a a a t c	3
6	a a t c c	3
7	a t a t c	3
8	a t g c g	3
9	c c g t g	3
10	c g a a a	3

Table 2: *Brassica* rapa: ITS marker, the 10 most frequently represented 5-grams and their frequencies

Rank (frequency order)	n-gram (word)	Frequency
1	t c t c g	6
2	c c a a a	4
3	c t c t c	4
4	g a a c g	4
5	a a a c a	3
6	a a a g t	3
7	a a c g a	3
8	a g a a c	3
9	a t g c g	3
10	c c g t g	3

Example of bag-of-words (BoW) models for *B. oleracea* and *B. rapa*:

BoW1 = {‘t c t c g’: 6, ‘t c g t c’: 5, ‘c t c t c’: 4, ‘t t t c g’: 4, ‘a a a t c’: 3, ‘a a t c c’: 3, ‘a t a t c’: 3, ‘a t g c g’: 3, ‘c c g t g’: 3, ‘c g a a a’: 3}

BoW2 = {‘t c t c g’: 6, ‘c c a a a’: 4, ‘c t c t c’: 4, ‘g a a c g’: 4, ‘a a a c a’: 3, ‘a a a g t’: 3, ‘a a c g a’: 3, ‘a g a a c’: 3, ‘a t g c g’: 3, ‘c c g t g’: 3}

Table 3: Comparison of exemplary bag-of-words models of *B. oleracea* and *B. rapa*

	B. oleracea	B. rapa
„tctcg“	6	6
„tcgtc“	5	0
„ctctc“	4	4
„tttcg“	4	0
„aaatc“	3	0
„aatcc“	3	0
„atatc“	3	0
„atgcg“	3	3
„ccgtg“	3	3
„cgaaa“	3	0
„ccaaa“	0	4
„gaacg“	0	4
„aaaca“	0	3
„aaagt“	0	3
„aacga“	0	3
„agaac“	0	3

Results

In the following discussion, we introduce the bag-of-words analysis of the ITS and matK markers of the chosen representatives of the Brassicaceae family. We proceed from the analysis of relationships at the intraspecific level to the analysis of relationships between species across the whole family.

In the analysis, *n*-grams with *n* ranging from 2 to 50 were tested. Apart from the 2-gram case, all analyses approximately correspond to the official taxonomy (see the following analytical results). As an example, we present the analysis using 3-, 5-, and 10-grams of bases. However, we cannot say clearly what sizes of *n*-grams are the best to use for analysing these samples because there is no valid statistical verification of the usability of different *n*-gram sizes.

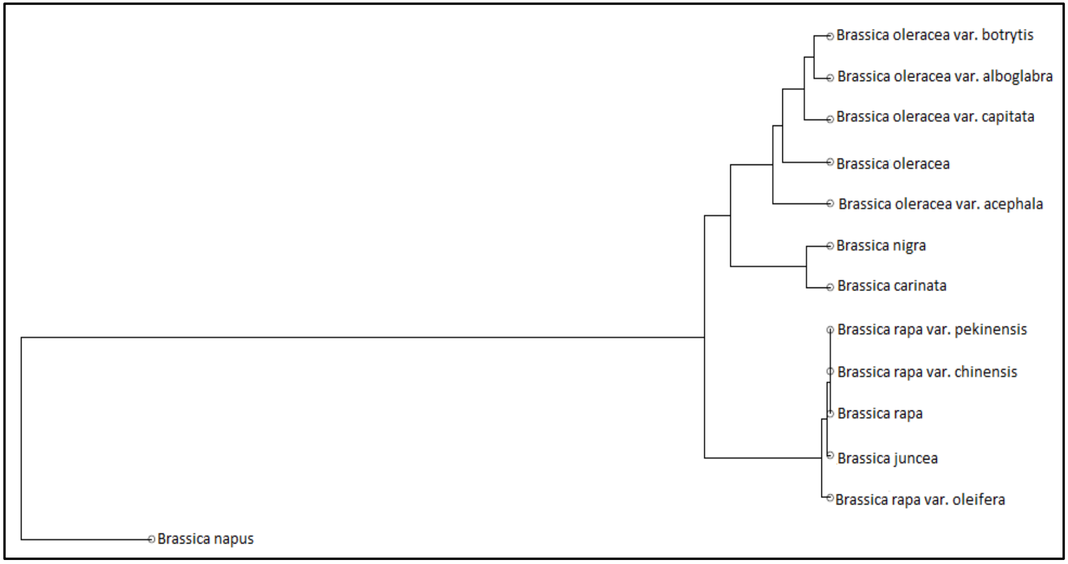
Brassica oleracea var. botrytis
Brassica oleracea var. alboglabra
Brassica oleracea var. capitata
Brassica oleracea
Brassica oleracea var. acephala
Brassica nigra
Brassica carinata
Brassica rapa var. pekinensis
Brassica rapa var. chinensis
Brassica rapa
Brassica juncea
Brassica rapa var. oleifera
Brassica napus

Figure 3: Dendrogram of the ITS marker and the used 3-grams for genus *Brassica*.

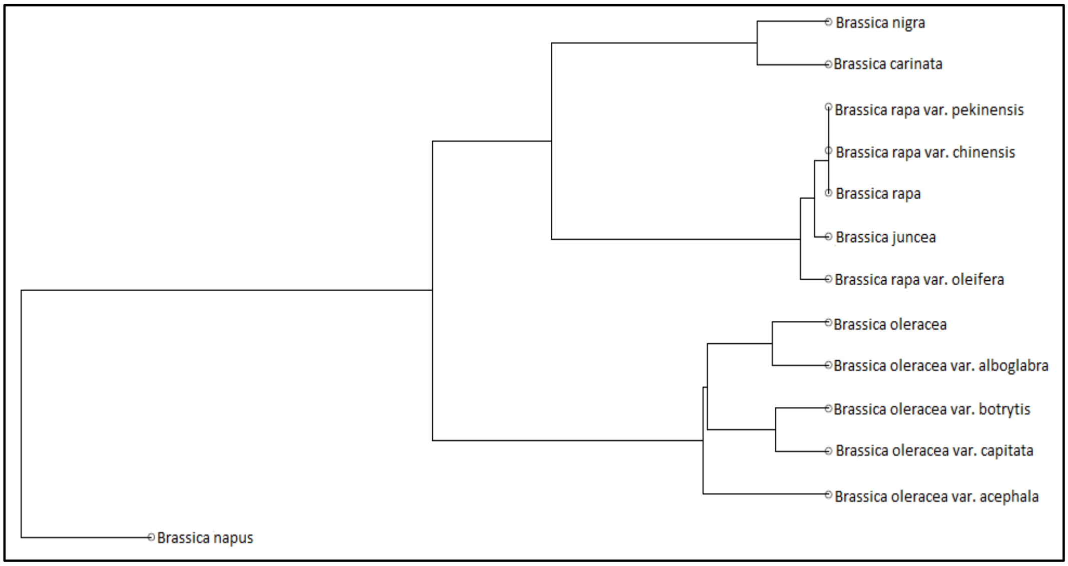
Brassica nigra
Brassica carinata
Brassica rapa var. pekinensis
Brassica rapa var. chinensis
Brassica rapa
Brassica juncea
Brassica rapa var. oleifera
Brassica oleracea
Brassica oleracea var. alboglabra
Brassica oleracea var. botrytis
Brassica oleracea var. capitata
Brassica oleracea var. acephala
Brassica napus

Figure 4: Dendrogram of the ITS marker and the used 5-grams for genus *Brassica.*

We can see that the bag-of-words model of 3-grams and 5-grams for the ITS marker expresses very well the relationship of species and varieties of the genus *Brassica* and the species resulting from their crossing (see U-model described earlier in the paper). On the bottom branch of the cluster, we find a hybrid *B. napus* separated from the remainder of the family. We suppose that the distance of this hybrid from the remaining genera is explained by its intensive cultivation for economic purposes. With further branching of the dendrogram, varieties of *B. oleracea* are separated from the *B. rapa* varieties and from *B. nigra* and *B. carinata* species. In general, we can conclude that species and their varieties are clustered in a common branch.

Within the branch of *B. rapa* species, the proximity of the *B. juncea* hybrid to *B. rapa* and its varieties is evident, which may indicate late separation of the *B. rapa* varieties and the late hybridisation of *B. juncea*. Within the branch of *B. oleracea*, wild species *B. oleracea var. oleracea,* and not wild species *B. oleracea var. alboglabra* (Chinese broccoli), is located. This confirms the premise ([28], p. 57) of the earlier separation of this variety of *B. oleracea* compared to other varieties.

The remaining species *B. nigra* and *B. carinata* are found in the case of 3-grams near the *B. oleracea* varieties and in the case of 5-grams near the *B. rapa* varieties. In the case of 3-grams corresponds to the predicted relationship of *B. carinata* to *B. nigra* and *B. oleracea*, which are progenitors of this hybrid (see U-model). However, a 5-gram analysis assesses this relationship to be the opposite. Thus, we can note that the change in *n*-gram size affects the results of analysis.

Based on the analysis presented herein, the bag-of-words model for genetic markers appears to be a sensitive method for reconstructing biological taxonomy at lower taxonomic levels. As we test the possibilities of taxonomy analysis based on the use of the bag-of-words model, we want to further verify the potential of this method for the same species using multiple molecular markers. In addition, the combination of different markers is a common procedure, and in taxonomy analysis, it refines the reconstruction of taxonomy relationships. For plants, a common combination of markers is the chloroplast gene and the ITS region of nuclear ribosomal DNA, due to their large variability even among closely related species (e.g. [29, 30]). For our chosen group of Brassicaceae family members, the chloroplast marker matK has been previously used, as is the ITS marker. On samples of these latter markers, we want to verify that the bag-of-words method is generally valid for different genetic material and to evaluate which marker is more appropriate.

In the following charts of hierarchical clustering of the bag-of-words analysis of maturase K, we see that even in the case of using this marker, the bag-of-words model is suitable for the reconstruction of interspecific relations. This time, we present the analysis on 3-grams and 10-grams to illustrate the difference in analytical results. In all cases, we can see the severance of species and their varieties into two separate branches corresponding to two vertices of the U-model – *B. oleracea* and *B. rapa*. In these branches, *B. oleracea* and *B. rapa* are associated with their hybrids, and *B. rapa* also contains its *B. oleifera* variety.

As can be seen from the graphs, the change in *n*-gram size used in the analysis may not only affect the clustering of species in the branches, it can also affect the length of individual branches, reflecting the degree of mutual similarity of the species and its varieties. Compared to the bag-of-words analysis of the ITS marker, the results of the analysis performed on the matK marker are less detailed because we only used samples that were available in the genetic database used by us. The difference between the two analyses in terms of expressing the biological taxonomy is that the matK analysis does not include *B. nigra* and *B. carinata* in a separate common branch. These two species are associated with the branches of both *B. oleracea* and *B. rapa.* Unlike the 5-grams of the ITS marker, there is no conflict with the taxonomy relationships described in the U-model – *B. nigra* has been associated with its hybrid *B. juncea* and *B. carinata* with its progenitor *B. oleracea*.

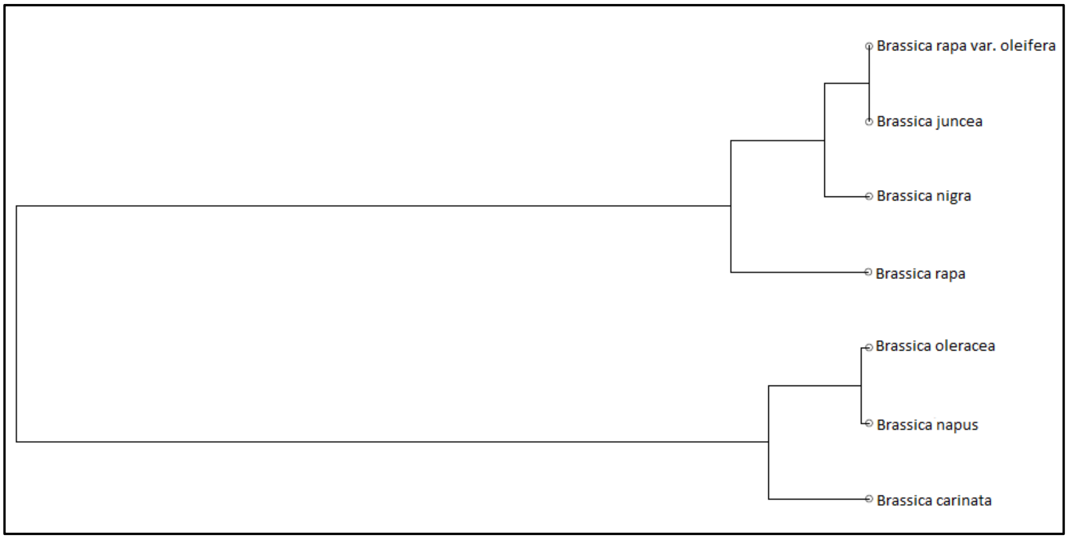

Figure 5: Dendrogram of the matK marker and the used 3-grams for genus *Brassica.*

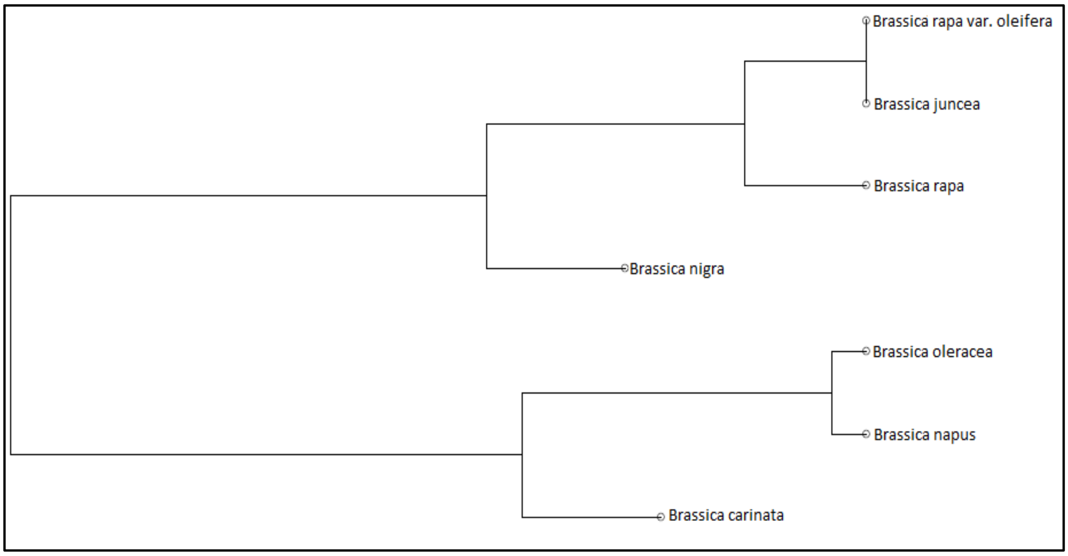

Figure 6: Dendrogram of the matK marker and the used 10-grams for genus *Brassica.*

The next step of the analysis is the extension of the sample of the Brassicaceae family to species not belonging to the genus *Brassica*. These species include *Sinapis alba* (white mustard), *Sinapis arvensis* (charlock mustard), *Eruca vesicaria* (rocket), *Raphanus sativus* (radish), *Orychophragmus violaceus* (Chinese violet cress), *Capsella bursa-pastoris* (shepherd's purse), *Allium ursinum* (bear's garlic), *Alliaria petiolata* (garlic mustard), *Arabidopsis* *thaliana* (thale cress), *Barbarea vulgaris* (bitter cress) and *Thlaspi arvense* (field pennycress). Their inclusion in the analysis makes it possible to illustrate the relationship of the individual species according to their classification into different genera.

In the ITS marker hierarchical cluster, we can see that *B. oleracea* is associated with its varieties in one branch, while another branch contains *B. rapa* with its varieties and its hybrid *B. juncea*. The third branch is populated by *B. nigra* and *B. carinata,* as well as *Sinapis, Eruca* and *Raphanus*. These genera, in comparison with others, seem to be closer to the *Brassica* genus and, at the same time, we can say that they are also among culturally exploited crops. In the last branch, we find *B. napus* and the other newly appointed representatives of the family Brassicaceae. The inclusion of *B. napus* in this cluster suggests that due to cultivation that has occurred in this species, it has a different character than other U-model hybrids. The location of the species *Orychophragmus violaceus* in the same cluster indicates its inclusion not only outside the genus *Brassica* but also outside the tribe Brassiceae (see [31]).

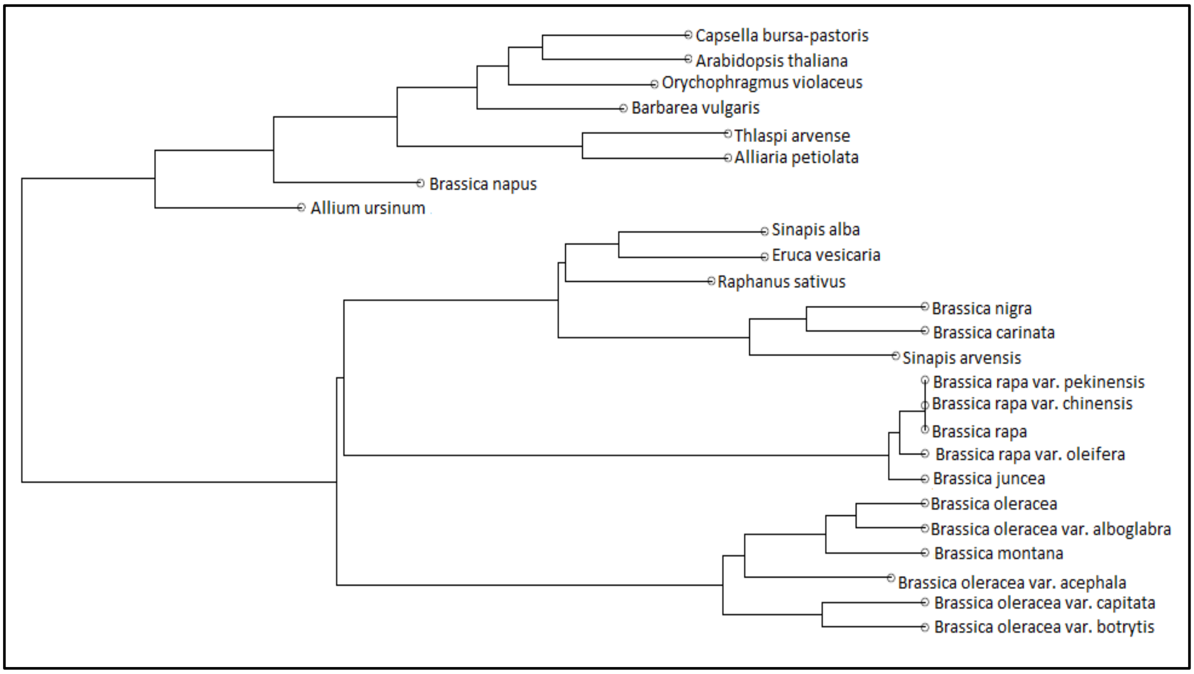

Figure 7: Dendrogram of the ITS marker and the used 5-grams for family Brassicaceae.

Conclusion and Discussion

Our goal was to introduce the use of the bag-of-words model for biological taxonomy analysis, while describing the relationships of chosen cultivated crops of the Brassicaceae family. For this purpose, we used two standard genetic markers – ITS and matK. At the same time, we ascertained that the *n*-gram size has an effect on the clustering results of taxonomy analysis, and higher *n*-grams (10-grams in our case) provide stable results that correspond to real taxonomy (affinity to species, genus and so on compared to the U-model). Using the hierarchical clustering charts, we have shown that the bag-of-words method is an appropriate yet simple tool for the reconstruction of taxonomy relationships at the species and intraspecific levels.

Acknowledgement

This work was supported by grants from the European Regional Development Fund: Project 'Sinophone Borderlands: Interaction at the Edges' (registration number: CZ.02.1.01/0.0/0.0/16_019/0000791).

References

[1] Soumya, G. K., Shibily, J., 2014. Text classification by augmenting bag of words (BOW) representation with co-occurrence feature. *OSR Journal of Computer Engineering (IOSR-JCE)*, 16 (1), 34–38.

[3] Boukhaled, M. A., Ganascia, J.-G., 2015. *Using Function Words for Authorship Attribution: Bag-Of-Words vs. Sequential Rules*. The 11th International Workshop on Natural Language Processing and Cognitive Science, Oct 2014, Venice, Italy. DE GRUYTER, Natural Language Processing and Cognitive Science Proceedings, 2014, 115–122, 2015.

[5] Serva, M., Petroni, I. F., 2008. Indo-European Languages Tree by Levenshtein Distance. *EPL (Europhysics Letters)*, 81, 680–685.

[7] Swadesh, M., 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of American Philosophical Society*, 96, 452–463.

[9] Swadesh, M., 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21, 121–137.

[11] Embleton, S., 2000. Lexicostatistics/ Glottochronology: from Swadesh to Sankoff to Starostin to future horizons. In: C. Renfrew, A. McMahon and L. Trask (eds.) *Time Depth in Historical Linguistics*, 1. Cambridge: McDonald Institute for Archaeological Research, pp. 143–165.

[14] Toldo, R., Castellani, U., Fusiello, A., 2009. A *bag of words* approach for 3D object categorization. In: Gagalowicz, A., Philips, W. (eds.) Computer vision/ computer graphics Collaboration techniques. MIRAGE 2009. *Lecture Notes in Computer Science*, Vol. 5496. Berlin: Springer.

[17] Zhang, Y., Jin, R., Zhou, Z. H., 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1 (1–4), 43–52.

[19] Bolshoy, A., Volkovich, Z., Kirzhner, V., et al., 2010. *Genome clustering from linguistic models to classification of genetic texts*. Berlin: Springer.

[21] Lovato, P., 2015. *Bag of words approaches for Bioinformatics*. Ph.D. thesis, Dept. of Computer Science, University of Verona, series TD-03-15.

[23] Harris, Z., 1954. Distributional structure. *Word*, 10 (2/3), 146–62.

[25] Huang, C. H., Sun, R., Hu, Y., et al., 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution*, 33(2), 394–412.

[27] Francisco-Ortega, J., Fuertes-Aguilar, J., Gómez Campo, C., et al., 1999. Internal transcribed spacer sequence phylogeny of *Crambe* L. (Brassicaceae): molecular data revealed two old world disjunctions. *Molecular Phylogenetics and Evolution*, 11, 361–380.

[29] Koch, M., Haubold, B., Mitchell-Olds, T., 2001. Molecular systematics of the Brassicaceae: evidence from coding plastidic matK and nuclear Chs sequences. *American Journal of Botany*, 88, 534–44.

[31] Koch, M., Sharma, A. K., Sharma, A., 2003. Molecular phylogenetics, evolution and population biology in Brassicaceae. *Plant Genome: Biodiversity and Evolution*, 1, 1–35.

[33] Warwick, S. I., Sauder, C., 2005. Phylogeny of tribe Brassiceae (Brassicaceae) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast trnL intron sequences. *Canadian Journal of Botany*, 83, 467–483.

[35] Warwick, S. I., Francis, A., Al-Shehbaz, A. I., 2006. Brassicaceae: Species checklist and database on CD-ROM. *Plant Systematics and Evolution*, 259, 249–258.

[38] Mummenhoff, K., Al-Shehbaz, I. A., Bakker, F. T., et al., 2005. Phylogeny, morphological evolution, and speciation of endemic Brassicaceae genera in the Cape flora of southern Africa. *Annals of the Missouri Botanical Garden*, 92, 400–424.

[40] Couvreur, T., Franzke, A., Al-Shehbaz, I. A., et al., 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Molecular Biology and Evolution*, 27, 55–71.

[42] Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., et al., 2011. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends in Plant Science*, 16(2), 108–116.

[44] Al-Shehbaz, A. I., Beilstein, M. A., Kellogg, E. A., 2006. Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Systematics and Evolution,* 259, 89–120.

[46] Hayek, A., 1911. Entwurf eines Cruciferensystems auf phylogenetischer Grundlage. *Beihefte zum Botanischen Centralblatt,* 27, 127–335.

[48] Nagaharu, U. 1935. Genome analysis in Brassica with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Journal of Japanese Botany*, 7, 389–452.

[50] Sadowski, J., Kole, C., 2011. *Genetics, genomics and breeding of vegetable Brassicas*. Enfield, NH, USA: Science Publishers.

[52] Schulz, O. E., Engler, A., Harms, H., 1936. Cruciferae, Die natürlichen Pflanzenfamilien. Leipzig, *Germany Verlag Von Wilhelm Engelmann*, 227–658.

[54] Bailey, C. D., Koch, M. A., Mayer, M., et al., 2006. Toward a global phylogeny of the Brassicaceae. *Molecular Biology and Evolution*, 23 2142–2160

[56] Liu, L., Zhao, B., Tan, D., Wang, J., 2012. Phylogenetic relationships of Brassicaceae species based on *mat*K sequences. *Pakistan Journal of Botany*, 44 (2), 619–626.

[58] Maggioni, L., 2015. *Domestication of Brassica oleracea L.* Doctoral Thesis No. 2015:74 Faculty of Landscape Architecture, Horticulture and Crop Production Science.

[60] Juniper, B. E., Watkins, R., Harris, S. A., 1998. The origin of the apple. *Acta Hor-ticulturae*, 484. 27–33.

[62] Crespo, M. B., Lledo, M. D., Fay, M. F., et al., 2000. Subtribe Vellinae (Brassiceae, Brassicaceae): a combined analysis of ITS nrDNA sequences and morphological data. *Annals of Botany*, 86, 53–62.

[63] 31. German, D. A., Friesen, N., Neuffer, B., et al., 2009. Contribution to ITS phylogeny of the Brassicaceae, with special reference to some Asian taxa. *Plant Systematics and Evolution*, 283, 33–56.

Appendix

	ITS1 5.8S ITS2	matK
Alliaria petiolata	KJ748666.1	–
Allium ursinum	KF419382.1	–
Arabidopsis thaliana	AJ232900.1	–
Barbarea vulgaris	AJ232915.1	–
Brassica carinata	DQ003690.1	AB354275.1
Brassica juncea	AF128093.1	AB354274.1
Brassica oleracea	AY722423.1	AB354271.1
Brassica oleracea var. acephala	GQ891869.1	–
Brassica oleracea var. alboglabra	GQ891870.1	–
Brassica oleracea var. botrytis	GQ891875.1	–
Brassica oleracea var. capitate	DQ003650.1	–
Brassica oleracea var. italic	KX709353.1	–
Brassica napus	AB456109.1	AB354273.1
Brassica nigra	DQ003644.1	AB354272.1
Brassica rapa	AF531563.1	AY541619
Brassica rapa var. chinensis	AF128095.1	–
Brassica rapa var. oleifera	GQ891873.1	AB354276.1
Brassica rapa var. pekinensis	AF128096.1	–
Capsella bursa-pastoris	KM892665.1	–
Eruca vesicaria	AY722459.1	–
Orychophragmus violaceus	AY722506.1	–
Raphanus sativus	AY722480.1	–
Sinapis alba	MG923992.1	–
Sinapis arvensis	AY722487.1	–
Thlaspi arvense	KM892656.1	–

Source: https://www.ncbi.nlm.nih.gov/genbank/.