AUTOMATED MOTIF IDENTIFICATION IN FOLKLORE TEXT CORPORA

Vilmos Voigt, Michael Preminger, László Ládi, Sándor Darányi

INTRODUCTION

Commissioned by the Hungarian Academy of Science, we carried out an experiment in 1996-1997 (grant no. 96-114/31). Aspects of the experiment fell into the categories of folklore text studies, automated classification, and visual presentation of information. The following results, of theoretical interest to both folklore studies and related fields and to information science, were reported and demonstrated at the 12th Congress of the International Society for Folk Narrative Research (ISFNR) in Göttingen, Germany on 26–31 July, 1998.

BACKGROUND CONSIDERATIONS

The variation of folklore texts is a universal phenomenon, as is shown by the many variants among collected texts. Type and motif indexes are used to register this variety. On the other hand, motifs as content markers for particular texts invite computer classification, preferably by multivariate statistical methods. Based on earlier attempts of more or less the same working group (Darányi & Ábrányi 1986; Darányi 1996a), our task was to use such statistical methods so that they yield artificial equivalents of traditional motifs, for the indexing of text collections. Further to this, we were interested in automating the process, from text input to their indexing by content extracts.

We note in passing that the concept of a motif goes back to classics of folklore and literary analysis (see the summary by Würzbach 1998). In our interpretation, a motif is a second-level aggregate of some first-level content criteria, e.g. the motif "Unpromising hero" (see Meletinsky 1958) is a compilation of 'hero', 'son', 'youngest', and the like. In other words, a motif is a broad concept related to those narrower terms which define it.

In library and information science however, it is an established practice to express such broader concepts from more detailed content criteria by automated classification, for example by singular value decomposition (SVD) (Deerwester *et al*. 1990), so that, as a prelude to information retrieval, the results can be used for an advanced type of indexing called latent semantic indexing (LSI) (Lochbaum & Streeter 1989). In short, we wanted to apply LSI in the domain of folklore.

Further, we knew from our earlier attempts that large-scale comparison of texts may result in output files of forbidding size, making personal computing obsolete and word processors crash. So in order to test the idea and improve the accuracy of the results, the first problem to be solved was dealing with relatively big text corpora in an automated manner.

THE FACTOR ANALYTIC MODEL

The factor analytic model is mainly used in psychometry to express unobservable variables of a certain setting, such as "mathematical talent" or "intelligence". This model assumes that the setting is composed of individual factors that characterize variables individually, with no interdependence among them, plus common factors that are shared by the variables (Mardia *et al*. 1979).

Operationally, we regarded a motif a stable correlate of word forms, extracted by statistical means from text corpora. In our implementation of the above approach, each common factor was regarded such a motif. In their entirety, common factors can be seen as the representation of the context as a whole.

THE TEXT VARIATION LABORATORY

A text variation laboratory was set up with computer-based classification and processing utilities. The idea was to generate output familiar to the folklorist, for inspection and quality assessment

(Figure 1). The laboratory concept was modular, to allow for future program extensions.

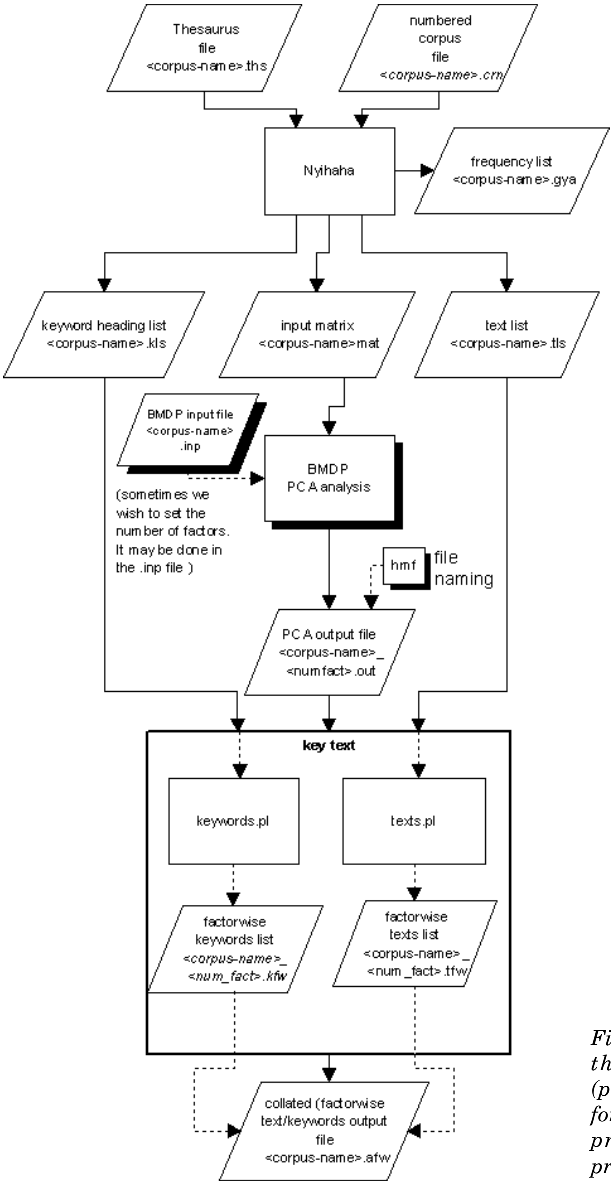


Figure 1. Flowchart of the text laboratory (parallelograms stand for files, rectangles for programs and/or processes).

For classification, we used a method called principal component factor analysis (PCFA) (Jackson 1991: 398, 402–403), as implemented in the BMDP statistical program package. PCFA is one method of estimating the factors for a given setting. All text processing utilities were written in Perl. For visualization of the results, we tested Manitou, a program written in Pascal by Zoltán Hajnal (Darányi *et al*. 1996b), MATLAB's mesh, waterfall and contour functions, and Microsoft Excel' 97 three-dimensional surface diagrams.

For test purposes, three corpora of traditional Hungarian texts from several genres were recorded in electronic format (2706 belief texts (Verebélyi 1998), 1500 political jokes (Katona 1994), 773 proverbs (Paczolay 1991), all in Hungarian). For input matrix generation, the following normalization procedure was followed: full texts were manually stemmed, orthographic and dialectal variants regarded as lexemes, and used for the derivation of keywords. A keyword was regarded a preferred expression for one or more orthographic or dialectal word forms, and declined nouns and inflected verbs. Because of the grammatical structure of negation in Hungarian, we distinguished between "positive" (affirmative) and "negative" keywords as well (ie. "eszik" (he/she/it eats) vs. "nem eszik" (he/she/it does not eat)).

Before processing, a list of stopwords was designed, to be excluded from the indexing procedure. Based on the stopword and keyword lists, a utility program wrote the input called a term-document matrix, which was subsequently exposed to principal component factor analysis (BMDP 4M). Based on the co-occurrences of keywords in documents as coded in the input matrix, PCFA created an n-dimensional vector space, the dimensions of which stood for higher-order content markers extruded from lower-order ones, and grouped documents in these, according to their content similarities. The program we used called such higher-order content markers principal components, so document and keyword membership was tabulated in them (Table 1 in the Appendix).

COMPUTATIONAL RESULTS

Using $n = 3, \ldots, k$ dimensional decomposition of the input matrix by PCFA, the following results were obtained:

Corpus	Belief texts	Jokes	Proverbs
Number of texts	2,706	1,5	773
Number of individual word forms	13,989	14,677	1,993
Number of keywords	1,837	771	239
Number of stopwords	1,52	997	189
Number of principal components ("motifs")	520	312	154

Our first experiments suggested that, due to their genre peculiarities, jokes and proverbs were more resistant to this approach. Therefore we concentrated our efforts on the belief texts instead.

INFORMATION VISUALIZATION RESULTS

As with factor analytical methods in general, so PCFA too puts documents in keyword space, where the numerical values of the documents and keywords in particular principal components correspond to their geometrical coordinates. In other words, PCFA computes an n-dimensional geometry which is of the same nature as any 3-dimensional Euclidean geometry, except for that it cannot be visualized in its totality. For example, as it was the case with the belief texts, 520 rectangular dimensions could not be shown in a Cartesian coordinate system. On the other hand, information visualization is known to help users in interpreting their findings. Therefore, to explore the membership of keywords in principal components, we opted for a planar conversion of the above 520-dimensional geometry in a 2-dimensional map. Earlier, similar attempts to map content included WEBSOM (Honkela *et al*. 1997), SPIRE (Wise 1999), VIBE (Olsen *et al*. 1993) and the use of 3-dimensional histograms (Häkkinen & Koikkalainen 1999: 71; Kurimo 1997: 56). Sample maps are shown in Figures 2-4, with different compression rates of content.

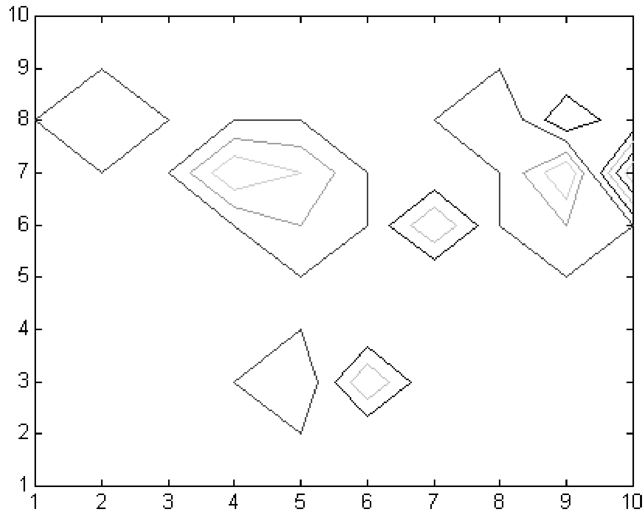


Figure 2. Beliefs corpus, groups of motifs (MATLAB contour diagram, 1837 keywords x 2706 texts; compression rate 1:10 x 1:10 [183 keywords : 52 motifs]).

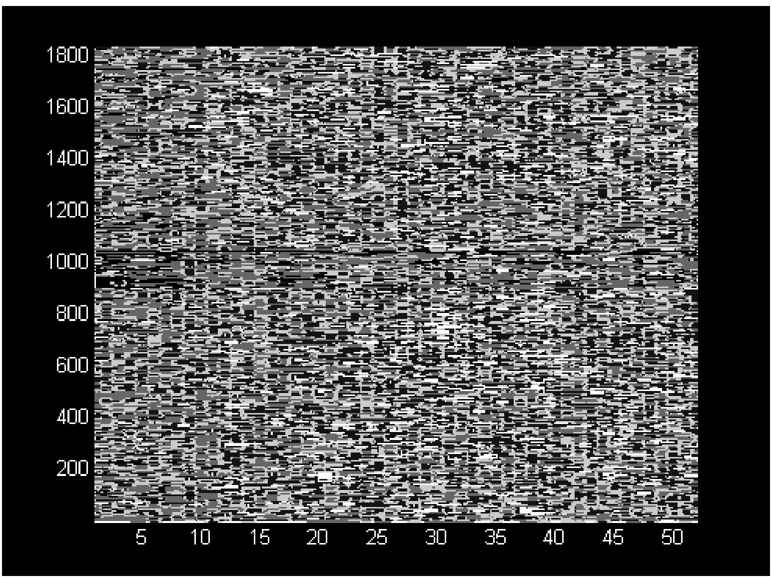
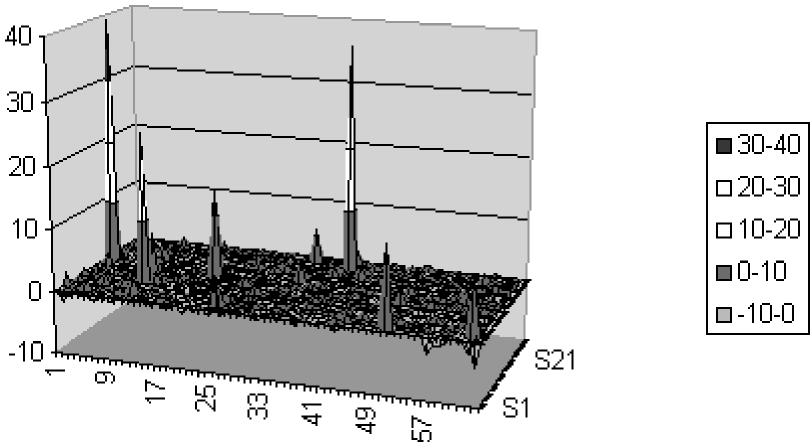


Figure 3. Beliefs corpus, groups of motifs (MATLAB contour diagram, 1837 keywords x 2706 texts; compression rate 1:0.98 x 1:10 [1800 keywords : 52 motifs])•.



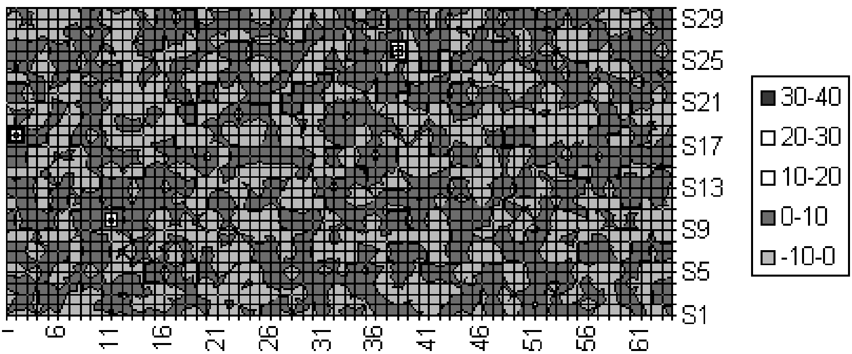


Figure 4. Enlarged segment of a "motif map" (Excel 3-dimensional histogram and its planar equivalent; motifs in columns and keywords in rows).

Based on similar experience, it seems increasingly possible to construct motif atlases for the display of content topologies. Such “thematic landscapes” could augment traditional motif indexes and become scholarly tools in an electronic environment. Further, inspired by the simplicity of displaying n-dimensional geometries in the plane, we started working on the interpretation of vector space word semantics in 1998. Some relevant theories, including semantic fields (Trier 1934), intensions and extensions (Carnap 1947), contextuality (Wittgenstein 1958), referential theories and a distinction between sense and meaning (Lyons 1968), will be discussed in a forthcoming publication by the fourth author (Darányi 2000, cf. Figure 5).

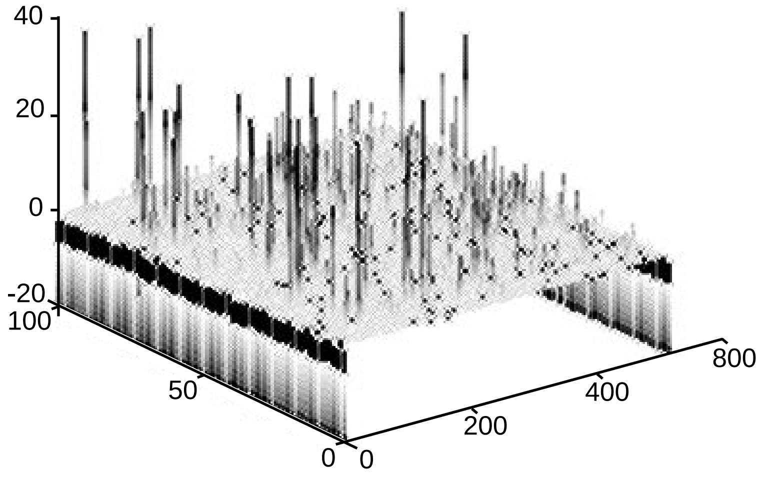


Figure 5. Beliefs corpus, n=520, semantic field of keywords 1-100 (MATLAB waterfall diagram, default angle).

CONCLUSION

Based on the above motif definition, the proposed new technique is capable of the large-scale comparison of original folklore texts (including Finno-Ugric ones), their automated grouping based on content similarities and differences, and the conclusions derived therefrom. More development work and the benchmarking of the results will be necessary. Furthermore, this model of thought offers a new angle on the study for word semantics and language philosophy as well.

Acknowledgements

The project was carried out in cooperation between the Department of Folklore, Eötvös Loránd University, Budapest, Hungary, and the Faculty of Journalism, Library and Information Science, Oslo College, Oslo, Norway, with computational and methodological help by the Center for Statistical and Mathematical Computing, Indiana University, Bloomington, USA. The authors are grateful to Robert Zawiasza (Central Library, József Attila University, Szeged, Hungary) for Perl programming and Zoltán Hajnal (Department of Theoretical Physics, Paderborn University, Germany) for Pascal programming and program development; to John Samuel and Dave Hart (Center for Statistical and Mathematical Computing, Indiana University, Bloomington, USA) for their constant availability and expert advice on BMDP and MATLAB; and to Kincsõ Verebélyi and Edina Batári (Department of Folklore, Eötvös Loránd University, Budapest, Hungary), for project management.

References

Because of the practical character of our paper, we give here only the source references, and do not refer to discussions of terms (e.g. motifs), or computer routines. It will be the task of a different paper to sum up similar attempts in recent folk narrative studies. The earlier summary (Voigt 1981) was written twenty years ago and refers only to the first, not in fact automated analyses of folklore texts. We note in passing that the full word-index to Verebélyi 1998 is now ready (Ládi 1999), it contains all the texts and is about 150 printed pages. We shall try to publish it in a separate volume.

Carnap, R. 1947. *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago.

Darányi, S. & Ábrányi, A. 1986. Ókori keleti epikus szövegek összehasonlítása számítógéppel [Computer Typology of Ancient Near Eastern Epic Texts]. *A mai folklór* [Folklore Today] 6. Budapest.

Darányi, S. 1996a. Hungarian Táltos Texts: Content Mapping from a Multivariate Perspective. *Folklore and the Encounters of Traditions: Proceedings of the Finnish-Hungarian Symposium, March 18–20, 1996, Jyväskylä, Finland. Research report 29.* Ed. by Suojanen, P. & Raittila, R. Jyväskylä, pp. 7–14.

Darányi, S. & Zawiasa, R. & Hajnal, Z. 1996b. Conceptual Mapping of a Database in the Humanities: First Results of an Experiment with *Sophia*. *Journal of Documentation*, 52(1), pp. 86–99.

Darányi, S. 2000. A Field Approach to Word Semantics. Accepted for publication in *Semiotica*.

Deerwester, S. & Dumais, S. T. & Furnas, G. W. & Landauer, T. K. & Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6), pp. 391–407.

Honkela, T. & Kaski, S. & Lagus, K. & Kohonen, T. 1997. WEBSOM – Self-Organizing Maps of Document Collections. *Triennial Report 1994-1996 of the Neural Networks Research Centre and Laboratory of Computer and Information Science.* Ed. by Alhoniemi, E. & Iivarinen, J. & Koivisto, L. Espoo, pp. 41–47.

Häkkinen, E. & Koikkalainen, P. 1999. The Neural Data Analysis Environment. Proceedings of the Workshop on Self-Organizing Maps [Espoo. Finland, June 1997], 69–74.

Jackson, E. J. 1991. A user's guide to principal components. New York.

Katona, I. 1994. Politikai vicceink 1945-tõl máig. "A helyzet reménytelen, de nem komoly". [Hungarian political jokes from 1945 onwards. "Our situation is hopeless but not serious".] Budapest.

Kurimo, M. 1997. Using SOM and LVQ for HMM Training. *Triennial Report of the Neural Networks Research Centre and Laboratory of Computer and Information Science.* Ed. by Alhoniemi, E. & Iivarinen, J. & Koivisto, L. Espoo, pp. 55–60.

Ládi, L. 1999. Mutató a magyar néphit szövegekhez. [Index to Folk Belief Texts collected by the Hungarian Section of Folklore

Fellows.] *Magyar Népköltési Gyûjtemény XX. kötet* [Collection of Hungarian Folk Poetry, vol. 20.] Budapest. *To appear.*

Lochbaum, K. E. & Streeter, L. A. 1989. Comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval. *Information Processing and Management* 25(6), 665–676.

Lyons, J. 1968. *Introduction to Theoretical Lingustics.* Cambridge.

Mardia, K. V. & Kent, J. T. & Bibby, J. M. 1979. *Multivariate Analysis.* London.

Meletinsky, E. M. 1958. *Geroi volshebnoi skazki. Proiskhozhdenie obraza.* [The hero of the Magic Tale. Its Origin.] Moskva.

Olsen, K. & Korfhage, R. R. & Sochats, K. M. & Spring, M. B. & Williams, J. G. 1993. Visualization of a document collection: the VIBE system. *Information Processing and Management* 29(1), 69–81.

Paczolay, Gy. 1991. 750 magyar közmondás és szólás. [750 Hungarian Proverbs. 2nd, enlarged edition – bilingual: Hungarian and English.] Veszprém.

Trier, J. 1934. Das sprachliche Feld. *Neue Jahrbücher für Wissenschaft und Jugendbildung* 10, 428–449, Leipzig.

Verebélyi, K. 1998. Néphit szövegek. [Folk Belief Texts collected by the Hungarian Section of Folklore Fellows.] Magyar Népköltési Gyûjtemény XIX. kötet [Collection of Hungarian Folk Poetry, vol. 19.] Budapest.

Voigt, V. 1981. Computertechnik und -analyse. *Enziklopädie des Märchens. Herausgegeben von Kurt Ranke.* Band 3. Berlin – New York, pp. 111–123.

Wise, J.A. 1999. The Ecological Approach to Text Visualization. *Journal of the American Society for Information Science* 50(13), 1224–1233.

Wittgenstein, L. 1958. *Philosophical investigations*. Oxford.

Würzbach, N. 1998. Motiv. *Enziklopädie des Märchens. Begründet von Kurt Ranke.* Band 9. Berlin, New York, pp. 947–954.

Appendix

Table 1: Sample output texts from Verebélyi 1998: Principal component ("motif") No. 452 ["Shaking one's beard causes storm"].

Keywords in Hungarian	Keywords in English	Occurr-ences	Weight in principal component (principal component scores)
vihar	[storm]	19	29.181
szakáll	[beard]	5	15.202
megráz	[shake]	7	11.885
gergely	[Gregory]	4	10.615
illés	[Elias]	4	2.919

indicate keywords, [ ] their translations. The individual word forms that were grouped under particular keywords are listed below them. Numbers refer to belief texts in the database, words in boldtype are keywords found in the texts and used for their indexing.

#vihar [storm]

vihar
viharos
vihart
vihartól
viharban
zivatarokat
zivatart
zivatar
zivataros
fergeteg

199 = Ha **vihar** van akkor szentült gyertyát kell égetni, hogy a villám a házba bele ne üssön, mert a villám az Isten ostora és ahová az leüt, arra a helyre valaki átkot mondott és nem tiszta hely.

852 = A **viharban** sárkányok szállanak s azok döntenek le mindent farkuk csapásával.

928 = Oly haranggal, mellyel már felakasztott embernek harangoztak, nem szabad **vihar** elé harangozni.

1074 = Ha úgy villámlik és dörög az ég, hogy **vihartól** kell félni, akkor baltát kell az udvaron egy fába vágni s nem lesz jég. Vagy ha már jég esik, hirtelen felkapni pár szemet s tüzbe dobni, hogy ne ártson a jég.

1396 = Mikor a halak a víz szinire dobájják magokat, akkor **vihar** lesz.

2062 = **Viharos** idôben meggyujtják a gyertyaszentelôkor szentelt gyertyát, a villám nem üt be a házba.

2367 = Ha **vihar** jön, akkor szentelt barkát égetnek a kemencében és szentelt gyertyát gyújtanak, hogy elmenejen másfelé.

2389 = Illés (júl. 20.) nagy **vihart** szokott hozni.

2516 = **Viharos** idô közeledtével a felhôk felé keresztet vetnek a levegôbe.

328 = **Illés** és szent Anna napján mindig nagy **zivatarokat** várnak. Innen azután **Illésnek** a neve: Szelethozó **Illés**.

387 = Mikor meg nagy **zivatar** van, fôleg ha jég esik, a baltát fokával a ház elô állítják, gondolván, hogy ez megvédi a házat.

490 = Ha a **zivataros** idôben szentelt gyertyát gyújtunk, az megvédi a házat a villámcsapástól.

1789 = A harangszó más faluba zavarja a jég**zivatart**.

1879 = **Zivatar** alkalmával a ház elibe baltát szokás tenni, hogy a villám a házba ne csapjon.

2246 = Ha a varjuk csoportosan keringenek **zivatar** lesz.

2385 = **Gergely** (márc. 12.) ha **szakállát** megrázza, **zivatar** lesz.

2492 = Mikor a barom az orrát “nehéz járású felhô után nyujtja és szagulja azt” – akkor egész biztosan **zivatar** várható, jéggel.

2515 = Urnapkor az utcán négy sátort állitanak fel zöld galyakból, melyeket kendôkkel, lepedôkkel, cserépvirágokkal diszitenek fel. Midôn a körmenet tovább megy, a nép a sátorokat fosztogatni kezdi; mindenki igyekszik egy-egy zöld ágat szerezni, mert ez épúgy mint a szentelt barka és búza tûzbe vetve, eloszlatja a **zivatart**. A melyik virág az urnapi sátorokban van elhelyezve, el fog száradni.

1147 = Ha télen valaki katonával álmodik, **fergeteg** lesz.

#*szakáll* [beard]

szakáll
szakállas
szakállos
szakálla
szakállát

514 = Ha valakinek a **szakálla** viszket **szakállas** vendége jön.

771 = A gyûjteményben két ízben is megkíséreltem bemutatni azt a társalgási modort, folyamatot, amely a falu egyszerû gyermekei között van, persze a beszéd tárgyát a gyûjtés szempontjából választva meg. A szavak lejegyzésénél lehetôleg kiejtés- és hangzási hûségre törekedtem.

Hely: Rátközberencsen a Jóni András háza. Augusztus van. A vájogból rakott spór vígon dúrozsol a pitvarban. Jóni Andrásné a vacsora készités körül forog, közben-közben élénken felel komja-asszonya szavaira, ki a küszöböt nyergelve tartja ôt szóval. Vendég-asszony: Hogy pattog a za tûz kifelé, Komámasszony! Még valami haragos vendëgfog jönni. Jóniné: Tán igaz a! Én sose hittem a zijet. Aszt mongyák, ha a zember szemôdöke viszket, új embert lát; ha a szeme viszket, sírni fog. Aszt elhiszem, hogy ha szarka csörög a ház tetején, hogy akkor vendëg jön. V[endégasszony]: Hát mit mongyëk a zember, komámasszony!? Nekem a multkor a zállam viszketett, asztán csakugyan **szakállos** vendëgem jött, a Ferenc bácsi. J[óniné]: Nem tán!? V.: Úgy-a!...(...)

1564 = Ha az állunk viszket, **szakállas** vendégünk érkezik.

2385 = **Gergely** (márc. 12.) ha **szakállát megrázza**, **zivatar** lesz.

#megráz [shake]

megrázza
megráz
megrázni
megrázzák
megrázkódik

16 = Ha a gyermek hideglelôs, az apja fogja a gyerek ingét és napfelkelte elôtt kiteszi az udvarra és azt mondja: “Alsó, felsô szomszéd, szégyeljétek magatokat, a én kis gyermekemet a hideg leli!” Ezután egy fát háromszor **megráz**, az inget hátra dobja és szerintük elmúlik a gyermek hideglelése.

430 = Feldebrô községben (Heves-m.) Nagyszombat napján azok, akik nem mennek templomba nagy szorongva várják a harang megszólalását. S midôn megszólal a harang, akkor az udvarban és a kertben lévô gyümölcsfákat mindet iparkodnak **megrázni**. Teszik pedig ezt azért, hogy az Úr Jézus sok gyümölcsöt adjon.

1922 = Mikor a tehén bornya egyhetes, meghivják a környék 8-10 gyermekét, közben a földre ültetik ôket s a tehén összegyûjtött s felforralt tejét egy nagy tálból kanalazva megetetik velök. Addig nem kelnek fel, míg rostán keresztül le nem öntik ôket vizzel. Ezután kimennek s az udvaron levô kerítés karóit **megrázzák**, hogy a bornyú “jó futós” legyen. (Futós alatt vidámságot, egészséget, virgoncságot értik) Az elfogyasztott tejet “fröccstej”-nek hivják.

2026 = Nagypénteken az elsô kerepeléskor szokták **megrázni** a gyümölcsfákat, hogy sok termés legyen.

2385 = **Gergely** (márc. 12.) ha **szakállát megrázza**, **zivatar** lesz.

2572 = Ha valaki pénzt akar szerezni, az három nap és három éjjel se ne egyék, se ne igyék egy cseppet sem, hanem imádkozzék, midôn a hold egy hónapban kétszer megujul; járjon arra, amerre a munkás emberek mennek, de ne szóljon egy fél betût sem hozzájuk, de ne is feleljen. akkor annak a harmadik éjjel megjelenik egy aranygyapjas bárány és **megrázkódik**, mire hullnak a csengô aranyak. Ha azonban az ember megszólamlik, vagy örömében fölsikolt, akkor a bárány ujra **megrázkódik** és elviszi az aranyakat.

15 = Aki elôször dagaszt tésztát és a tésztával egy kis gyerek ora alját megkeni, szerintük annak a gyereknek nem lesz bajusza.

#*Gergely* [Gregory]
Gergely
Gergelytôl

389 = A tavaszi idôkre van a népnek bizonyos általános tapasztalata. Igy tart **Gergelytôl** és a “fagyos szentektôl”. A “fagyos szentek”, akiknek napjáig még mindig féltik a gyönge termést: “Pongrác, Szervác és Bonifác.”

2238 = Ha Mátyás napján hideg van, hamarosan lesz az olvadás, ha meleg, akkor késôbben. Mátyás hidat ront vagy hidat épít. Mátyás, **Gergely** két rössz ember. **Gergely** azt mondta, hogy ha ô annyira belenyúlna a februárba mint Mátyás, akkor a tehénben még a borjut is megfagyasztaná.

2385 = **Gergely** (márc. 12.) ha **szakállát megrázza**, **zivatar** lesz.

#*Illés* [Elias]
Illésnek
Illés

328 = **Illés** és szent Anna napján mindig nagy **zivatarokat** várnak. Innen azután **Illésnek** a neve: Szelethozó **Illés**.

2389 = **Illés** (júl. 20.) nagy **vihart** szokott hozni.