Hebrew University of Jerusalem

Department of Linguistics

Seminar paper

“Statistical methods in lexical typology: clustering of cutting and breaking events”

Name: Daniel Levin

ID: 336462874

Course of studies: B.A in Functional Linguistics

Email: daniil.levin@mail.huji.ac.il

Academic Supervisor: Dr. Alena Witzlack-Makarevich

Date of submission: 01.10.2020

Keywords: lexical typology; cut and break; separation events; verb semantics; Korean; Hebrew; Russian

1. Introduction

Events of “cutting and breaking” (henceforth C&B events) occur in every language of the world. It can be taken for granted that the broadest features of separation events are not particular to any culture. Nevertheless, the C&B events are perceived differently, which may arise from the diversity of tools and methods used for cutting and breaking throughout the world. One could expect that differences in “separation” techniques should form the way we encode semantic categories of C&B verbs.

In this paper I am going to strengthen or weaken this assumption by addressing the domain of C&B verbs from the perspective of lexical typology. My goal is to replicate the research, conducted by Majid, Boster, and Bowerman (2008). The research will be conducted on three typologically diverse languages: Modern Hebrew, Russian and Korean. On the basis of these languages I will verify whether the lexical categorization is constrained across languages universally (Majid et al. 2008) or free to vary according to particular ecology, evolution, and practices of a community. The main task is to classify the C&B domain based on 61 Cut and Break non-verbal video stimuli (Bohnemeyer, Bowerman, and Brown 2001). First, this will be done within each language in order to examine whether they comply with the general structure proposed in Majid et al. (2008). After that, the collected data will be considered as a whole in order to make a broader typological conclusion and again compare it to the results implied in the previously conducted research on C&B verbs.

Lexical typology studies organization of semantic fields. Languages tend to lexicalize semantic zones in different ways. Lexical typologists investigate how many lexical items can be in a language to cover all the meanings of a given field, what meanings are frequently opposed in languages, and what meanings, on the contrary, are usually expressed in one word. These questions can be divided into 3 groups (Koptjevskaja-Tamm, 2012: 374):

- Onomasiological: What meanings can or cannot be expressed by a single word in different languages?

- Semasiological: What are the different meanings that can be expressed by the same lexeme?
- Lexicon-Grammar Interaction: What are the typologically relevant features in the grammatical structures of the lexicon?

Two first approaches concern this paper directly. Nevertheless, the methodology is predominantly onomasiological, in that participants were required to describe each of the 61 videoclips. Eliciting the corresponding semasiological data would have been more difficult. It would have involved asking each informant to identify, for each of the examined verbs, which of the videoclips is the best illustration of the meaning of the verb. That would have been very time-consuming and probably imprecise due to the lack of correspondence between a situation depicted by the video and a certain verb, presented in the language.

Verbs are the main category addressed in this paper. This is supported by their relatively high lexical variability as opposed to semantically more robust parts of speech, e.g., nouns. The semantic behavior of verbs is proved to be unpredictable even in closely related languages (Majid, Gullberg, Van Staden, and Boster 2007). Therefore, it is rather plausible to expect lexical dissimilarity of the verbs among the languages observed in this research.

Section 2 begins with an introduction to the previous research in the field. Section 3 describes the methods that were applied for collecting the data and consecutive statistical analysis. The results of the quantitative research are presented in Section 4 along with visualization sample. The paper ends with a discussion and further interpretation of the results. They are compared to the results in Majid et al. (2008) formerly.

2. Previous research

The major research that concerns semantic categorization of C&B events from the typological perspective, was conducted and published in a special issue from 2007-2008. This special issue of Cognitive Linguistics examines crosslinguistic universality and variation in the encoding of the events, involving ‘‘separation in the material integrity of objects’’ (Hale, and Keyser 1987). This domain was chosen for several reasons. First, the centrality of the actions of C&B to hominid cognition played a role. Second, the C&B verbs had been prominently addressed in discussions of universals of verb semantics and syntax. It is claimed that the underlying semantic structure of cutting-type verbs is distinct from that of breaking-type verbs (Majid, Bowerman, Van Staden, and Boster 2007). Moreover, this distinction is associated with different argument structure and syntactic privileges (Guerssel, Hale, Laughren, Levin, and Eagle 1985; Levin, Hovav, and Keyser 1995). The syntactic aspect of the C&B verbs was examined too (Gaby 2007; Essegbey 2007; Chen 2007; Ameka, and Essegbey 2007; Brown 2007; Bohnemeyer 2007). Despite of the crosslinguistic universality, preliminary crosslinguistic work (Pye 1993; Pye, Loeb, and Pao 1995) shows that C&B verbs have different extension patterns in different languages. Notwithstanding, Majid investigates universal dimensions according to which, speakers choose a lexeme, most of the languages tend to show unique distinctions. E.g., the *Yélî Dnye* (Papuan language of Rossel Island) verbs covering the C&B domain are all based on 'exotic' distinctions in mode of severance—coherent severance with the grain vs. against the grain, and incoherent severance (Levinson 2007).

The paper from the special issue, which is the most relevant for this research, is Majid et al. (2008). The authors conduct an extensive statistical overview of the data, considered in the research, and make a broader typological conclusion. Event descriptions of this extensive research were collected from speakers of 28 typologically, genetically, and geographically diverse languages. For each language there were between one and seven consultants. The data were collected using the aforementioned 61 videoclips (see Bohnemeyer et al. 2001; Appendix) depicting a wide range of C&B events with such variables as agent, instrument, object acted upon and manner of destruction. The analysis was focused solely on the categorization imposed by verbs, the use of other parts of speech was left aside.

After identifying standing out events (e.g., opening, peeling) and leaving them aside, four major dimensions were elicited, which the vast majority of the languages showed agreement with:

1) The predictability of the locus of separation in the affected object.
2) Separation of only two videoclips from the rest – tearing a piece of cloth with the hands.
3) Differentiation between "snapping" and "smashing" events.

4) Separation of only two videoclips from the rest – poking a hole in a piece of cloth stretched tautly between two tables.

After that, the suggestion that individual languages may differ from statistical averages – from the general structure, has been disproven by showing the high correlation between dimensions of the sample languages and the general solution. Further, the factor-analytic method (Romney, Weller, and Batchelder 1986) was used for capturing how similar languages correlate to their overall categorization pattern. The analysis has consistently affirmed the hypothesis that there is a shared structure across languages. Summing it all up, the authors made a conclusion that the extent to which the precise categories of C&B events encoded across languages forms a limited set. Regarding the differences between categories among single examined languages, it is proposed that discrete categories come from the fact that humans recognize lexical categories based on observation of correlations in the distribution of the features in the environment. Although, the precise categories recognized by the languages in the sample vary, they are highly constrained by the four dimensions proposed by the researchers . This statement is the main focus of the research conducted in this paper.

3. Data and method

3.1. Participants

Event descriptions were collected from native speakers of 3 typologically, genetically, and geographically diverse languages – Modern Hebrew, Russian and Korean. For each language there were between one and seven consultants (M = 4). In order to reduce influence of other languages, monolingual speakers were preferred.

3.2. Materials

The data were collected using a set of 61 videoclips depicting a wide range of events (Bohnemeyer et al. 2001). The clips' length varied from 2 to 34 s. Since the goal of the stimuli is to capture the widest range of separation variations in terms of instruments, agent, object and manner of destruction, the selection of the events shown in the clips was influenced by two factors. First of all, the previous cross-linguistic work by Pye et al. (1995) was taken into consideration, which highlighted potentially important distinctions in the domain of C&B events. The accent of the study was on the dimensions of distinction that go beyond those obvious from English and other familiar Indo-European languages. A second influence was the overview of children's errors of verb use in this domain. The final set of videoclips included a ''core" set of C&B events involving non-reversible separations, and a smaller set of reversible separations such as ''opening a teapot" and ''pulling apart paper cups", as well as two ''peeling" events (Majid et al. 2008; Appendix).

3.3. Data collection

The data collection took place in the framework of the course "Lexical typology". The task was divided between the participants of the course, in a way that each student was responsible for one informant and formatting the final table with inserted data. The video stimuli (Bohnemeyer et al. 2001) and a table with list of the clips were sent to the informants and they were asked to analyze the data as follows:

- display the clips in the fixed order they appear in the folder
- for each video make sure that the object of destruction and the instrument used are recognizable
- answer the question: "What happened in this clip?" or "What did the agent do?"
- extract the verb used in the answer and insert its infinitive form into the corresponding to the current video cell

The entire communication including the questions was held using the language being researched with the involved informant. The speakers were allowed to insert additional comments into the table, if necessary.

3.4. Coding

The research focuses on verbs since we are interested in precise classification of the change in an object from a state of integrity to a state of separation or material destruction. Particles, affixes, or other verbal constructions (e.g., serial verbs) tend to have a marginal relation to semantic domain of separation events. When these additions were used to describe the clips, they often formed “meta” information, e.g., in many serial verb languages (e.g., Ewe, Kilivila, Likpe), consultants’ descriptions of the C&B events included mention not only of the state-change but also of the subevent of taking control of the instrument – even if the instrument was already held by the agent at the start of the clip (Majid et al. 2008). Nevertheless, the assumption that verbal modifications (e.g., affixes, serial verb constructions) do not correlate with the caused state-change is not taken for granted. In order to address this issue, I will conduct an additional analysis of the prefixes in Russian, separately from verbs (see sections 3.5.3, 4.1.2.1) in order to identify what are the dimensions along which the clips are separated from the standpoint of prefixes in Russian.

The outlook of the categorization is that each different verb used by the informants is taken to define a category or “group” of events for them. The clips are further grouped according to these categories.

Technical tools used in this research are: programming language – Python (Van Rossum and Drake Jr 2009); data preprocessing – Pandas (McKinney 2010); clustering, dendrograms – SciPy (Virtanen et al. 2020), Scikit-learn (Pedregosa et al. 2011); plotting – Matplotlib (Hunter 2007); handling multi-dimensional arrays and matrices – NumPy (Harris et al. 2020).

3.5. Data preprocessing

3.5.1. General

Most of the notes added by the respondents had to do with expressed uncertainty or hesitation between two or more different lexemes. In this case both options were considered. In order to do that, an additional, mock speaker was added to the language with one of these options being adopted as a response for the relevant clip. Usually the speakers matched in their hesitation, which allowed to limit this artificial extension by single column. Empty cells of this “speaker” were filled with the most frequent answer for the given video. This strategy prevents odd data patterns because the variance it adds to the data is minimal. It imitates the data that already exists and at the same time counts all the verbs associated with the video among the informants. This method is especially useful in case of shortage of speakers. It is very plausible to assume that if there were more speakers, one of them would use the verb that came up to the respondents as an option. Otherwise such an artificial way of expanding the data should be avoided.

3.5.2. Hebrew

The responds of Hebrew required one additional step of preprocessing. Some of the informants related to the outcome of the action, which does not align with the question they were asked. For instance, clip 45 (poking a hole in cloth stretched between two tables with a twig) was described as ‘to make a hole’ (*la'asót khór* – לעשות חור). Such reaction is natural for speakers due to the fact that in everyday life the result of poking a hole is more important than how it was done. Instead of ignoring these samples, it was decided to take the semantics associated with the clip and encode it with a colloquial verb *‘לחורר’* (*lekhorer*) that embraces the semantics of making a hole into one lexeme. This allows retaining the focus of the description on the core of the event without neglecting the natural association of the respondents. The same steps were taken for other clips, where applicable.

After further inspection of Hebrew verbs, I decided to minimize grammatical influence of different patterns (binyanim – “conjugations”). Hebrew verbs are inflected according to specific patterns or derived stems, called forms or בניינים (/binjaˈnim/ *binyanim*, "constructions"); where vowels patterns (משקלים /miʃka'lim/ *mishkalim,* "scales"), prefixes, and suffixes are put into the (usually) three-letter roots from which the vast majority of Hebrew words are made ("Modern Hebrew verbs", n.d.). The core of this study lies in lexical constituent of verbs, whereas binyanim in Hebrew determine grammatical categories, such as voice and reflexivity. In the survey the same root was often used with different grammatical conjugations (binyanim) for the same clip. At first glance, it produces different verbs but knowledge of the essence of this variation that lies behind the Hebrew verbal system leads to the logical implication, namely, that three literals of the root determine the lexical component of the verb. This observation is crucial for the main goal of this research; therefore, I have extracted three-letter roots for each verb and executed the analysis over them. It decreased the overall number of the lexemes for Hebrew but significantly improved lexical correspondence of the data.

3.5.3. Russian

First thing that was modified for Russian verbs is omitting the reflexive suffix '*ся*' (*sja*). It does not contribute to the lexical aspect of the descriptions. The functions of this suffix are exclusively grammatical: it occurs with reflexive, reciprocal, intransitive or passive verbs.

To examine whether particles, adpositions, or other verbal constructions correlate with the core event of destruction, for each clip I composed a set of prefixes that were used for this clip. Cluster analysis of the clips with these sets as features defining points of distinction between the clips will classify C&B events into clusters according to the prefixes.

Due to the fact that prefixes were treated separately, the lexemes for the analysis of the verbs are plain roots without additional modifications. Thereby, resembling the result of breaking down the Hebrew verbs into roots, less lexemes and more sophisticated lexical correlation are obtained.

4. Results

To analyze the correspondence of the languages with the dimensions proposed by Majid et al. (2008), I created a videoclip-by-verb matrix for each language separately (with clips as rows and verbs as columns). For each clip, if a particular verb was used by the speakers X times for that scene then a ratio of this verb was coded; otherwise a zero. For instance, if a clip was described four times by the verb A then A divided by number of speakers/verbs used for this clip was inserted in the corresponding cell. In this manner the data shows how robust is the linkage between each verb and clip and not merely whether this verb has ever been used for this video. In other words, the confidence of the speakers was also encoded. This approach does not suffer from variations in the number of respondents, given that languages are analyzed separately. This is more desired to have 3-7 speakers for each language, but it is not obligatory for the approach that I have chosen. Either way the only language with a low number of speakers – 1 – is Korean. For Hebrew and Russian the variance should be even lower.

The main technique, used for the analysis of the matrices is cluster analysis, a statistical technique that groups items together based on their similarity. The variance of each cluster is minimized, while variance across the clusters is maximized. The encoded matrices of each language were analyzed using hierarchical clustering, a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This family of algorithms is particularly useful for the purpose of encoding lexical domains due to its hierarchical representation of the data – this hierarchy of clusters is represented as a dendrogram, which shows the sequence of cluster fusion and the distance at which each fusion took place. The root of the dendrogram is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample – video stimuli. The hierarchical structure is pertinent to this study due to the assumption that verbs of material destruction have a hierarchical relationship, triggered by hyponyms and hypernyms. My goal is to capture this structure where it exists. Because of the structural asset of clustering, this type of unsupervised learning was preferred over other multivariate techniques, which are yet more powerful computationally.

For this research, agglomerative clustering was chosen, determined by the bottom-up approach: each clip starts in its own cluster, and clusters are successively merged together based on their similarity. The linkage criteria determines the metric used for the merge strategy. Linkage choice can influence the resulting structure of the dendrogram and the size of the clusters. Assuming that each language divides the domain of C&B events to several subdomains, each containing multiple items (videos), complete linkage, also known by the Farthest Point Algorithm (Everitt et al., 2001; Defays, 1977), is optimal. Complete-linkage clustering is one of several methods of agglomerative hierarchical clustering. This algorithm minimizes the maximum distance between observations of pairs of clusters. This is equivalent to choosing the cluster pair whose merge has the smallest diameter. This complete-link merge criterion is non-local; the entire structure of the clustering can influence merge decisions, which is very important for capturing the global picture of division of the C&B domain. This results in a preference for compact clusters with small diameters over long, straggly clusters. Moreover, complete linkage clustering avoids a drawback of the alternative single linkage method – the so-called *chaining phenomenon*, where clusters formed via single linkage and therefore, clustering may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very distant from each other. In other words, single linkage is not robust to noisy data, which is expected to be steadily present in the dataset of C&B verbs. Complete linkage tends to find compact clusters of evenly distributed diameters and thus, to produce bigger groups from the early stage on. Hence, complete linkage manages to capture several groups of similarity at once. This feature may be very helpful for identifying bigger subdomains of "cutting and breaking" instead of focusing on outliers and joining them one by one to bigger clusters.

Agglomerative clustering ensures that clips, described with the same verb(s) and consequently more similar to one another, are more likely to be in the same cluster. Clips that are never described by the same verb(s) will end up in separate clusters.

The results of the cluster analyses for the three languages are presented in Figures 1a–1d. These are dendrograms, where each videoclip is depicted as a separate row and identified by the clip number followed by a short abbreviation, describing the event captured in the clip. Colored clusters capture the main groupings based on the (dis)similarity of the verbs across the whole dataset. Videoclips that were described by a verb that was not used for other clips are clustered with the longest lines linking from left (leaf) to right (root) without intermediate splits. Clusters that were split closer to the left are denser due to the fact that the distance between its elements is shorter. This density equals to a group of closely related events, that were often described by the same verbs. Order and location of different clusters in the dendrogram is meaningless due to the deterministic nature of the algorithm. In other words, two clusters placed in the bottom do not have special correlation as opposed to clusters placed in the opposite parts of the tree. The most important visual indicator of similarity is splits and their x coordinate.

On each dendrogram there is a vertical dotted line plotted, indicating the longest horizontal distance without any horizontal line passing through it. This coordinate on the x axis is selected manually in order to simplify visual interpretability and also to divide the domain into well-defined clusters. A horizontal line is drawn through this coordinate. The number of horizontal lines this newly created vertical line passes is equal to the number of clusters.

If a cluster is embedded within a larger cluster, one can conclude that there is a hierarchical relationship amongst the verbs that were used to describe clips: there was at least one verb that was used for all the clips in the most encompassing cluster, and at least one verb used across the clips in the sub-cluster. Such structure is equivalent to hyponyms, being the embedded verbs and hypernyms – the embedding ones.

For all three languages the events of peeling are clustered separately from all the rest (clips 29, 30). This means that each of the examined languages features a special verb for peeling, that was not used for any other event. The same holds for clip number 7 – pushing a chair back from the table. Such clips form the entropy group, they are henceforth called data outliers.

For Russian and Hebrew there is an additional clip which groups a cluster of its own – clip number 15 – sawing a stick propped between two tables in half. For both languages, this clip was described by a verb that was not used for any other event. Such verbs cannot constitute the basis for further grouping; hence they stand alone. In Korean, however, this clip ended up in a group of two events – together with 'slicing a carrot'. This may have happened due to the identical manner of the action – sawing across the object back and forth repetitively. Neither Hebrew nor Russian showcase this distinction.

For Hebrew and Korean there is another outlier – clip number 11 has formed a single item cluster – pulling two paper cups apart by hand. This clip was described by verbs that were not used for any other event in both languages. Nonetheless, in Russian, this clip was classified to the cluster, that included the most distanced events that were not close enough to any other of the major groups (see the purple branch of the figure 1c). Most of the events of this dispersed cluster contain a sudden separation of the object into two parts, e.g., poking with a sharp long instrument.

4.1. Clustering of the individual languages

4.1.1. Hebrew

Hebrew has shown a huge diversity of used verbs – 54 roots were extracted. This might have occurred due to the highest number of respondents – 7. Hebrew speakers tended to be precise in their descriptions and used hyponyms that are usually replaced by their hypernyms in colloquial register. Such tendency is represented visually in organization of 5 bigger clusters. The overall structure is quite flat and even. At first glance, the 61 clips depicting C&B events were

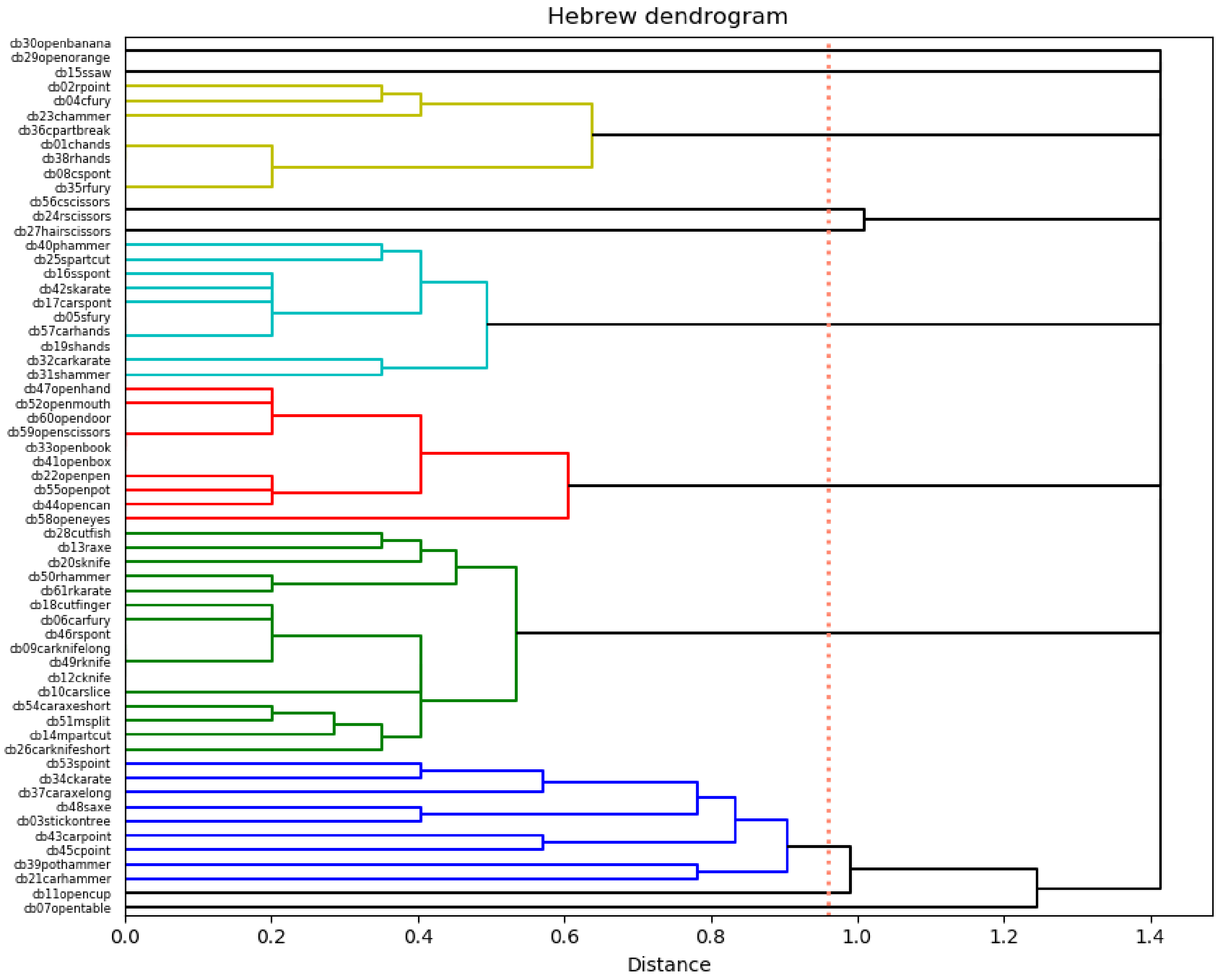

Figure 1a. *Hebrew dendrogram*

translated into 5 categories, the ones colored differently (see Figure 1a). Turquoise and blue branches are the most well-defined, since their split has occurred close to the left side on the x axis, which means that the distance between the components is short.

First let us address the clips that were clustered separately and have not been discussed yet. Hebrew treats three clips that included the use of scissors separately (black branch between the yellow and turquoise ones). None of the remaining languages make this distinction. Therefore, we can conclude that the verb ‘לגזור’ (*ligzor*) implicitly encodes events of cutting-with-scissors. A further question, whether this verb has to do with number of blades rather than with the specific instrument, can not be answered based on the chosen sample. Another interesting distinction observed only in Hebrew is the opposition of the event of eyes opening to all the rest of the opening events (see the red cluster of the Figure 1a).

Otherwise, Hebrew has five distinct clusters (see Figure 1a). From top to bottom, the yellow cluster includes events of tearing, the turquoise events of breaking with relatively low predictability of the locus of separation, red events of opening, green events of cutting that also include chopping events – events that are located on an intermediate location on the scale of predictability, and lastly the blue ones – smashing events with the lowest predictability, probably due to the power required to separate the objects of this cluster that cannot be combined with accuracy. Two subclusters of the green branch can be described in terms of predictability, the upper one contains hitting movements that are usually associated with lower predictability, although the locus of separation is directly connected to the place where the blow landed. The lower one, in contrary, almost exclusively contains events with sharp blades and precisely predictable place of separation. Thus, we see that the resulting clustering corresponds to the scale of predictability of the locus of separation. Additionally, the events of tearing are treated separately, as well as those of opening.

4.1.2. Russian

4.1.2.1. Clustering by prefixes

The main purpose of the analysis of the prefixes is to identify whether they influence the lexical dimensions, usually shaped by verbs in many languages (Majid et al. 2008). I will track whether the way they organize the clusters corresponds to the organization implied by verbs, both in Russian and in the rest of the languages. The Russian respondents used 10 different prefixes in the whole survey, with one of them being present in two its allomorphs (раз-, рас- : *raz-, ras-).* The dendrogram (see figure 1b) shows that the prefixes formed one big cluster of events – the turquoise branch in the top and several smaller ones – red, green, and blue in the bottom. The blue subcluster contains events that have a shared quality, albeit quite abstract – two parts of the object are being moved against each other, with eventual detachment or without. This secondary distinction is encoded in the internal subdivision of the blue cluster. The smaller green branch concerns mostly irreversible events. The red branch clearly defines events of partial breaking – making single incision in watermelon (14), tearing a cloth about halfway through with two hands (36), snapping twig with two hands, without halving it completely (25),

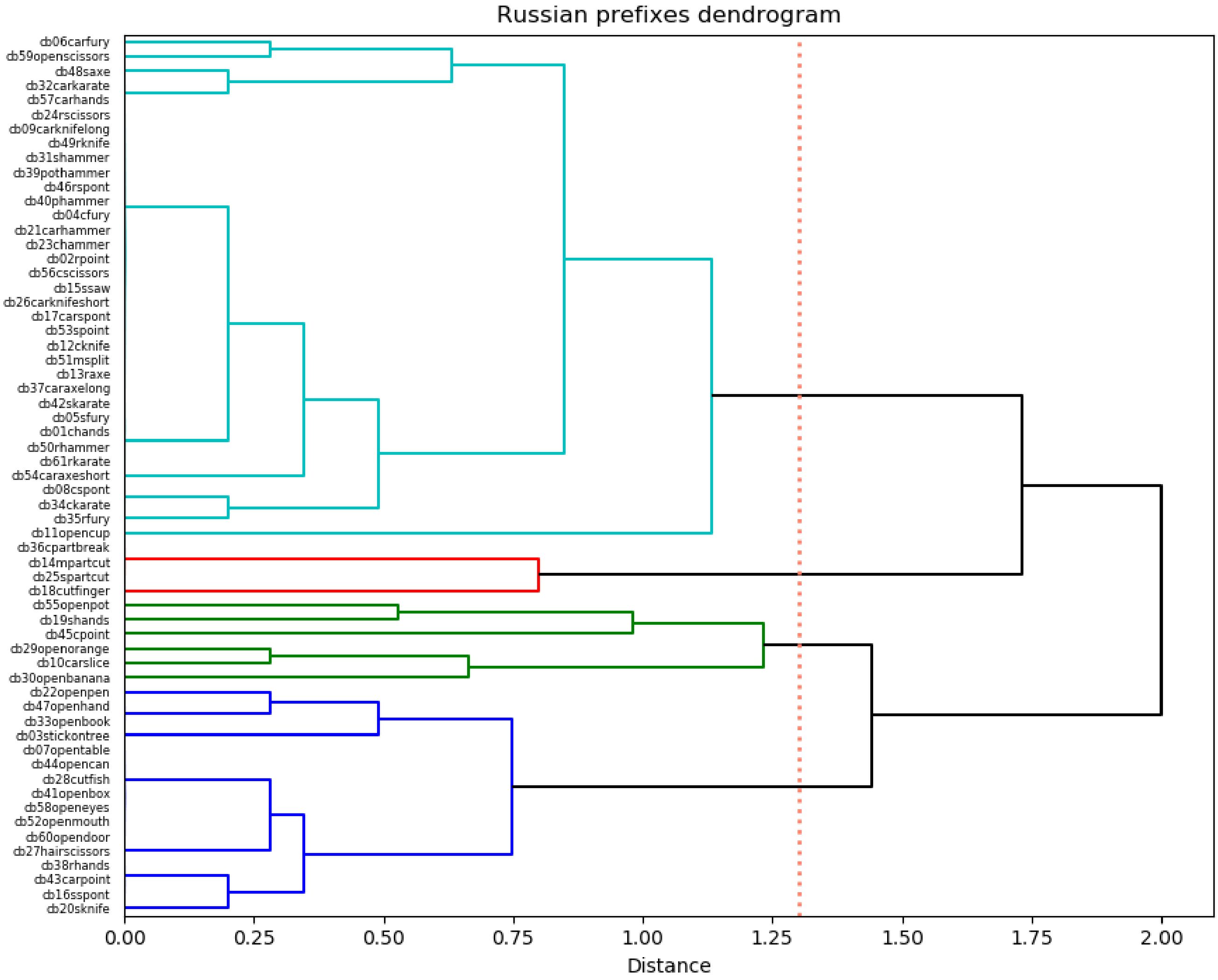

Figure 1b. *Russian prefixes dendrogram*

and cutting a finger superficially (18). This distinction correlates to the final state of the object being destructed. The last and the biggest group – turquoise cluster – combines various events. Most of them include irreversible separation. All the clips, where the object was broken into multiple pieces, belong to this cluster.

In general, the clusters formed by prefixes as features are not absolutely consistent. There is only one small cluster that defines a very clear group of the events without exceptions – the red branch with events of partial separation. The blue and the turquoise branches cannot be distinguished clearly. This may have happened due to the fact that prefixes are less robust than verbs and can be interchangeable. Verbal prefixes are also more prone to semantic shifts – both synchronically and diachronically.

To sum it up, there is an exception in almost every cluster, in other words, the same prefix can encode different lexical refinements depending on the verb it is attached to. This note affirms that the global classification of the C&B events should not rely on such verbal modifications as prefixes. Consequently, verbs should be treated separately from prefixes, particularly in Russian. This will guarantee lexical clearness expressed by the verbal stems and thus, noise triggered by the prefixes will be reduced.

4.1.2.2. Clustering by verbs

Russian speakers have used 21 different verbal stems. In this regard, they were more consistent than Hebrew speakers. From the other side there were less Russian informants than in Hebrew. Either way it seems like the core group of C&B verbs in Russian is smaller than that of Hebrew.

Except for the aforementioned outliers (7, 15, 29, 30) Russian has not elicited any outstanding events, i.e. events described uniquely (see figure 1c). The overall structure of Russian is similar to that of Hebrew. However, the maximal number of clusters among all the languages was formed in Russian – 7. They are distributed evenly over the sample. The blue (bottom), turquoise, yellow, and blue (top) are denser – their leftmost splits have occurred within distance of 0.6. The purple one is the most diverse one, since its primary split occurs on distance of 1.2; thus, it resembles the blue cluster formed in Hebrew (see figure 1a). Closer inspection of the branches and analysis of the events contained in them leads to the following classification (from top to bottom): blue (top) cluster is made of snapping events. The objects of this cluster are always long and thin objects – sticks and carrots. For the separation either hands or no instrument was used. This group is made of events that contain actions that can be generalized as actions of breaking suddenly and completely, with a sharp cracking sound. The predictability of the locus of these events is relatively low. The next cluster, the yellow one, combines almost every clip where hammer was used as an instrument. The only event that included hammer as the tool of destruction and was not included in the yellow cluster is event number 23 (chop a cloth stretched between two tables into two pieces with two hammer blows). This clip stands out because it

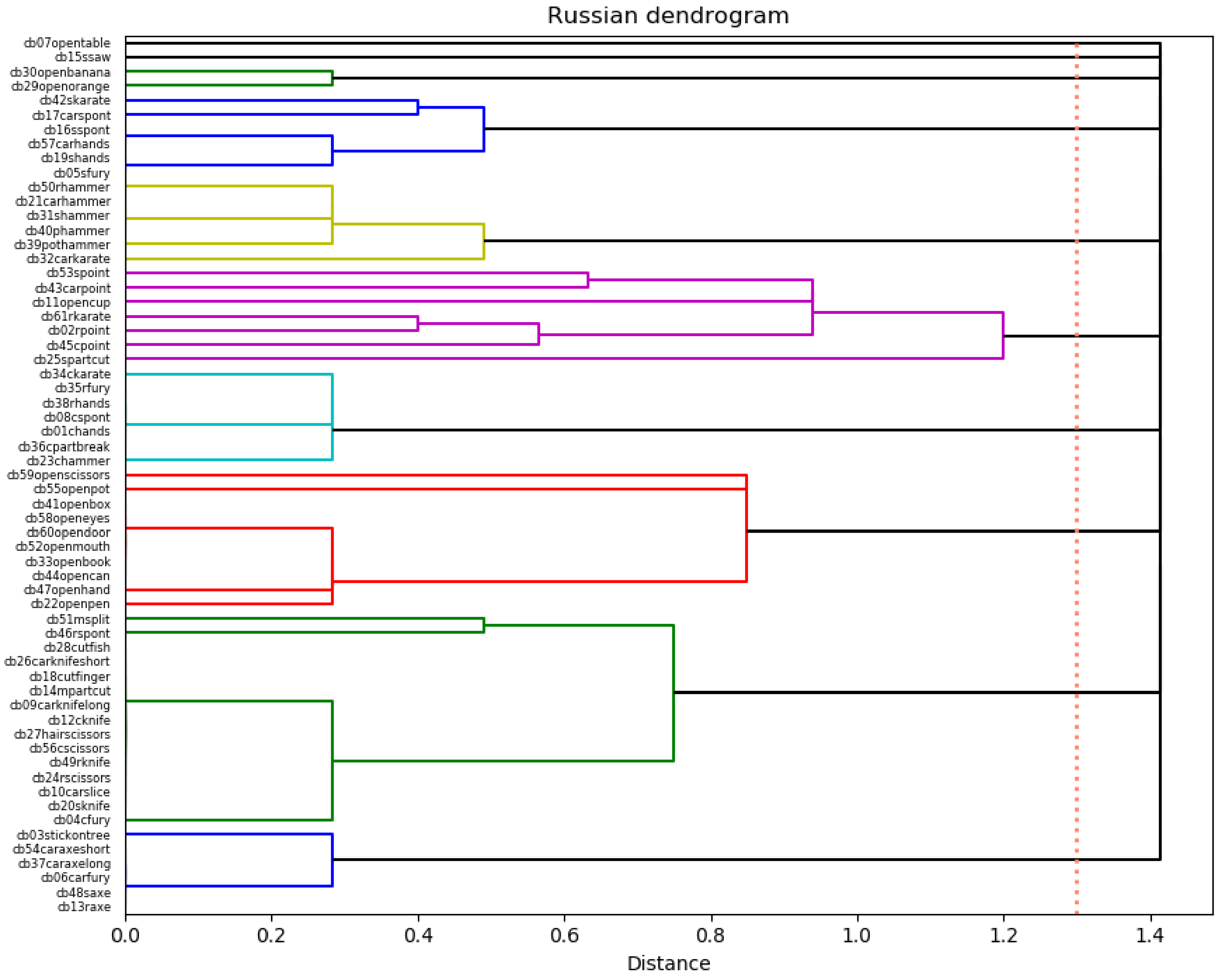

Figure 1c. *Russian dendrogram*

contains an event of tearing and all of the tearing events have been unambiguously assigned to one cluster (turquoise). Clip number 32 (cutting a carrot in half crosswise with single karate chop) constitutes the only exception of the yellow cluster. The level of predictability of the locus for the whole group is constant. The clip of cutting a carrot was attached to this cluster because of the non-deterministic nature of the algorithm. Although, the manner of the action is very similar, it does not match the cluster in other aspects according to the overall structure of the language. The next cluster, purple, is made of events with one shared property – the objects are being halved. Another subgroup of this cluster is group of events of poking – a sudden, sharp single blow from up to down towards a tightly stretched object. This cluster contains two exceptions from this description – 11 (pulling two paper cups apart by hand) and 25 (snapping twig with two hands partly). The reason for such an imprecision is hidden in the fact that these clips could not be assigned to any cluster with a better defined property and therefore they ended up in a cluster where they show the minimal similarity, namely, only single verb used for them was the one, that shaped this cluster. This observation is visible on the plot as the longest links from right to left without intermediate splits that represent these events. The following turquoise cluster is clearly made of tearing events. All of them were classified together on the very early phase of the algorithm. The red cluster consists of opening events, which is self-explanatory too. The green cluster contains cutting events. Most of the cutting events, where a knife or another sharp blade was used, have been grouped into this cluster. This cluster is the biggest one. The last cluster, blue (bottom), combines events of chopping, either with an axe or with a machete. Again, there is a strong correspondence to the instrument used in the cluster rather than to predictability of the locus.

4.1.3. Korean

Korean was the language with the least speakers – only one. Despite of that, the speaker used 30 different verbs to describe the video stimuli. That means that on average only two clips were described by the same verb. This fact forecasts that the dendrogram of Korean will have multiple outliers and every cluster will not contain many verbs. The upper part of the figure 1d, colored in yellow, represents clusters of maximal size of 2. Therefore, this part is not pertinent for discussion. As I already mentioned the events of peeling are described using specific for this action verb. The cluster of four clips at the very bottom of the dendrogram is made of events of halving objects into two pieces. The cluster above, that consists of 11 videoclips, contains mostly breaking events with low predictability of the locus. Nevertheless, clips from this cluster vary in such variables as agent, instrument, change in the state of the object, material of the object and number of blows. For instance, both clips depicting the slicing a watermelon have also been classified to this cluster. It is not obvious what generalization could be suitable for this cluster. The groups above are smaller and defined better. There is an emerging group of events of opening, events, where an axe was used and events, where no instrument was applied.

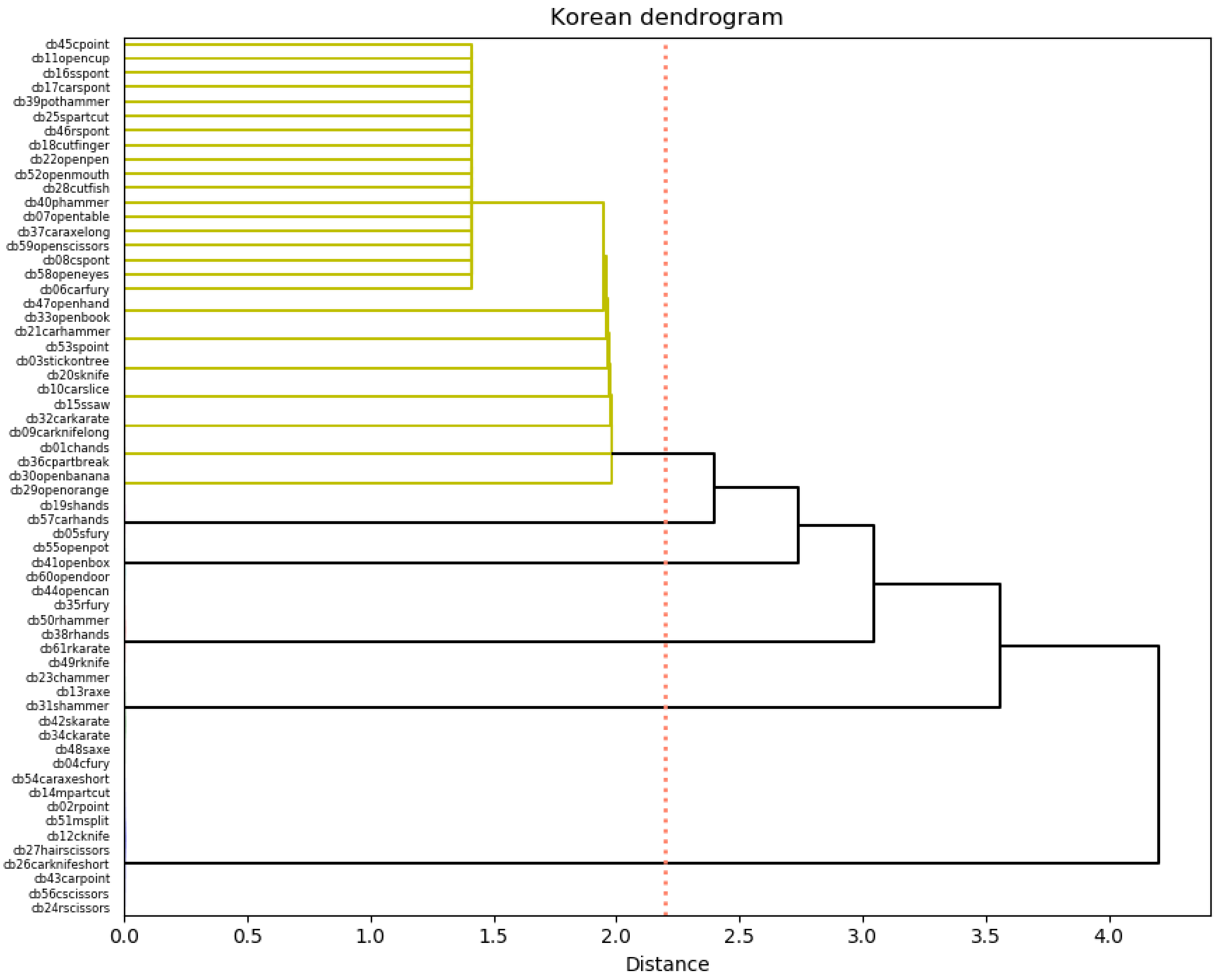

Figure 1d. *Korean dendrogram*

In contrary to other examined languages, Korean shows no special treatment of the events of tearing.

4.2. Clustering across languages

For revealing a general structure lying behind all the languages, the three clip-by-verb matrices were stacked along the x axis. This combined verbs of all the languages into one long array of features defining relations between the items – 61 videoclips. The same algorithm was run on this new matrix (see figure 2). The overall structure of the tree is different from those obtained by single languages. The clusters are more complex, and they are merged on the last steps of the algorithm. Most of the primary clusters have two major subclusters. Corresponding

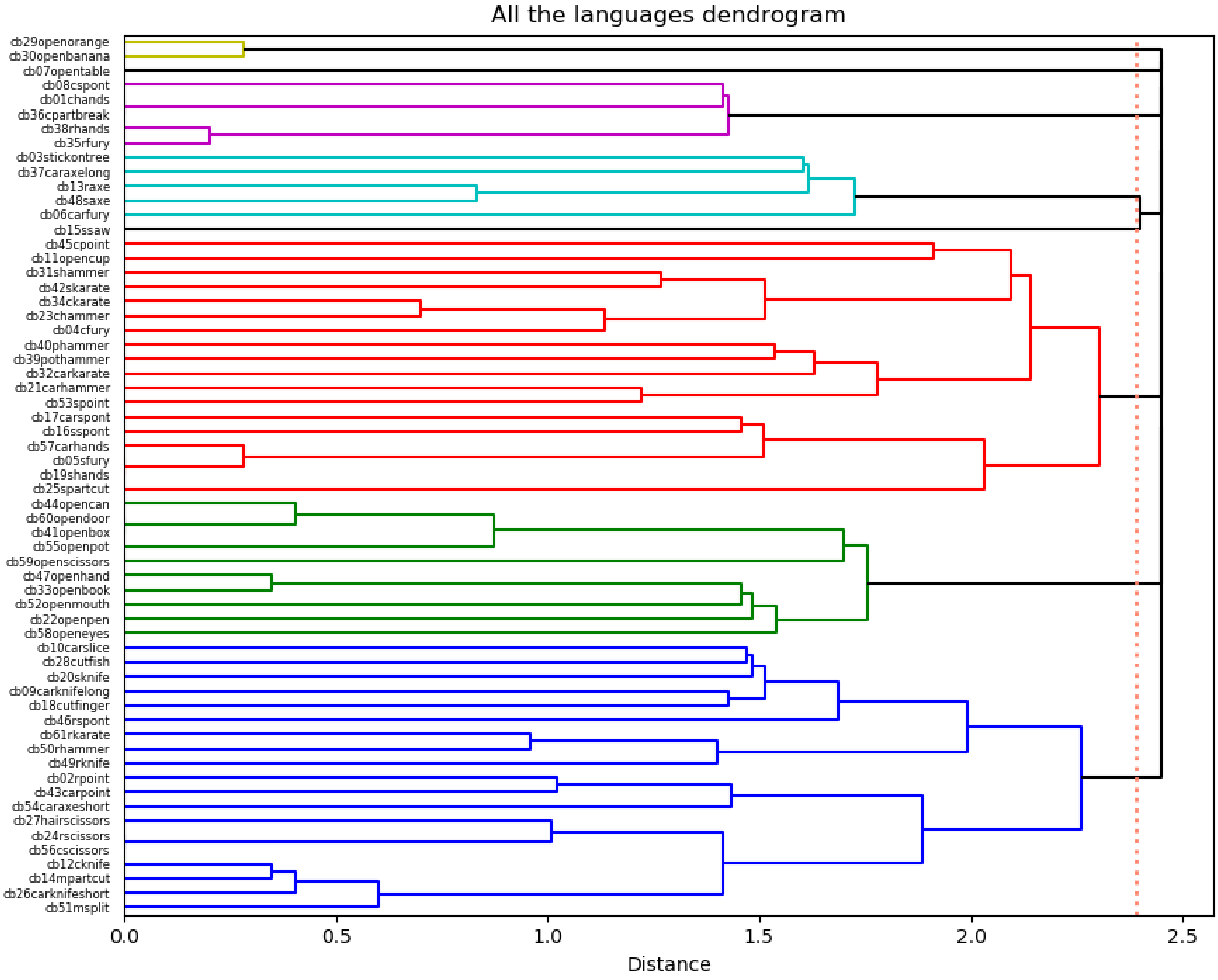

Figure 2. *Combined dendrogram*

to the outliers elicited by all the languages, the events of peeling (29,30), pushing a chair back from the table (7) and sawing a stick (15) each formed a branch by themselves; it is not grouped with the other clips until the very end of the procedure.

Now, that outliers are discussed and put aside, the clusters that emerged, based on all the languages, can be analyzed independently from top to bottom. The purple one contains only five verbs which represent all the tearing events of the dataset. These tearing events do not involve any instrument, the separation is made by hands. So, this cluster represents the dimension of tearing events, as it was proposed by Majid et al. (2008). The following branch, the turquoise one, combines the events of chopping. Consequently, the instruments presented in this cluster are axe and machete. The predictability of the locus of these events is very low.

Moreover, the actions performed in these clips require several blows to destruct the object. This dimension was not spotted in the research conducted by Majid et al. (2008). The next cluster (red) is the biggest one formed in this dendrogram. This group combines different “breaking” events. Two subgroups organized within this branch are snapping events (17,16,57,5,19,25) in the lower subbranch and smashing events in the upper one. The events of poking a hole in tautly stretched cloth (45) and pulling two cups apart (11) were classified to this cluster too. Moreover, they have merged together relatively late. In contrary to the results elicited by Majid et al. (2008) the clip depicting poling a hole (45) has not been classified separately in none of the languages researched in this paper. Thus, the fourth dimension suggested by Majid et al. (2008) is not demonstrated in Hebrew, Russian, or Korean. The green branch is self-explanatory – opening events. These events have been treated separately by every language. The last and the second biggest cluster is the blue one. In general, it is made of cutting events. Nonetheless, there are two major subclusters. The upper one (10,28,20,9,18,46,61,50,49) includes events with the most predictable locus of separation (10,28,20,9,18) and the rest represents a slight transition on the scale of predictability – less predictable events that include separating a rope, either with instrument or without. The lower subbranch of the blue cluster combines cutting that involves scissors (27,24,56), knife (12,26) or machete (14,51) – cutting with a very predictable locus of separation. And these clips together form the lower part of the cutting subbranch, whereas the upper one (2,43,54) demonstrates slightly lower extent of predictability but at the same time shares the manner of the action – sharp sudden move from up to down with single blow. Summing it up, the blue branch shows different degrees along the first dimension suggested by Majid et al. (2008).

5. Discussion

In the present study, I tackle the question of universality by systematically examining the lexical categories employed by speakers of Hebrew, Russian and Korean in describing a standardized set of events. If certain distinctions or groupings recur across a wide range of languages, it is plausible to assume that these reflect conceptualizations that are fundamental to human cognition (Majid et al. 2008). Hebrew and Russian demonstrated similarity in their underlying structure of the C&B domain with slight deviations in how they group events together. Korean will not be discussed from this perspective due to the several issues which evolved in the process. These issues will be addressed in the end of this chapter.

The structures of Hebrew and Russian were analyzed and examined from the perspective of the classification proposed by Majid et al. (2008). This classification has elicited four main dimensions, along which the events of "cutting and breaking" are divided typologically:

1) The predictability of the locus of separation in the affected object.
2) Separation of tearing events from all the rest
3) Differentiation between "snapping" and "smashing" events.
4) Separation of poking a hole in a piece of cloth stretched tautly between two tables from all the rest of the events.

First, I would like to mention that the third dimension is an extension of the first one. It represents a distinction that is already contained in a somewhat more abstract first dimension. The act of "snapping" is by its nature more predictable than "smashing", which represents an undefined and less oriented action. This dimension was treated differently by Hebrew and Russian as follows: Hebrew correlates snapping and chopping events to the events with higher predictability of the locus, namely, cutting events, whereas Russian forms multiple unrelated groups for snapping, chopping, cutting, and smashing events. This classification in Russian is rather based on the instrument used for separation but it is obviously true that there is a strong correlation between the tool and the manner of destruction, in a sense that one can predefine the other.

This leads to the assumption that languages differ not only in the number of clusters they recognize, but also in how they group events together. An additional example in this sense is cutting-with-scissors events. Hebrew treats them separately whereas in Russian they are not distinguished from other cutting events.

As Majid claims, predictability is a continuous dimension, with events involving separation with knives and scissors at one end, events of snapping and smashing at the other, and events in which a sharp blow causes the separation—chopping events—in the middle. In crosslinguistic perspective then, chopping events seem to be intermediate in the predictability of the location of separation. This is reflected in Hebrew as well, although Russian does not specify which end of the scale these events are closer to.

In contrary to the first and the third dimensions, the second dimension is reflected very clearly in all the languages examined in this paper. All of them group the events of tearing together without mixing them with another clusters on the very early stages of the algorithm.
The fourth dimension, namely, the event of poling a hole was never clustered separately, but rather merged to the cluster of cutting events for Russian and to the cluster of breaking events for Hebrew. One possible explanation is that this dimension is of a lower importance in these languages. However, I suggest a technical explanation to this phenomenon. Since the algorithm used for clustering is non-deterministic, this event might have been classified to one of the bigger clusters that have merged in the earlier phases of the run. The algorithm is greedy; therefore, it tries to minimize the maximal distance between each point of two given clusters. Consequently, depending on the order in which the algorithm passes over the clusters, this event can be merged with different cluster during different executions. For clearer analysis of the core C&B events, such outliers should be excluded. This was not done in my research because its primary goal was to capture the entire picture, identify such outliers and to try to explain why and in which environment they occur. Such approach allowed me to analyze outliers, events of opening, peeling, tearing and core C&B events altogether. The resulting dendrograms visualized all these dimensions separately without mixing them up significantly. Few exceptions that occurred in the process were spotted and discussed separately.

Now I will refer to another problematic aspect that has arisen in the process of analyzing Korean. The dendrogram of Korean was not interpretable enough to make broad implications. Moreover, the structure and clusters did not comply with none of the earlier proposed dimensions. This is plausible that Korean is a typologically unique language, or at least in the lexical domain of C&B verbs. However, this assumption has to be proved by a more thorough research.

The case of Korean has proven that one respondent is not enough for a comprehensive analysis. The resulting dendrogram has not captured any hierarchical relations. Therefore, the results of Korean cannot be considered for a proper elicitation, whether it complies with certain lexical dimensions or not. For example, to measure coherent predictability of the locus for the emerged clusters, there is not enough structure in this particular dendrogram. In order to resolve this issue, more data should be collected. The influence of the noise points will be minimized and then, more stable structure, analogously to Hebrew and Russian, will be achieved. Furthermore, the quantity of verbs should be normalized as more data is introduced. Hebrew, with the highest number of verbs – 54, has an average number of 9 verbs per respondent. In contrary, the number of verbs per respondent in Korean is 30 which cannot be transferred into any meaningful statistical implication.

To conclude, the semantic categories of typologically diverse languages can be very different from one another. Yet, at least in the domain of C&B events, this variation is often played out within a common structure. The attempt of defining this structure made in Majid et al. (2008) set a very abstract and typologically respected dimensions. Thus, the assumption that languages make a distinction between events depending on the predictability of the locus of the performed action is strengthened by this research. However, there are deviations from the structure defined by these dimensions as the examples of Hebrew and Russian show. These deviations can be researched further and gathered together in order to refine the common structure of the C&B events throughout the languages of the world.

References

Ameka, F. K., & Essegbey, J. (2007). Cut and break verbs in Ewe and the causative alternation construction. Cognitive Linguistics, 18(2), 241-250.

Bohnemeyer, J. (2007). Morpholexical transparency and the argument structure of verbs of cutting and breaking. *Cognitive Linguistics*, *18*(2), 153-177.

Bohnemeyer, J., Bowerman, M., & Brown, P. (2001). Cut and break clips, version 3. In S. C. Levinson & N. J. Enfield (Eds.), *Field manual* (pp. 90–96). Language & Cognition Group, Max Planck Institute for Psycholinguistics.

Brown, P. (2007). 'She had just cut/broken off her head': Cutting and breaking verbs in Tzeltal. *Cognitive Linguistics*, *18*(2), 319-330.

Chen, J. (2007). 'He cut-break the rope': Encoding and categorizing cutting and breaking events in Mandarin. *Cognitive Linguistics*, *18*(2), 273-285.

Essegbey, J. (2007). Cut and break verbs in Sranan. *Cognitive Linguistics*, *18*(2), 231-239.

Everitt, B.S., Landau, S., Leese, M. (2001). *Cluster Analysis* (Fourth ed.). London: Arnold.

Defays, D. (1977). "An efficient algorithm for a complete link method" (PDF). *The Computer Journal*. British Computer Society, 20 (4), 364–366.

Gaby, A. (2007). Describing cutting and breaking events in Kuuk Thaayorre. *Cognitive Linguistics, 18*(2), 263-272.

Guerssel, M., Hale, K., Laughren, M., Levin, B., & Eagle, J. W. (1985). A cross-linguistic study of transitivity alternations. *Cls*, *21*(2), 48-63.

Hale, Kenneth L. and Keyser, Samuel J. (1987). A View from the Middle. *Lexicon Project Working Papers 10.*, Cambridge, MA: Center for Cognitive Science, MIT Press.

Harris, C., Millman, S., Gommers, P., Cournapeau, E., Taylor, J., Berg, N., Kern, R., Picus, S., Kerkwijk, M., Haldane, J., Wiebe, P., Gérard-Marchant, K., Reddy, T., Weckesser, H., & Gohlke, T. (2020). Array programming with NumPy, *Nature, 585*, 357–362.

Hunter J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering, 9*(3), 90-95.

Koptjevskaja-Tamm, M. (2012). New directions in lexical typology.

Levin, B., Hovav, M. R., & Keyser, S. J. (1995). *Unaccusativity: At the syntax-lexical semantics interface* (Vol. 26). MIT press.

Levinson, S. C. (2007). Cut and break verbs in Yélî Dnye, the Papuan language of Rossel Island. *Cognitive Linguistics*, *18*(2), 207-218.

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*, *109*(2), 235-250.

Majid, A., Bowerman, M., Van Staden, M., & Boster, J. S. (2007). The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive linguistics*, *18*(2), 133-152.

Majid, A., Gullberg, M., Van Staden, M., & Bowerman, M. (2007). How similar are semantic categories in closely related languages? A comparison of cutting and breaking in four Germanic languages. *Cognitive linguistics*, *18*(2), 179-194.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 56 - 61).

Modern Hebrew verbs. (2020). Retrieved September 27, 2020, from https://en.wikipedia.org/wiki/Modern_Hebrew_verbs#Regular_conjugation.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Pye, C. (1993). Breaking concepts: Constraining predicate argument structure. In *Kansas Linguistics Workshop, Lawrence, Kansas, USA*.

Pye, C., Loeb, D. F., & Pao, Y. Y. (1995). The acquisition of breaking and cutting. In *The proceedings of the twenty-seventh annual child language research forum* (pp. 227-236). Center for the Study of Language and Information Stanford.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. American Anthropologist, 88, 313–338.

Van Rossum, G., & Drake, F. (2009). *Python 3 Reference Manual*. CreateSpace.

Virtanen, P., Gommers, R., Oliphant, T., Haberland, M., Reddy, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Walt, S., Brett, M., Wilson, K., Mayorov, N., Nelson, A., Jones, E., Kern, R., Larson, C., Polat, ., Feng, Y., Moore, E., Vand erPlas, J., Laxalde, J., Cimrman, R., Henriksen, E., Harris, C., Archibald, A., Ribeiro, A., Pedregosa, P., & Contributors, S. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*.

Appendix

Below is the list of video stimuli used in the research. The list with short descriptions was designed by Henkemeyer et al. (2001). *Italics* indicates “open”, “take apart” and “peel”. **Bold** indicates the “spontaneous” actions.

1. Tear cloth into two pieces by hand.

2. Cut rope stretched between two tables with single downward chisel blow.

3. Hack branch off tree with machete.

4. Chop cloth stretched between two tables with repeated intense knife blows.

5. Break stick over knee several times with intensity.

6. Chop multiple carrots crosswise intensely with big knife.

7. *Push chair back from table.*

8. **Piece of cloth tears spontaneously into two pieces.**

9. Slice carrot lengthwise into two pieces with knife.

10. Slice carrot crosswise into multiple pieces with knife.

11. *Pull two paper cups apart by hand.*

12. Cut strip of cloth stretched between two people’s hands in two with knife.

13. Cut rope stretched between two tables with axe blow.

14. Make single incision in melon with knife.

15. Saw stick propped between two tables in half.

16. **Forking branch of twig snaps spontaneously off.**

17. **Carrot snaps spontaneously.**

18. Cut finger accidentally while cutting orange.

19. Snap twig with two hands.

20. Cut single branch off twig with sawing motion of knife.

21. Smash carrot into several fragments with hammer blows.

22. **Take top off pen.**

23. Chop cloth stretched between two tables into two pieces with two hammer blows.

24. Cut rope in two with scissors.

25. Snap twig with two hands, but it does not come apart.

26. Cut carrot crosswise into two pieces with a couple of sawing motions of knife.

27. Cut hair with scissors.

28. Cut fish into three pieces with sawing motion of knife.

29. *Peel an orange almost completely by hand.*

30. *Peel a banana completely by hand.*

31. Smash a stick into several fragments with single hammer blow.

32. Cut carrot in half crosswise with single karate chop.

33. *Open a book.*

34. Chop cloth stretched between two tables with single karate chop.

35. Break yarn into many pieces with intensity.

36. Tear cloth about halfway through with two hands.

37. Cut carrot in half lengthwise with single axe blow.

38. Break single piece off a length of yarn by hand.

39. Smash flowerpot with single hammer blow.

40. Smash plate with single hammer blow.

41. *Open a hinged box.*

42. Break vertically held stick with single karate chop.

43. Cut carrot crosswise into two pieces with single chisel blow.

44. *Open canister by twisting top slightly and lifting it off.*

45. Poke hole in cloth stretched between two tables with a twig.

46. **Rope parts spontaneously, sound of a single chop.**

47. *Open hand.*

48. Chop branch repeatedly with axe, both lengthwise and crosswise, until a piece comes off.

49. Cut rope in two with knife.

50. Chop rope stretched between two tables in two with repeated hammer blows.

51. Split melon in two with single knife blow, followed by pushing halves apart by hand.

52. *Open mouth.*

53. Break stick in two with single downward chisel blow.

54. Cut carrot in half crosswise with single axe blow.

55. *Open teapot/take lid off teapot.*

56. Cut cloth stretched between two tables in two with scissors.

57. Snap carrot with two hands.

58. *Open eyes.*

59. *Open scissors.*

60. *Open door.*

61. Break rope stretched between two tables with single karate chop.