



Dissecting the aero-acoustic parameters of open articulatory transitions

Mark Gibson¹, Oihane Muxika¹, Marianne Pouplier²

¹ Universidad de Navarra, Spain

² LMU-Munich, Germany

mgibson@unav.es

Abstract

We capitalize on previously recorded kinematic and acoustic data for three languages (Georgian (GE), Spanish (SP) and Moroccan Arabic (MA)) that exhibit open articulatory transitions between the consonants in clusters in order to dissect the aero-acoustic parameters of the transitions in each language. These particular languages are of interest because they show similar patterns of interarticulatory timing in clusters, offering the unique opportunity to examine the acoustics of open transitions cross-linguistically. Our analysis centers on word initial clusters (/k/ and /g/), from which we extract relativized temporal values relevant to clusters and spectral parameters related to open articulatory transitions. We report baseline results using linear mixed effects models, then train a Random Forest model in a supervised learning environment on the significant variables. After training, test tokens are introduced in order to test whether the model can categorize the language based on the spectral and temporal parameters, and rank variables in terms of their feature importance. The results show that the model can categorize the data to the correct language with a 95,59% accuracy rate, where normalized zero-crossing (nzcr), modifications of the amplitude envelope (ΔE), and intensity ratio ranked highest in feature importance.

Index Terms: speech timing, consonant clusters, spectral and temporal parameters

1. Introduction

Previous studies have revealed two basic patterns of intergestural timing in word initial consonant (C1C2) clusters. On one hand, the consonant gestures in languages like German and English overlap such that the constriction of C2 begins before the release of constriction of C1 [1-3]. This type of transition from C1 to C2 has become known as a close transition. On the other hand, languages such as Georgian (GE), Spanish (SP) and Moroccan Arabic (MA) have been found to exhibit so called open transition [4] (for MA see [5] for SP see [6]), whereby the constriction of C2 begins well after the release of constriction of C1, producing what has become known as a non-zero interconsonantal (or inter-plateau) interval, as shown in Figure 1 (where the interconsonantal interval is simply labeled *Lag*).

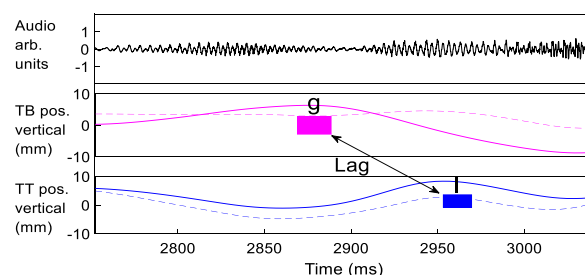


Figure 1. EMA recording of a representative production of /gl/ in SP. The time lag appears between the plateaus (maximal constriction phases) of the two consonants, shown by the separation between the shaded boxes in the lower two rows of each panel. The panels from top to bottom show the wave form, tongue body (TB) and tongue tip (TT) trajectories.

Due to the appearance of a vocoid-like segment between C1 and C2 in languages that exhibit open transitions, it is not altogether obvious whether the instantiation of phonation for C2 is phased in relation to any specific articulatory landmark of the C2 (see, for example, [7] on Catalan; [8] on Hungarian; [9-10] on Modern Greek; [11] on Polish; [12] on Romanian; [13] on Serbo-Croatian; [14] on Slovak; [6] for Spanish. For studies on the articulatory basis of such transitional vocoids, see [5] and [15] on Moroccan Arabic as well as [16]). Most studies investigating this vocoid-like segment consider its appearance to be a passive consequence of the aerodynamic conditions that develop from the open articulatory transition, and not an active objective of articulatory (or laryngeal) timing. This stance is buttressed by the fact that the vocoid generally lacks a durational and gestural target [17] and is usually considered optional and unpredictable (both in its frequency of appearance and the aero-acoustic/temporal parameters) (see previous citations above). Thus from this perspective, the aero-acoustic properties of the transition are not phonologically directed. Of course, if this is the case, then differences across languages that cannot be attributable to the specific articulatory parameters of the segments in the clusters should not abound, since there is no specific acoustic target (thus differences in the aero-acoustic parameters fallout from the language-specific phonetic implementation of surrounding gestures).

At the same time, prior studies have shown that the duration of the interconsonantal interval, and as a consequence the transitional vocoid that appears between the constriction plateaus of C1 and C2, is modulated by the durational properties of C2 [18-19]. In addition, anecdotal evidence for SP suggests that speakers, both consciously and subconsciously, can manipulate the articulatory timing of C1 and C2 in order to produce longer or shorter transitions of gradient levels of

audibility in order to achieve a metrical or prosodic target (see the following web version of an article that appears in a national newspaper in Spain, El País, that documents everyday cases of vocoid insertion between C1 and C2 in SP: https://verne.elpais.com/verne/2018/02/20/articulo/1519116469_169937.html). These studies support the idea that the timing of the consonants in clusters is phonologically directed, though, to date, no cross-linguistic study has examined the spectral parameters of open transitions for evidence of phonologically driven aero-acoustic targets. If it is the case that some acoustic parameters of the transitions are phonologically encoded, then evidence for differences in the acoustic parameters of the transition across languages, independently of the identity of the consonants in C1C2 clusters, should be robust (because they are language-specific).

In the following sections, we outline a current pilot study in which we dissect the acoustic parameters of open transitions in three languages in order to tease apart the phonological and phonetic constraints that modulate the oro-laryngeal timing in clusters.

2. Speech materials

Speech materials for Georgian, Spanish and Moroccan Arabic were collected previously at the Universität Potsdam and LMU-Munich. Both kinematic and acoustic files were originally recorded but only the acoustic files were used for the present study. For detailed descriptions of the speakers, dialects, corpora and data collection procedures please see [20] for Georgian, [6] for Spanish, and [5] for Moroccan Arabic (as properly explaining each study's procedures would be impossible here due to space restrictions).

2.1. Acoustic labeling and processing

The acoustic files were labeled by hand using Praat. Five interval tiers (cluster+vowel, cluster, segment, VOT, transition) were used to extract temporal aspects of the /kl-gl/ clusters. Raw temporal values were relativized to the whole (1) cluster+vowel and (2) cluster, rendering a percentage for that variable in relation to the entire cluster+vowel/cluster. The transition interval was extracted and uploaded into Matlab for phonation profile analysis. To register modulation of the amplitude envelope, we used a measure (see Figure 3), ΔE , based on the difference between two envelope time points: the envelope amplitude at the beginning of the transition and the maximum of the envelope (see [20] for a full review).

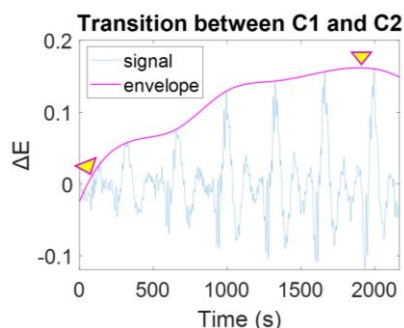


Figure 3. Oscillogram with amplitude envelope of transition intervals for stop-lateral productions by a speaker of SP. The triangles in the graph indicate E1, E2 which were used to calculate ΔE .

An intensity ratio to capture differences due to C1 manner was collected by dividing the minimum intensity of the cluster+vowel (which was always during C1) by the maximum intensity (which was always at the vowel). Thus, the lower the number the more stop like the C1 is, while higher values signal a more approximant-like production (as can be the case in Spanish). Finally, we obtained a normalized zero crossing rate for a given transition as an index of degree of voicing by obtaining the zero crossing rate and normalizing for the number of samples in a given transition. Higher zero crossing rates imply less voicing.

2.2. Statistical analysis

2.2.1. Linear mixed effects analyses

We used linear mixed effects models to investigate whether language (henceforth, Language, when referred to as a specific factor) and C1 voice (henceforth, C1 Voice) had an effect on the temporal and spectral parameters of the clusters. Both random intercepts by subject as well as random by-subject slopes were included. Maximum Likelihood Chi-squared tests based on the deviance statistics were performed in order to determine significance. For post-hoc comparisons, significance was determined using the Tukey adjusted contrast using the multcomp package in R. Variables and factors are listed in Table 1, along with a description of the measurement and levels.

Table 1. Variable and factor names, descriptions and levels

Variable name	Description
Relativized durational variables	
% C1/C2 of total cluster duration	C1, C2 duration÷duration cluster
% VOT of total cluster duration	VOT÷duration cluster
% transition of total cluster duration	Transition÷duration cluster
Spectral parameters	
ΔE	Change in amplitude from start of transition to max. amp.
Intensity ratio	Min inten.÷max. inten.
Nzcr	Normalized zero crossing rate
Factors	
Language	GE, SP, MA
C1 Voice	±voice

2.2.2. RF models in a supervised learning environment

Random Forest classifiers are a machine-learning model that takes the consensus of a number of decision trees to determine the probability of a single instance of data belonging to a particular class. Each decision tree uses a set of Boolean conditions on features (such as $VOT \leq 45$) to classify data into one of a number (in this case 3) classes (GE, SP, MA). They are a powerful machine-learning model that may be trained on relatively few data points when compared with other models such as Deep Neural Networks. Random Forests, like other machine learning models trained to perform data classification, partition the feature space into a number of categories. However, many other classification models (including a standard statistical analysis that relies upon the data being normally distributed within each class) do not have the

flexibility of the Random Forest classifiers when partitioning the feature space, as they are often limited to partitioning via hyperplanes. Finally, since, there are a set number of decision trees in the Random Forest, and each has a set number of Boolean conditions, one may simply take the percentage of Boolean conditions related to each feature as that feature's importance in classification.

For our models, we used R's randomForest package [21]. We trained the model on 70% of the data, and subsequently introduced test data for classification. Each of our Random Forest classifiers consisted of 500 decision trees, each of which had a maximum depth of 3 (only three Boolean conditions were allowed to classify any piece of data).

3. Results

3.1 Linear Mixed Effects Models

For the normalized zero crossing (nzcr), there is a main effect of Language ($\chi^2[1, N=374] = 18.91, p<0.001$), as well as for the C1 Voice ($\chi^2[1, N=374] = 3.95, p=0.046$) and their interaction ($\chi^2[2, N=374] = 25.04, p<0.001$). MA has the highest mean nzcr ($M=0.2201$), followed by GE ($M=0.1777$) and SP with the lowest mean nzcr ($M=0.0652$). This means that Spanish shows a relatively stronger degree of voicing of the transition compared to Georgian and MA.

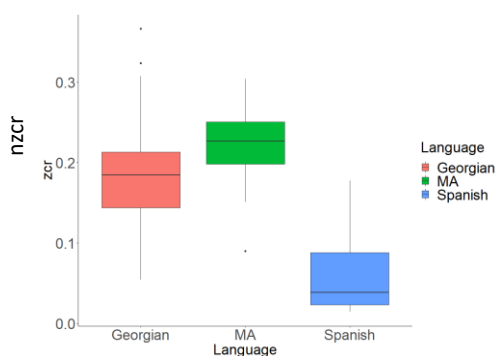


Figure 4. Boxplots of nzcr (y-axis) across languages (x-axis).

With regard to ΔE , there is a slight, though significant, main effect of Language ($\chi^2[1, N=374] = 18.91, p=0.043$) on amplitude modulation, but no effects of C1 Voice were found ($\chi^2[1, N=374] = 1.99, p=0.15$) nor an interaction between C1 Voice and Language ($\chi^2[1, N=374] = 9.64, p=0.09$). Post-hoc testing by language show significant differences in ΔE for MA with SP and GE, but not between SP and GE, where means were quite similar (SP: $M=0.0781$; GE: $M=0.0763$).

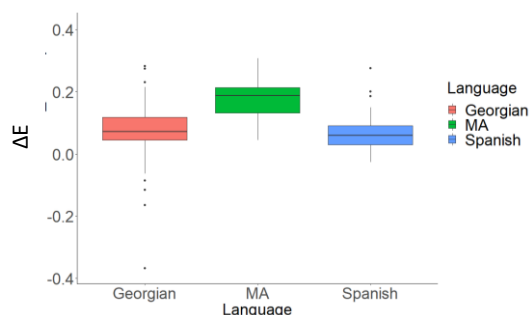


Figure 5. Boxplots of ΔE (y-axis) across languages (x-axis).

Concerning the intensity ratio, a strong main effect of Language ($\chi^2[1, N=374] = 31.05, p<0.001$) was found, as was a strong main effect of C1 Voice ($\chi^2[1, N=374] = 18.74, p=0.001$) and their interaction ($\chi^2[2, N=374] = 49.52, p=0.001$). For pooled data, GE had the lowest mean intensity ratio ($M=0.43$), SP had the highest ($M=0.72$), and MA exhibited a mean ratio ($M=0.57$) between GE and SP. This result was expected since part of the voiced stops in SP were produced as approximants, as per the rule of spirantization that was variably applied by our subjects due to the experimental protocol we employed (we explain this in full in section 4.).

Finally, with regard to the relativized durational variables, an effect of language was found for % of C1 duration ($\chi^2[1, N=374] = 19.37, p<0.001$), % of C2 duration of total cluster ($\chi^2[1, N=374] = 3.54, p=0.04$), % of VOT of total cluster ($\chi^2[1, N=374] = 11.17, p=0.003$), and % of transition of total cluster ($\chi^2[1, N=374] = 6.14, p=0.01$), with strong interactions with C1 voicing for % of C1 of total cluster ($\chi^2[1, N=374] = 24.82, p<0.001$), % VOT of total cluster ($\chi^2[1, N=374] = 10.71, p=0.003$), and % transition of total cluster ($\chi^2[1, N=374] = 5.87, p=0.02$).

Although main effects (and interactions) were found for these relativized durational variables, patterns were not consistent across languages, though we only report the results in summary due to space limitations. SP exhibits transitions that account for a significantly ($p<0.05$ for pairwise contrasts with GE and MA) lower percentage of the cluster than do GE and MA, while GE and MA show lower overall proportional duration of C1 and C2. With regard to the proportional transition duration, GE and MA showed commensurate percentages (31% and 29% for total cluster), which a post-hoc Tukey HSD revealed was not significant ($p=0.85$). For SP, though, the transition only accounted for 22% of the total cluster duration, which post-hoc pair-wise analyses did prove significantly different from GE ($p<0.001$) and MA ($p<0.001$). This means that transitions in SP occupy a lower percentage of the total cluster as compared to transitions in GE and MA.

An opposite effect, however, was found for the way the three languages patterned with regard to relative C1, C2 duration. For GE and MA, both C1 and C2 occupy lower percentages of the total cluster than SP (GE: C1=24%; C2=34%; MA: C1=28%; C2=33%; SP: C1=33%; C2=44%). Again, post-hoc analyses reveal like patterning for GE and MA (i.e., no group differences, Tukey HSD: $p=0.78$ for C1; Tukey HSD: $p=0.83$ for C2), though both languages showed significant differences with the SP patterns (Tukey HSD: $p<0.01$ for all contrasts). These differences in patterning are of interest given recent results for SP syllable timing in the literature where a compensatory effect was found between C2 laterals and the interconsonantal interval, such that the duration of the latency between C1 and C2 is modulated by the duration of C2 [19].

3.2 Random Forest analysis

We fitted the Random Forest models with the variables that showed effects of Language and interactions between Language and C1 Voice. The variables used to train the models were: nzcr, ΔE , intensity ratio, % of C1 of total cluster, % of C2 of total cluster, % of VOT of total cluster, and % of transition of total cluster. The models were trained on 70% of the data set before test data were fed in, whereby the model was directed to categorize the test tokens by Language. We built three models in order to observe the importance of the temporal and spectral variables in categorization. In one model, we combined the

temporal and spectral parameters, and then built separate models for the temporal and spectral parameters. Results of the models were as follows.

Accuracy rate for categorization was highest for the model that only contained the spectral parameters of the transition (95.59% accuracy). For the mixed model (where the relativized temporal and spectral parameters were programmed), the accuracy rate fell to 86.96%, while accuracy was lowest for the model specified only for the temporal parameters (79.17%).

For any given decision tree, the most important variables are ranked by the amount they reduce the error when they appear, the error reduction being weighted by the number of observations on the node. In a Random Forest, this is performed for every tree in the forest, and then averaged to obtain the importance of a given feature. For our maximally accurate model (using only spectral parameters), *nzcr* had the highest feature importance value, followed by the intensity ratio, then ΔE . These results were expected given the results of the linear mixed effects models that showed no differences between SP and GE for ΔE , and main effects for *nzcr* (as well as significant differences for all pairwise post-hoc analyses).

For the spectral-parameters-only model, feature importance was highest for *nzcr*, followed by intensity ratio with ΔE showing the lowest feature importance (as shown in Table 2). These results were also reflected in the mixed model, where spectral and temporal parameters were fitted.

Table 2. Feature importance (max. 1) by language

	GE	MA	SP	MDA*	MDG [#]
<i>nzcr</i>	0.22	0.21	0.32	0.27	28.29
Intensity ratio	0.13	0.42	0.12	0.16	22.20
ΔE	0.04	0.20	0.02	0.05	13.15

* Mean Decrease Accuracy

[#] Mean Decrease Gini¹

In sum, the results of our models show that the spectral parameters of the transitions are more reliable indicators of language categories than the relativized temporal parameters, though the temporal parameters also provide robust cues by which to distinguish languages. The percentage of the total cluster/cluster+V that the transition occupies in SP, for example, had, after *nzcr*, the highest feature importance of all the variables (0.11, *nzcr* = 0.19). Recall, that SP has shorter overall transitions, but longer C1 and C2 than GE and MA. These temporal differences in C1 and C2 across the languages may play a role in molding the precise acoustic parameters of the transitions by constraining the aerodynamics needed for voicing and the intraoral pressure that affects VOT.

4. Discussion and Conclusions

We dissected the acoustic parameters (both temporal and spectral) of open articulatory transitions in three languages. The

¹ Mean Decrease in Gini expresses a variable's total decrease in node impurity as a mean, which is weighted according to the proportion of samples that reach that node in successive individual decision trees in the forest. Higher variable importance directly relates to a higher Mean Decrease in Gini.

results of the linear mixed effects models were by and large consistent with the results of the Random Forest analyses. First, we found significant differences in the noisiness of the transition, represented by *nzcr*, across the languages whereby GE and MA exhibited nearly double the *nzcr* values as SP, with significant differences as well between GE and MA. This means that SP has the strongest degree of voicing among our languages. However strong this language effect is, nonetheless, it did not seem to have significant consequences on the degree of amplitude modulation.

Independently, however, high variation in *nzcr* across languages and relatively stable amplitude modulation is not strong evidence in support of an acoustic target in open transitions. Nevertheless, following up this point, we surmised that if within the same language, we could find substantial modulations in clusters where the C1 manner was perturbed, then this may also indicate the presence of an acoustic target in the transition. Thus, we examined a subset of our Spanish data where C1 voicing stayed constant, but C1 manner fluctuated on a continuum between stops and approximants. In Spanish, the rule is that voiced stops following vowels (as in our carrier phrases) are produced as spirants of approximants, while stops following pauses are produced as stops. Due to the formal context of the experiments, some speakers paused before repeating the target token, producing voiced stops, while others read the sentence as a continuous phrase, and thus produced an approximant, as per the rule for intervocalic voiced stops in Spanish. As there is no complete occlusion for approximants, intraoral pressure is lower than for voiced stops, which may have an effect on amplitude modulation. Thus, we performed a correlation analysis using Pearson covariance for the amplitude modulation and the intensity ratio (which is in effect a reflection of manner) and found no correlation ($r(88) = 0.12$, $p = 0.39$). We found that the intensity ratio varied independently of amplitude modulation, providing some preliminary evidence for the modulation of the transition not being merely a function of voicing variation during C1.

Finally, with regard to ΔE , SP and GE showed commensurate means (though significantly different *nzcr*) while MA exhibited a different pattern with ΔE being significantly higher than GE and SP. This is noteworthy given the differences in how these particular clusters pattern in syllables across the different languages. In SP and GE, the stops and lateral belong to the same syllable onset, while MA reportedly does not permit consonants to cluster in syllable onsets [5]. Thus, ΔE may be an acoustic correlate of syllable structure differences, though to confirm this, future studies addressing ΔE for /kl,gl/ in different prosodic contexts are necessary. Additionally, in future studies we will consider the timing of the changes in modulation by obtaining a measure of the steepness of the slope from E1 to E2, which may be acoustically relevant.

5. References

- [1] D. Byrd, "Influences on articulatory timing in consonant sequences," *Journal of Phonetics*, vol. 24, no. 2, pp. 209–244, 1996.
- [2] L. Bombien and P. Hoole, "Articulatory overlap as a function of voicing in French and German consonant clusters", *The Journal of the Acoustical Society of America*, vol 134, no.1, pp 539–550, 2013.
- [3] S. Marin and M. Pouplier, "Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model," *Motor Control*, vol 14, no. 3, 380–407, 2010..

- [4] J.C. Catford, "English Phonology and the Teaching of Pronunciation," *College English*, vol. 27, no. 8, pp. 605–613, 1966.
- [5] A. Gafos, "A grammar of gestural coordination," *Natural Language and Linguistic Theory*, vol. 20, no. 2, pp. 269–337, 2002.
- [6] M. Gibson, S. Sotiropoulou, S. Tobin, and A. Gafos, "Temporal Aspects of Word Initial Single Consonants and Consonants in Clusters in Spanish," *Phonetica*, vol. 76, no. 6, pp. 448–478, 2019.
- [7] D. Recasens and A. Espinosa, "An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2288–2298, 2009.
- [8] R. Vago and M. Gósy, "Schwa vocalization in the realization of /r/," *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 505–509, 2007.
- [9] M. Baltazani and K. Nicolaidis, "Production of the Greek rhotic in initial and intervocalic position: An acoustic and electropalatographic study," in: Z. Gavriilidou, A. Efthymiou, E. Thomadaki, & P. Kambakis-Vougiouklis (Eds.), *Selected papers of the 10th International Conference on Greek Linguistics*, Komotini, Greece: Democritus University of Thrace, pp. 141–152, 2011.
- [10] M. Baltazani and K. Nicolaidis, "The many faces of /r/," in L. Spreafico & A. Vietti (Eds.), *Rhotics: New data and perspectives* Bozen: University of Bozen-Bolzano, pp. 125–144, 2013.
- [11] L. Stolarski, "Vocalic elements in the articulation of the Polish and English /r/," Paper presented at Languages in Contact, University of Wrocław, Poland, 2011.
- [12] A. Avram, "Cercetări experimentale asupra consoanelor lichide din limba română [Experimental re-search on liquid consonants in Romanian]," *Fonetică și dialectologie*, vol. 12, pp. 8–20, 1993.
- [13] S. Gudurić and D. Petrović, "О природи гласа /r/ у српском језику [The nature of the sound [r] in the Serbian language]," *Зборник Матице српске за филологију и лингвистику*, vol. 48, no. 1–2, pp. 135–150, 2005.
- [14] R. Pavlík, "K niektorým otázkam kvalitatívnych a kvantitatívnych vlastností slovenských vibránt [To the question of qualitative and quantitative characteristics of the Slovak vibrant consonants]," *Jazykovedný časopis*, vol. 52, no. 1–2, pp. 65–97, 2008.
- [15] A. Gafos, P. Hoole, K. Roon and C. Zeroual, "Variation in timing and phonological grammar in Moroccan Arabic clusters," in: C. Fougeron (Ed.), *Laboratory phonology 10: Variation, detail and representation*. Berlin, Germany: Mouton de Gruyter.
- [16] N. Hall, "Cross-linguistic patterns of vowel intrusion," *Phonology*, vol. 23, no. 3, pp. 387–429, 2006.
- [17] J.A. Bellik, "Vowel Intrusion in Turkish Onset Clusters", PhD dissertation, University of California Santa Cruz, 2019.
- [18] B. Blecua, "Compensación temporal en los elementos del ataque silábico," *Proceedings from the II Congress of Experimental Phonetics*, March 5–7, Sevilla, Spain, pp. 101–108, 2000.
- [19] S. Sotiropoulou, M. Gibson, and A. Gafos, "Global organization in Spanish onsets," *Journal of Phonetics*, vol. 82, pp. 1–22, 2020.
- [20] M. Pouplier, T. Lentz, I. Chitoran and P. Hoole, "The imitation of coarticulatory timing patterns in consonant clusters for phonotactically familiar and unfamiliar sequences," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 11, no. 1, pp. 1–41, 2020.
- [21] A. Liaw and M. Wiener, "Classification and Regression by Random Forest," *R News*, 2, pp. 18–22.