INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Using Games to Augment Corpora for Language Recognition and Confusability

Christopher Cieri, James Fiumara, Jonathan Wright

University of Pennsylvania, Linguistic Data Consortium, USA
{ccieri, jfiumara, jdwright}@ldc.upenn.edu

Abstract

We present a Game with a Purpose to elicit judgements of the language spoken in short audio clips of broadcast and conversational telephone speech, the resulting corpus and their potential use in research on language recognition and confusability.

Index Terms: speech corpora, games with a purpose, language recognition, language confusability

1. Introduction

Language Recognition is a critical early speech processing step in real world, multilingual settings [1] [2] that has received steady, if not intensive, attention for the past two decades. An initial foray in 1996 lead to NIST's Language Recognition Evaluations (LRE) continuing biennially, with occasional skips, since 2003 [3]. This contrasts with other evaluation campaigns recurred more frequently over shorter durations [4]. Despite fewer papers in recent meetings compared to topics such as Paralinguistic Analysis [5], Language Recognition has been a technical area since the first Interspeech [6] and predecessor conferences such as Eurospeech 1995 [7] and continues to attract research.

Corpora for Language Recognition (also LRE) vary along several dimensions in addition to size. Source data may be telephone conversation, news broadcast, prompted speech or less commonly, and perhaps less realistically, read speech. Broadcast is generally 'found' data while the other source types are often created specifically for LRE purposes. Most importantly, the degree to which speech is controlled or audited, and thus known to be in the target language, also varies. One of the earliest and most tightly controlled examples, the CSLU 22 Language corpus [8] recorded 2067 speakers responding to prompts via the telephone. Participants were asked to speak in the target language only. All responses were verified and more than 1/3 were transcribed. The Multilingual Corpus for Language Identification [9] contains speech from 300 speakers for each of four languages responding to written prompts intended to elicit short answers, reading and spontaneous speech. Recordings were verified for language and native speaker competence [10]. The fifteen CALLFRIEND corpora, published in 1996, e.g. [11], collected ~60 5-30 minute telephone conversations, audited for channel quality, number of speakers and distributed them with metadata on speaker sex, age and education. However the audit was performed at the file level and language issues were reported with general observations such as: "*there are a few sentences in English, but overall, is in the Hindi*" [12]. In contrast, preparation for the NIST LRE 2011 evaluation lead to the development of the Multi-Language Conversational Telephone Speech 2011 series with very careful auditing for the language of individual segments. Auditors were carefully screened and tested for inter-auditor agreement, and confusability of mutually intelligible languages was measured. Even so, the campaign's focus on avoiding repeat speakers in telephone conversations and broadcast narrow band speech lead to: "*very large volumes of unaudited audio*" [13]. The very recent VoxLingua107 [14] contains 6628 hours of speech extracted from YouTube videos for 107 languages validated automatically with parts human annotated via crowd-sourcing.

Notwithstanding the large number of LRE corpora available, instances of researchers employing data designed for speech-to-text applications [15] to develop language recognition technologies [16] as well as corpora created originally for LR being enhanced to support speech-to-text [17] confirm that the supply of available corpora has not kept pace with demand. In addition, a pattern that has emerged is that corpora focusing on read and prompted speech tend to have higher percentages of their content verified for language while those focusing on telephone conversation and broadcast tend to have greater overall volumes with only a subset verified for language. CommonVoice [18], reporting nearly 13k hours of read speech in 74 languages for which 10K hours are validated, is a good example of the former [19]. For applications where the latter sources are more appropriate, there could be is a rich supply of additional data if the languages of the unaudited portions could be verified.

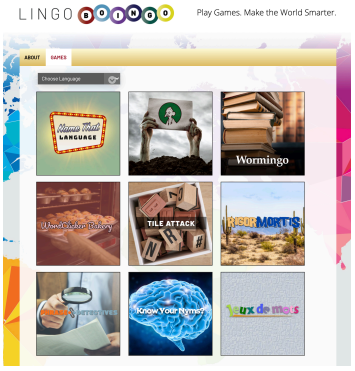

Figure 1: *LingoBoingo Game Grid*

2. NIEUW

The Novel Incentives and Workflows (NIEUW) project at the University of Pennsylvania Linguistic Data Consortium and Department of Computer and Information Science, sponsored by the US National Science Foundation's Community Research Infrastructure program (see §9), addresses the dearth of Language Resources by building infrastructure to create and share resources among multiple research communities. NIEUW offers novel (i.e. non-monetary) incentives to attract

http://dx.doi.org/10.21437/Interspeech.2021-1611

new workforces, providing custom workflows to supplement existing data collection and annotation approaches. NIEUW has created a toolkit with the aspirational name, Universal Annotator (UA), based upon LDC's WebAnn framework, which has been used to collect more than 1 million data samples and annotations across more than 100 different projects. UA generalizes many WebAnn features and runs in more environments: traditional web and database servers, dockerized configurations on VMs, laptops that can be carried into the field. NIEUW uses UA to create multiple portals each of which presents a logical cluster of incentives with an appropriate workflow to attract workforces to the challenge of filling resource gaps, including those in Language Recognition.

3. Outreach via LingoBoingo

The most immediately relevant of the NIEUW portals is LingoBoingo, hosting links to language games to increase discoverability, pool recruiting resources and measure the impact of outreach efforts. LingoBoingo is implemented as a WordPress site with: home page, About page that informs consents and lists partners and a Games Grid with titles, images, descriptions and links to games hosted on other sites.

LingoBoingo outreach began with posts to the LDC Newsletter [20] and social media, then expanded via LinguistList posts to their listserv and social media [21]. Subsequent advertisements broadened outreach to 'external' audiences via the SciStarter [22] portal promoting Citizen Science opportunities, as well as social media such as Twitter and Facebook. LingoBoingo was named a SciStarter Top Project for 2018 [23] which lead to mention in the online edition of Discovery Magazine [24]. Madge et al. [25] reports on the relative yield of some of these outreach efforts using metrics common to the free-to-play game markets and highlights the necessity of measuring impact not only in terms of the number of players or HITs but also in the number of data items fully annotated. We follow this advice in §7

4. Name That Language!

NameThatLanguage (NTL) was inspired by the Great Language Game (GLG) [26] which published ~16 million judgements [27] and may have collected many more over its five year life span. GLG presented short audio clips in multiple languages and asked players to guess the languages spoken from a small list of possibilities of which one was always the correct answer. Players gained points for correct answers and lost one of their three 'lives' with an incorrect answer. As players progressed through a game the number of choices for each audio clip increased. Over time, the number of languages grew and the developer encouraged players to add audio in new languages via the Wide Language Index project on GitHub [28]. The game site also added and later removed performance statistics and a map of player locations. The author blogged about the game and players could discuss game play and scores on a dedicated forum. GLG's purpose was to elicit data on the confusability of languages [29]. Presumably for this reason, the number of audio clips per language was relatively small and the language of each was already known. Two games inspired by GLG, LanguageSquad (LS) [30] and LingYourLanguage (LYL) [31] appear to follow the original model relatively closely. Both seem to give feedback for all plays suggesting the language of all clips is known or assumed to be so. LYL mentions having more than 2500 clips for almost 100 languages, (~25 clips/language on average). Both games have added multiple 'difficulty levels'. LS has also added an alphabet detection option from written text.

NTL has different goals that impact the game play, aggregation and possibly also popularity. NTL seeks to build corpora also for Language Recognition. To support this goal, data on confusability is useful but reliable ground truth language labels for many new audio clips collected under varying circumstances was critical to support robustness. NTL was seeded with two kinds of audio: *known* clips drawn from published corpora subjected to expert language annotation and *suspected* clips drawn from broadcasts or conversations purported to be in the target language but not verified.

NTL players listen to ~10 second audio clips and indicate which language they believe is spoken. Initially, NTL presents a score board, audio controls, decision buttons and a New Game button. Audio controls include a timer, a play/pause button and a stop button that resets the timer to the beginning of the clip. Once the audio clip has been played to the end, decision buttons are activated. The known or suspected language of the clip is always included among the buttons as distractors that increase in number as play continues. To progress the player must make a choice to which the game responds with a judgement, the language chosen and the actual language and a Next button. Currently, players receive 10 points for each correct answer and lose one of three 'lives' for each incorrect answer. The goal is to maximize points earned before losing all three lives. At any time, the player may click the New Game button to restart with a score of 0 and 3 lives.

The number of distractors increases by 1 with every 3 correct answers given until it reaches the maximum of 6. After every 6th normal play, the game enters a Bonus Round of three clips where the there is no feedback or penalty for an incorrect answer but the player receives 20 points per answer. Because the language of the bonus clips is not known, the game also provides an Other button, placed in a different, random position each time.

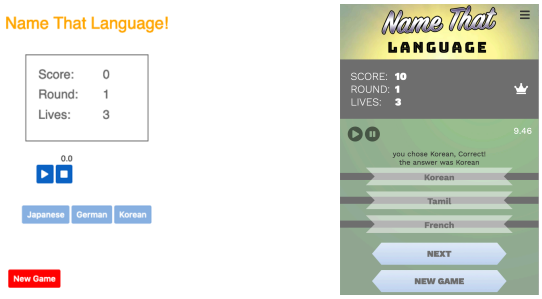

Figure 2: *NTL Game Screen: Original & Responsive Design*

Game parameters can be adjusted such as the audio clips and languages available, the number of language choices and the rate at which difficulty increases, the scoring system and the relationship between the player's native language or location when known, and the frequency of audio clips in related languages. Any such changes are noted in the data the game reports to researchers. For the data described here, the factors were held constant to provide a consistent corpus.

5. Input Data

The initial NTL data included a small number of known and a larger number of suspected clips. For any language in the known or suspected set, we wanted the same number of clips so that the priori probability hearing any language was equal.

The *known* clips are all conversational telephone speech (CTS) drawn from the *2003 NIST Language Recognition Evaluation* [32]. We had planned to include 80 clips for each language in the original corpus; however, due to clerical error or misunderstanding, there are more known clips in Japanese (160) and English (240)

Table 1: *Known and Suspected Clips by Language*

Language	Known	Suspected
Arabic	80	
English	240	600
French	80	600
German	80	
Hindi	80	600
Japanese	160	
Korean	80	600
Mandarin	80	600
Persian	80	600
Russian	80	600
Spanish	80	600
Tamil	80	
Vietnamese	80	600

The *suspected* clips are all extracted from a mixture of CTS and Voice of America (VOA) broadcasts with a preference for the former. We extracted clips from VOA using the same method that was used to create LDC's *broadcast narrow-band speech* (BNBS) corpora. Speech activity and bandwidth detectors identify speech within narrow band audio expecting the intersection to be embedded telephone conversations and thus more comparable to CTS. A small number of narrow band segments of the desired length are then extracted, typically one each from the beginning, middle and end of the broadcast, to reduce the probability of repeat speakers. 600 suspected clips per language were selected, though the proportion of CTS versus VOA differed across languages due to differences in supply.

Although the BNBS clips originate from Voice of America broadcasts purported to be in the target language, we are less than 100% confident in the language of these clips for multiple reasons. First, like any broadcast source, VOA is subject to pre-emptions and schedule changes that might have escaped notice during collection. In addition code-switching is common in many VOA language services. Finally, when a news program includes an audible quote originally in a different language, the practice seems to be to begin playing that quote in the original language before fading out the original voice and adding a voice-over in the language of the broadcast. This leads to audio clips with multiple languages present.

6. NTL Outreach, Players, HITs

NTL was deployed on October 23, 2018, several months after the language games portal LingoBoingo. Because NTL launched too late to benefit directly from the first LingoBoingo advertisements, we posted announcements specifically about NTL to LDC members and linguists in early 2019. NTL also benefitted from a link added to the Great Language Game's farewell page encouraging its former players to try one of three similar games.

Figure 3 shows the number of HITs NTL presented to players per day on the log scale y-axis from October 2018 through March 2021. The red vertical lines represent outreach efforts: posts on 1) Facebook 2) Twitter and 3) SciStarter, 4) our nomination as a SciStarter Top Project, 5) the addition of a link on the Great Language Game's farewell page, 6) mentions in LDC social media posts and its newsletter and, 7) a post on Linguist List and mentions in its social media. The peak just before 7 results from a serendipitous feature on the YouTube Positive Mongolians vlog [33]. As might be expected, the number of HITs/day seems to vary along a weekly cycle, pushed upward to various degrees by outreach effort and trending downward in their absence.

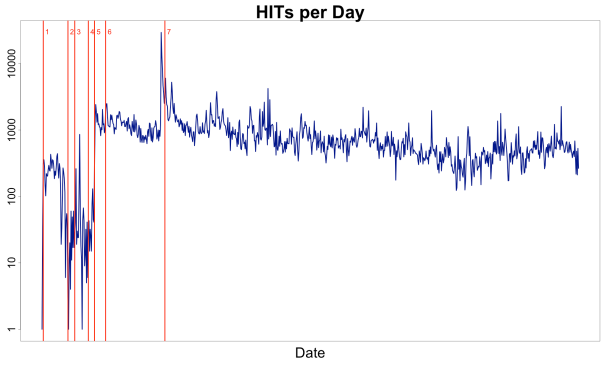

Figure 3: *NTL HITs/day Oct 2018 - March 25, 2021*

An analysis of LingoBoingo traffic revealed that many of the players recruited from social media accessed the site from mobile devices rather than desktops and often left without engaging with the games ("bounced"), particularly those whose interfaces did not adapt to handheld screen geometry. To accommodate, the NTL design was revised to be more responsive (Figure 2, right image).

As of March 22, 2021, NTL had presented 720,339 HITs to players who have returned useable responses 86% of the time (621,420). By algorithm, a response is judged unusable if it lacks a specific decision as to the language being spoken. Players may avoid submitting a specific guess by clicking the New Game button, by logging out or by changing browsers or machines. During bonus round they may also indicate that the language spoken is none of the choices offered by clicking the Other button. One can play anonymously or create and account. 45,740 UserIDs have been logged to date. This may overestimate the number of players because using different machines or browsers will create additional UserIDs. During account creation, players are asked to provide: user name (preferably not their real name), email address used only for verification and notification of game changes, and password. The player may optional provide: year of birth, gender, cities where the player has lived and language the player speaks. To date only 343 players have registered accounts.

As shown in Figure 4, NTL has players worldwide. Each point represents one or more players in a given city. The geographic diffusion suggests that we could expand the range

of languages covered and still find players able to identify those languages.

Figure 4: *Locations with at Least One NTL Player*

7. Results and Corpus Design

NTL game play assures that while any HIT completed provides information about confusability and the player's own knowledge no more the 1/3 of all HITs completed add to our knowledge of the language spoken in those clips. However, the real number of HITs on suspected clips is actually just above 20% (88,625) perhaps because players end games before completing the bonus rounds. However, we use data from known clips to understand the behavior of the players and their skills and to project performance onto suspected clips. A critical question is how many players must judge a clip before we are confident in its language. The answer depends in part upon the aggregation method used and in part upon the language. Starting from known clips and using simple voting, we see in Figure 5 the percentage of clips assigned to the correct language for a given number of HITs.

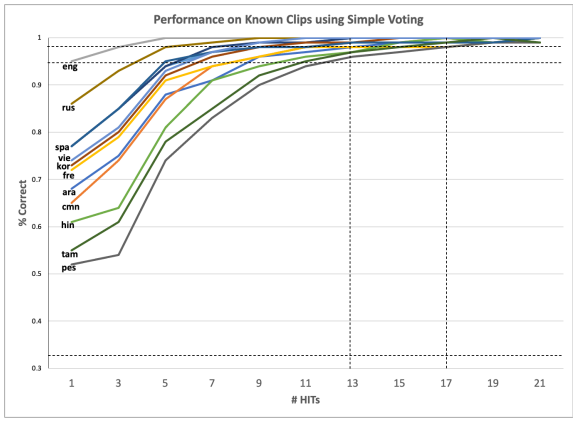

Figure 5: *Player Performance on Known Clips by # HITs Judged Aggregated with Prior Probability*

The uppermost horizontal dotted line marks an accuracy of 98% of clips correctly assigned; the lower line marks the 95% threshold. The vertical dotted lines show the number of judgements required, 13 and 15 respectively, to reach 95% and 98% accuracy for **all** languages. However, simple voting returns the correct answer at these accuracy thresholds with many fewer judgements for some languages. Our English language clips need only 1 judgement for 95% accuracy and 3 judgements for 98%. Finally, the dotted line at the bottom of the graph shows chance performance under the best of circumstances, when there are only three decision buttons and the a priori probability of guessing the correct answer by chance is .33. Individual players perform much better than chance. Based on Figure 5, using a very modest aggregation algorithm that only considers the prior probability that a player would guess a language by chance, we could set the threshold for declaring a clip effectively annotated once it has 13 or 15 judgements or we could can use a language specific threshold.

For ~540 clips, players submitted enough decisions to expect convergence on a specific language but did not converge. A review of 12 clips randomly selected from each of the 9 suspected languages revealed that ~96% were bad clips containing music or introductions or brief snippets of quoted speech in a language other than that of the broadcast.

Table 2 shows the number of clips considered useable according to the different criteria described above. The *NTL Language Recognition* corpus resulting from this effort will be released via LDC at no cost. It contains 6680 audio files and >720,000 database records indicating the file name, known or suspected language, other language choices offered during game play, city and country of the player, date and time of the HIT and other fields necessary for game administration. This data would support experiments not only in LRE but also in studies of confusability and experiments in the aggregation of crowd sourced data.

Table 2: *# Clips Annotated according to Different Criteria*

	Audio clips with enough judgements to meet accuracy				Audio clips assigned			
	Cross Language		Per Language		Known		Suspected	
Language	15 (98%)	13 (95%)	98%	95%	Total	% Skipped	Total	% Skipped
ara					38138	13%		
cmn	**506**	533	**549**	554	24621	19%	16226	17%
eng	**484**	522	**538**	538	89660	12%	16249	12%
fre	**462**	492	**506**	511	37960	12%	16165	16%
ger					38071	13%		
hin	**536**	567	**567**	586	38024	13%	16075	18%
jpn					64115	12%		
kor	**487**	510	**528**	530	37945	12%	16305	17%
pes	**503**	540	**503**	540	37778	12%	16216	23%
rus	**542**	572	**584**	584	37807	12%	16081	11%
spa	**458**	481	**493**	495	38215	13%	16394	18%
tam					38057	13%		
vie	**471**	489	**505**	505	37857	13%	16249	23%

8. Conclusions

We presented a Game with a Purpose that elicited judgements of languages spoken in clips from telephone conversation and broadcast to support Language Recognition and confusability research. Using simple aggregation, players responses identify clip languages with high accuracy and signal problematic clips when they do not converge on a single language. This proof of concept has yielded a corpus that helps to fill the gaps in available data for LRE and suggests new directions including addition languages and improved game play and aggregation methods.

9. Acknowledgements

The authors acknowledge the generous support of the US. National Science Foundation via grant CRI CI-NEW 1730377 that enabled the NIEUW project and its outcomes: LingoBoingo, NameThatLanguage, the resulting corpus and its free distribution, and other community-relevant portals such as LanguageARC, discussed elsewhere.

6. References

[1] Sarthak, S. Shukla and G. Mittal, "Spoken Language Identification using ConvNets," European Conference on Ambient Intelligence, 2019.

[2] P. Rangan, S. Teki and H. Misra, "Exploiting Spectral Augmentation for Code-Switched Spoken Language Identification," in *Proceedings INTERSPEECH 2020 - First Workshop on Speech Technologies for Code-switching in Multilingual Communities*, Shanghai, 2020.

[3] National Institute of Standards and Technology, "Language Recognition," [Online]. Available: https://www.nist.gov/itl/iad/mig/language-recognition. [Accessed 1 April 2021].

[4] National Institute of Standards and Technology, "Past HLT Evaluation Projects," 18 October 2016. [Online]. Available: https://www.nist.gov/itl/iad/mig/past-hlt-evaluation-projects. [Accessed 1 April 2021].

[5] Interspeech, "Interspeech 2019 Program Overview," [Online]. Available: https://interspeech2019.org/program/overview/. [Accessed 2 April 2021].

[6] "DBLP Computer Science Bibliography: INTERSPEECH 2000," DBLP, [Online]. Available: https://dblp.org/db/conf/interspeech/interspeech2000.html. [Accessed 2 April 2021].

[7] "DBLP Computer Science Bibliography: EUROSPEECH 1995," DBLP, 2 April 2021. [Online]. Available: https://dblp.org/db/conf/interspeech/eurospeech1995.html. [Accessed 2 April 2021].

[8] T. Lander, "CSLU: 22 Languages," Linguistic Data Consortium, Philadelphia, 2005.

[9] LIMSI-CNRS, "Multilingual Corpus for Language Identification," ELRA/ELDA, Paris, 1998.

[10] L. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy, J. Gangolf and J. Gauvain, "A Multilingual Corpus for Language Identification," in *First Language Resources & Evaluation Conference*, Granada, 1998.

[11] A. Canavan and G. Zipperlen, "CALLFRIEND German," Linguistic Data Consortium, Philadelphia, 1996.

[12] A. Canavan and G. Zipperlen, "CALLFRIEND Hindi," Linguistic Data Consortium, Philadelphia, 1996.

[13] S. Strassel, K. Walker, K. Jones, D. Graff and C. Cieri, "New Resources for Recognition of Confusable Linguistic Varieties: The LRE11 Corpus," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, 2012.

[14] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[15] H. Van den Heuvel, L. Boves, A. Moreno, M. Omologo, G. Richard and E. Sanders, "Annotation in the SpeechDat Projects," *International Journal of Speech Technology,* vol. 4, pp. 127-143, 2001.

[16] D. Caseiro and I. Trancoso, "Spoken Language Identification using the SpeechDAT Corpus," in *5th International Conference on Spoken Language Processing (ICSLP)*, Sydney, 1998.

[17] D. Miller, D. Graff, C. Cieri, K. Jones, S. Strassel, "CALLFRIEND Farsi Second Edition Transcripts," Linguistic Data Consortium, Philadelphia, 2014.

[18] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," *CoRR,* 2019.

[19] "Common Voice," [Online]. Available: https://commonvoice.mozilla.org. [Accessed 15 06 2021].

[20] J. Fiumara, "Linguist List," 8 January 2018 . [Online]. Available: https://linguistlist.org/issues/29/29-389/. [Accessed 1 April 2021].

[21] Linguistc Data Consortium, "Linguist List," 16 December 2017. [Online]. Available: https://linguistlist.org/issues/28/28-5338/. [Accessed 1 April 2021].

[22] Linguistic Data Consortium, "SciStarter - LingoBoingo," May 2019. [Online]. Available: https://scistarter.org/lingoboingo. [Accessed 1 April 2021].

[23] L. Shell, "Top 18 Projects of 2018," 14 January 2019. [Online]. Available: https://blog.scistarter.org/2019/01/top-18-projects-of-2018/. [Accessed 1 April 2021].

[24] J. Fiumara, "LingoBoingo: Play Games, Make the World Smarter," *Discover Magazine,* 21 December 2018.

[25] C. Madge, M. Poesio, U. Kruschwitz and J. Chamberlain, "Testing TileAttack with Three Key Audiences," in *Proc. LREC 2018 Workshop on Games and Gamification for Natural Language Processing*, 2018.

[26] L. Yencken, "TheGreatLanguageGame," [Online]. Available: http://greatlanguagegame.com. [Accessed 1 April 2021].

[27] L. Yencken, "Data: Language Confusion," 2 March 2014. [Online]. Available: https://lars.yencken.org/datasets/languagegame/. [Accessed 1 April 2021].

[28] L. Yencken, "Wide Language Index," [Online]. Available: https://github.com/larsyencken/wide-language-index. [Accessed 1 April 2021].

[29] H. Skirgård, S. G. Roberts and L. Yencken, "Why are some languages confused for others? Investigating data from the Great Language Game," *PLoS ONE,* vol. 12, no. 4, 2017.

[30] "Language Squad," [Online]. Available: https://www.languagesquad.com. [Accessed 15 06 2021].

[31] "LingYourLanguage," [Online]. Available: https://lingyourlanguage.com. [Accessed 15 June 2021].

[32] A. Martin and M. Pryzbocki, "NIST Language Recognition Evaluation (LDC2006S31)," Linguistic Data Consortium, Philadelphia, 2006.

[33] Positive Mongolians, "YouTube Positive Mongolians Channel," [Online]. Available: https://www.youtube.com/channel/UCXr6n3L8EDVQR-_EUJc-2Nw. [Accessed 1 April 2021].