APPENDIX: CORPORA TO BE REANNOTATED / MINED AND REFERENCES

AixOX
Longdale
PELCRA for spoken data

SPOKEN DATA
ANGLISH:
http://sldr.org/voir_depot.php?id=733&lang=en&sip=1
The corpus is multimodal and compiles several tasks; reading task, shadowing task and two-minute spontaneous speech. Though limited in numbers of speakers (20 native speakers, 20 intermediate and 20 advanced speakers), this corpus is multitask: having data for read speech, shadowing task and spontaneous data can improve the models by investigating
We will use the multimodal property of the ANGLISH corpus in so far as the continuous production differs from the read speech. The most important aspects of the corpus is the possibility to model the effect of task as the same text is also used for reading and shadowing task. The strategy is to extract specific data points responding to research questions we have identified and existing specific corpora will to help us in this undertaking to fine tune our models.

FRENCH
The French IPFC corpus can be used for the analysis of interphonology and especially of French vowels due to its emphasis on wordlist. The same applies to the English counterpart which has less data points. The AixOx corpus can be used to monitor Anglophones speaking French and English and vice versa and therefore help us build models based on transfer learning. We will use the neural network trained with levels of English to learn how to classify learners of French.
Racine, I., Detey, S., Zay, F., Y. Kawaguchi (2012). Des atouts d'un corpus multitâches pour l'étude de la phonologie en L2: l'exemple du projet « Interphonologie du français contemporain» (IPFC). In: Kamber, A., Skupiens, C. (éds). Recherches récentes en FLE. Berne: Peter Lang, pp. 1-19.

We have identified a key player in the NN analysis of learner data. NUS has a comparable expertise. We have made some preliminary contacts with the School of computer science and with the School of English. When the NUS-USPC CfP is re-launched, a solid partnership could be established allowing us to fine-tune our models with their annotated NUCLE corpus, a reference corpus based on Chinese learners partially used in data challenge in world-class conferences (the Building Educational Application workshop at ACL in 2019).

The ISLE corpus can be used to model the difference realisations of our L1s as Italian German and Spanish speakers carried out the same task. University of Paris has acquired the corpus through ELRA and the conversion of the files is currently underway.

WRITTEN DATA

REALEC (Russian as L1)

PLEC (Polish as L1)

HonkKong corpus for learners

NUCLE

Write&Improve

LOCHNESS

First Certificate in English dataset

EFCAMDAT

ICLE

ISLE CORPUS FOR LEARNERS
http://catalog.elra.info/product_info.php?products_id=568&language=fr
Menzel, W, Atwell, E, Bonaventura, P et al. (4 more authors) (2000) The ISLE corpus of non-native spoken English. In: Gavrilidou, M, (ed.) *Proceedings of LREC 2000: Language Resources and Evaluation Conference*, vol. 2. LREC 2000: Language Resources and Evaluation Conference, 31 May - 02 June 2000, Athens, Greece. European Language Resources Association, 957 - 964.

References

[1] Helen Yannakoudakis, Øistein E. Andersen, Ardeshir Geranpayeh, Ted Briscoe and Diane Nicholls. 2018. Developing an Automated Writing Placement System for ESL Learners. Journal of Applied Measurement in Education.

[2] Helen Yannakoudakis, Marek Rei, Øistein E. Andersen and Zheng Yuan. 2017. Neural Sequence-Labelling Models for Grammatical Error Correction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

[3] Youmna Farag and Helen Yannakoudakis. 2019. Multi-Task Learning for Coherence Modeling. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

[4] Simon Flachs, Ophélie Lacroix, Marek Rei, Helen Yannakoudakis and Anders Søgaard. 2019. A Simple and Robust Approach to Detecting Subject-Verb Agreement Errors. In

Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics.

[5] Youmna Farag, Helen Yannakoudakis and Ted Briscoe. 2018. Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input. In Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics.

[6] Dimitrios Alikaniotis, Helen Yannakoudakis and Marek Rei. 2016. Automatic Text Scoring Using Neural Networks. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics

[7] Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker and Tim Parish. 2013. Developing and Testing a Self-Assessment and Tutoring System. In Proceedings of the NAACL 2013 Workshop on Innovative Use of Natural Language Processing for Building Educational Applications.

[8] Grundkiewicz R, Junczys-Dowmunt M, Heafield K. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications.

[9] Zaidi AH, Caines A, Davis C, Moore R, Buttery P, Rice A. Accurate modelling of language learning tasks and students using representations of grammatical proficiency.

[10] Knill K, Gales M, Kyriakopoulos K, Malinin A, Ragni A, Wang Y, Caines A. 2018. Impact of ASR performance on free speaking language assessment. In Interspeech. International Speech Communication Association (ISCA).

[11] Moore R, Caines A, Graham C, Buttery P. 2015. Incremental dependency parsing and disfluency detection in spoken learner English. In International Conference on Text, Speech, and Dialogue.

[12] Craighead H, Caines A, Buttery P, Yannakoudakis H. Automated grading of learner English speech. (under review).

[13] Ballier, N , Gaillat T Simpkin, A Stearns B, Bouyé M Zarrouk M (2019) A supervised learning model for the automatic assessment of language levels based on learner errors, 14th European Conference on Technology Enhanced Learning (http://www.ec-tel.eu/), Delft, the Netherlands, 16-19 September 2019. Published In Scheffel M., Broisin J., Pammer-Schindler V., Ioannou A., Schneider J. (eds) *Transforming Learning with Meaningful Technologies*. EC-TEL 2019. Lecture Notes in Computer Science, vol 11722. Springer, Cham, Pages 308-320.

[14] Ballier, N, Gaillat, T & Pacquetet, E. (2019) Prototype de feedback visuel des productions écrites d'apprenants francophones de l'anglais sous Moodle, Actes de la 9ème Conférence sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH2019), édités par Julien Broisin, Eric Sanchez, Amel Yessad, Francoise Chenevotot, 395-398.

[15] Michaud D, Ballier N. 2018 Perception et production de /y/ et /u/ en français L2 chez l'apprenant anglophone débutant : étude de cas de leur catégorisation chez quatre locuteurs, JEP2018, Aix-en-Provence, 5 juin 2018. Proc. XXXIIe Journées d'Études sur la Parole, 231-239, doi.10.21437/JEP.2018-27