Palabras: Crowdsourcing Transcriptions of L2 Speech

Eric Sanders, Pepi Burgos, Catia Cucchiarini, Roeland van Hout
CLS/CLST, Radboud University Nijmegen, the Netherlands
{e.sanders, j.burgos, c.cucchiarini, r.vanhout}@let.ru.nl

Abstract

We developed a web application for crowdsourcing transcriptions of Dutch words spoken by Spanish L2 learners. In this paper we discuss the design of the application and the influence of metadata and various forms of feedback. Useful data were obtained from 159 participants, with an average of over 20 transcriptions per item, which seems a satisfactory result for this type of research. Informing participants about how many items they still had to complete, and not how many they had already completed, turned to be an incentive to do more items. Assigning participants a score for their performance made it more attractive for them to carry out the transcription task, but this seemed to influence their performance. We discuss possible advantages and disadvantages in connection with the aim of the research and consider possible lessons for designing future experiments.

Keywords: crowdsourcing, transcription, L2 speech

1. Introduction

In their research, (Burgos et al., 2013; Burgos et al., 2014; Burgos et al., 2015) studied the pronunciation of Dutch by Spanish L2 learners. Judgements of pronunciation quality were obtained from experts (Burgos et al., 2013; Burgos et al., 2014). However, judgements by nonexpert Dutch native listeners are also relevant and informative, as they can reveal which features of the learners' vowel realizations may lead to confusions in perception. To get large numbers of transcriptions, it was decided to use crowdsourcing for data collection (Burgos et al., 2015). The use of crowdsourcing to obtain annotations or scorings of intelligibility or accentedness of non-native speech is not new (Evanini et al., 2010; Cooke et al., 2013; Wang et al., 2013). For our purpose, we built a web application that allows participants to listen to utterances and transcribe what they hear. In their crowdsourcing experiment (Cooke et al., 2013) had observed that limited feedback could lead to low task engagement. For this reason, we decided to build in a few feedback parameters, to see whether the presence or type of feedback would impact the number and nature of the transcriptions. We also asked for metadata such as gender, age and completed education to be able to study how these variables affect crowdsourcing behaviour/participation.
The results of the analyses of the data that were collected with the application are reported in (Burgos et al., 2015). The current paper describes the application, the feedback parameters, metadata and their influence on the results.

2. Application

Before designing the application, we defined a number of criteria aimed at maximizing response:

- The application had to be easy to use.
- The task had to be "fun" to do.
- It had to be shared on Facebook to attract new participants.
- Transcribers had to participate voluntarily.
- Participants should be able to return and continue from where they left.

This led to the web application that we called *Palabras*, the Spanish word for 'words'. Our conditions were met in the following way:

- The task is very easy: the participant listens to a word that is played (with option to repeat), enters what (s)he hears and the next sound is played. The login procedure is also very easy.
- We added a score, so participants could compare how well they did and share the score by posting it on Facebook.
- After completing 50 items, participants could share their score on Facebook. By clicking on the picture (figure 1) their contacts were directed to the application.
- No (monetary) compensation was given to the participants.
- By using a login procedure, the application remembers which items a participant had transcribed and can continue from there.

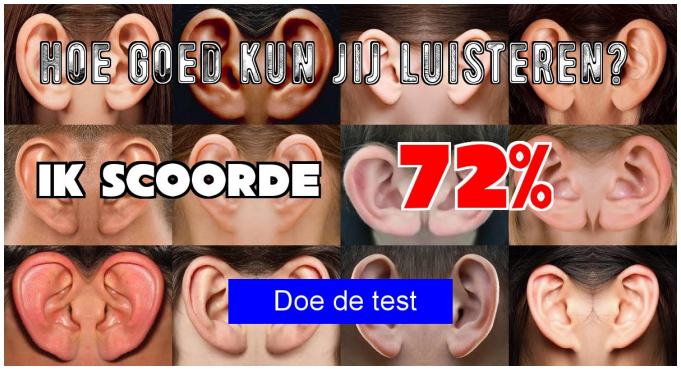

Figure 1: Image on Facebook to share score and link.

The application consists of two screens:

- A screen with basic explanation of the task and the two login options (see figure 2). Participants can either use their Facebook account to login or can register and login at the same time by choosing a username/password combination.
- A screen with the main application including feedback, the main parts of the explanation and a metadata fillin form (see figure 3). As soon as the metadata are filled in, the form disappears. The metadata that are asked from the participants are age group (10-20, 21-30, 31-40, 41-50, 51-60, >60), gender and level of completed education (4 levels of Dutch education: lbo, havo/vwo, hbo, wo) and their mother tongue, if different from Dutch.

The application gives three types of feedback:

- A score with percentage correct. Each transcription is compared to the majority transcription of all transcribers (the transcription that is transcribed most often). If it is the same the transcription is counted as "correct". Mind that there is not a correct or incorrect transcription; one hears what one hears. But this was the closest we could get to a simple score. This type of feedback was given to all participants.
- Transcription of previous utterance by the participant and by the majority. This gives the participant the possibility to compare his/her transcription with the majority and check if it is "correct". This type of feedback was given to half of the participants (either always or never).
- Information about the number of items. This contains the total number of items and the number of items that are already transcribed or still have to be done. Half of the participants did not get this type of feedback. Of the other half, half (a quarter of the total) got the number of items already transcribed and half got the number of items that still had to be done.

The speech material to be transcribed consisted of 29 monosyllabic Dutch words pronounced by 28 Spanish learners of Dutch. Six items were unusable and were removed. The 29 words contain the 15 vowels of Dutch followed by /s/ or /t/, as these consonants are known to alter the preceding vowel least (van der Harst, 2011; van der Harst et al., 2014). Only the sequence /y/ + /s/ is missing, since there are no Dutch monosyllabic nouns with this combination (except proper names). See table 1 for all words presented as stimuli.
The utterances were randomly chosen and presented to the participants in such a way that they did not get the same utterance (same word spoken by same speaker) twice. However every 30th utterance was a randomly chosen utterance that had already been transcribed and which was presented a second time to be able to calculate intratranscriber agreement. We also took care that the utterance that had to be transcribed was of a different word type than that of the last 20 transcribed utterances to prevent a carryover or learning effect.

Vowel	s-word	phonemic	t-word	phonemic
/ɪ/	kies	/kɪs/	riet	/rɪt/
/ɪ/	vis	/vɪs/	fit	/fɪt/
/ɛ/	zes	/zɛ/t/	vet	/vɛ/t/
/y/	-	-	fuut	/fyt/
/ʏ/	zus	/zʏs/	put	/pʏt/
/u/	poes	/pus/	voet	/vut/
/ɔ/	vos	/vɔs/	vlot	/vlɔt/
/ɑ/	gas	/xɑs/	rat	/rɑt/
/a/	aas	/as/	staat	/stat/
/e/	mees	/mes/	beet	/bet/
/ø/	neus	/nøs/	neut	/nøt/
/o/	boos	/bos/	boot	/bot/
/ɛi/	ijs	/ɛis/	spijt	/spɛit/
/œy/	huis	/hœys/	fluit	/flœyt/
/ɔu/	kous	/kɔus/	fout	/fɔut/

Table 1: Word stimuli used in crowdsourcing transcriptions.

3. Results

Almost 200 people participated and produced an average of 100 transcriptions. See figure 4 for a distribution of the number of items transcribed. About 70% of the participants transcribed more than 50 items and 3 participants transcribed more than all 806 words. Over 90% did only 1 session (a new session is started when there is more than 1 hour between two items). Three participants did five sessions. About 90% of the participants filled in their metadata.

3.1. Quality control

We checked the quality of the data in several ways. We applied filters to remove the following transcribers and transcriptions from our data set:

- Testers of the application and the authors of the paper.
- Transcribers that had indicated to have another native language than Dutch.
- Transcribers with less than 10 transcriptions, these are not regarded as serious participants.
- Transcribers that did not fulfill our quality criteria (inter and intratranscriber agreement below threshold).
- Transcriptions that were entered more than once (when the server was slow in response).
- Transcriptions that were produced after the whole set of stimuli had been completed.

3.2. Transcriptions

The main goal of the data collection was to find out how the pronunciation of the Dutch words by Spanish learners was perceived by nonexpert listeners. In 62% of all 17534 transcriptions the canonical transcription of the target word was used. In 19% of the cases the most often used alternative was selected and in another 18% another variant was

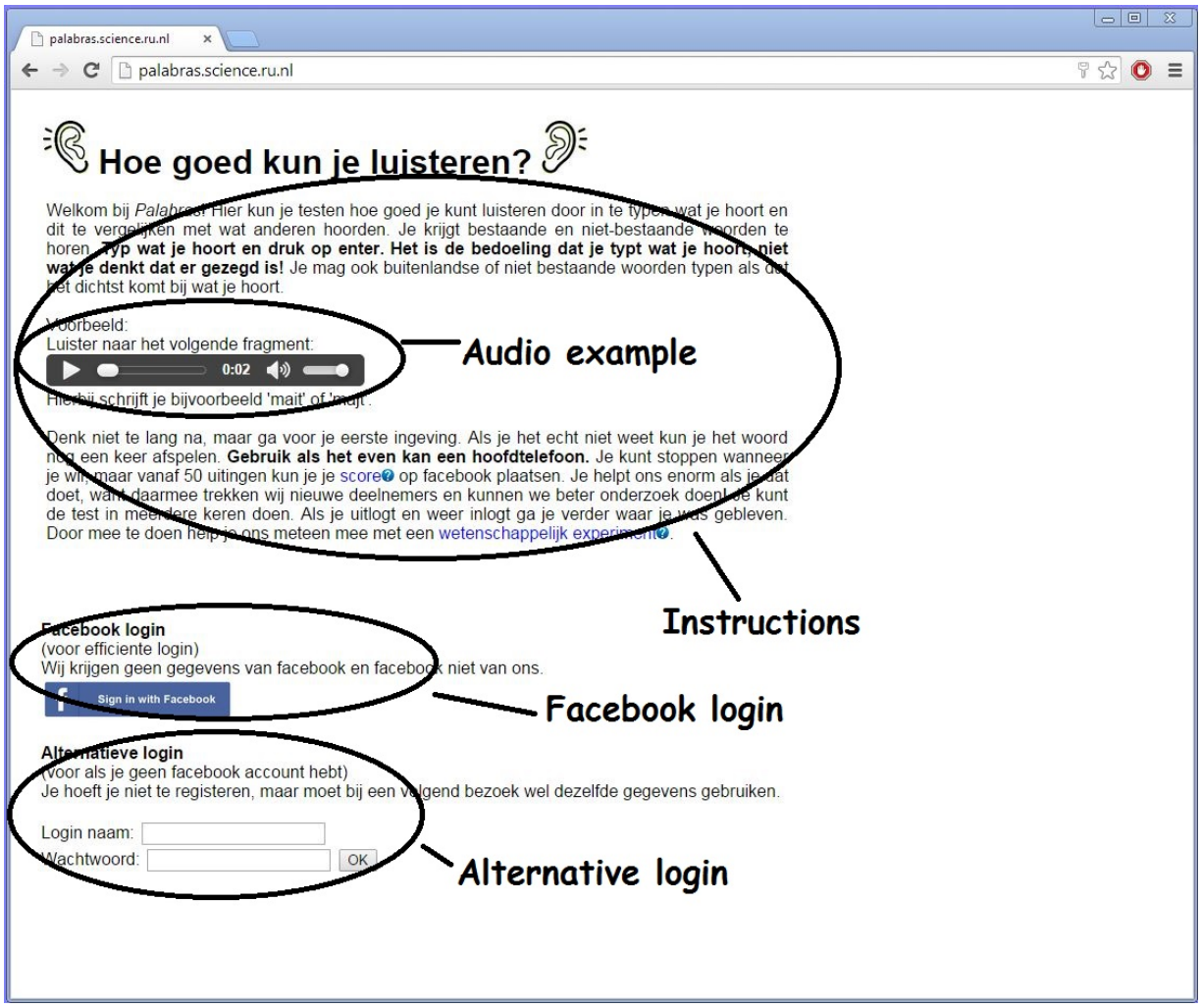

Figure 2: Introduction screen of *Palabras*.

chosen. See (Burgos et al., 2015) for concrete results on the transcription variants. The main conclusion is that the nonexperts found in general the same effects as experts.
The score of a transcriber is defined as the percentage of "correct" transcriptions. A transcription is "correct" if it is the same as the most frequently used transcription. In 81% of the cases the canonical transcription was the correct transcription. Figure 5 shows the percentages correct transcriptions that the participants scored. The mean score is 67% and the medium score is 69%.

3.3. Metadata

In this subsection the relation between the metadata and the number of transcribers, transcriptions and scores are presented. This is only done for those transcribers that filled in the metadata, thus the numbers do not add up to the total numbers of participants and transcriptions.
What stands out from this table is the high percentage of female participants (table 2). Almost three times as many women participated and they transcribed 1.5 times more items on average than men, which is a significant difference (t' (130.328) = 2.203, p =.029). The percentage of "correctly" transcribed words was not significantly different.

gender	#participants	#words	average #words	%correct words
men	35	3056	87	69
women	104	13370	128	68

Table 2: Results for men and women.

age group	#participants	#words	average #words	%correct words
10-20	22	2097	95	63
21-30	56	7529	134	70
31-40	22	1984	90	68
41-50	10	799	79	66
51-60	21	2129	101	67
>60	8	1917	239	66

Table 3: Results for different age groups.

One third of the participants are in age group of 2130. This group also produced a high average of transcribed items. The large average of transcribed words in the group of participants older than 60 is mainly due to one person who

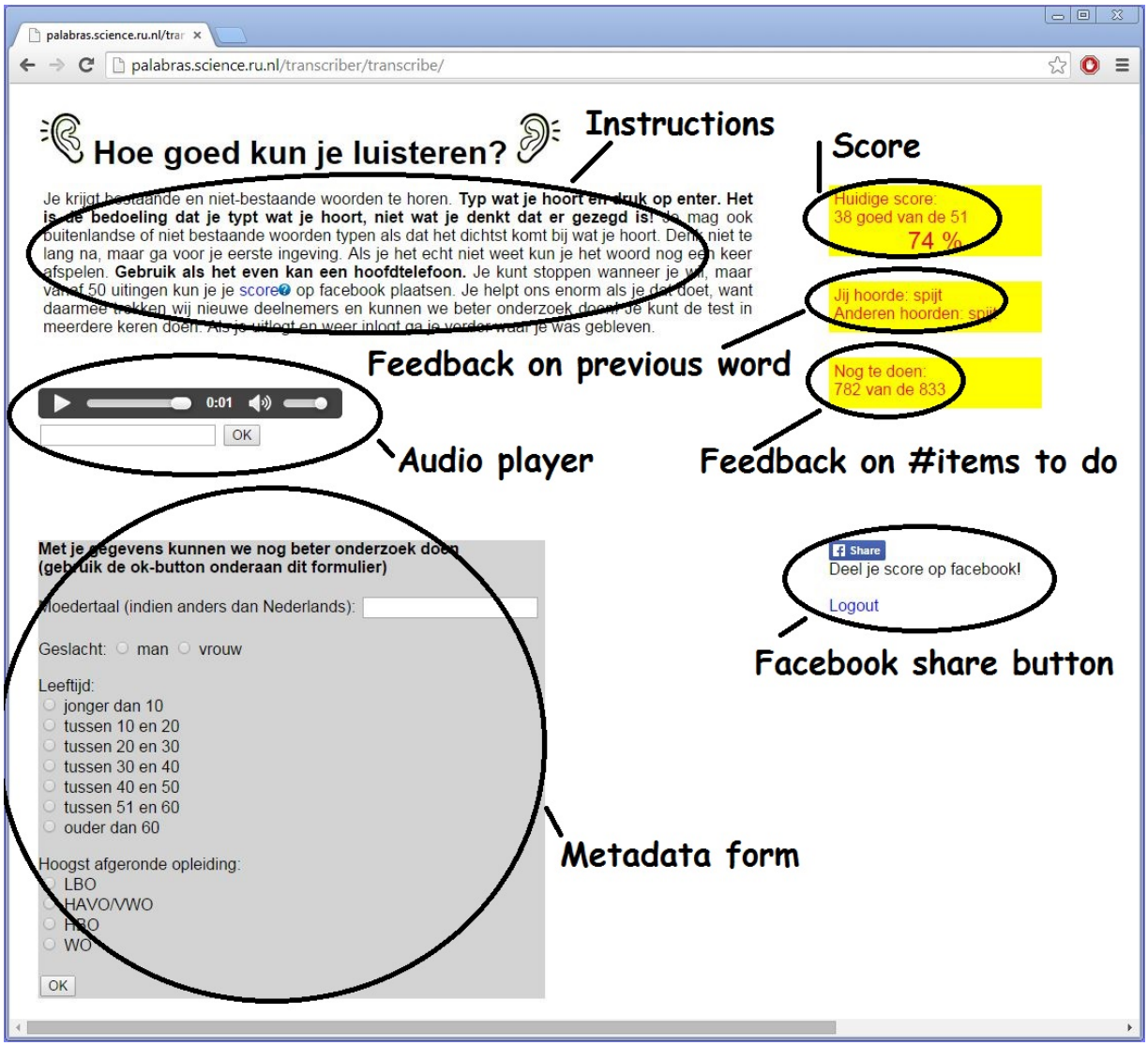

Figure 3: Transcription screen of *Palabras*.

transcribed all items. The word correct scores are more or less the same, except for the participants in the youngest category who score a bit lower than the rest.

education	#participants	#words	average #words	%correct words
lbo	3	210	70	60
havo/vwo	42	5427	129	70
hbo	30	3334	111	65
wo	61	6963	114	69

Table 4: Results for different education levels.

The lowest education level has only few participants, who transcribed few words and scored low on percentage correct. The other three groups behave similarly to each other. The group of participants with a university degree is relatively large. This is not surprising since recruitment started from people in this category.

3.4. Feedback

In this subsection the different feedback parameters in relation with the number of participants, transcriptions and scores are presented.

feedback	#participants	#words	average #words	%correct words
with	77	8044	104	72
without	82	9490	115	64

Table 5: Results for different feedback on previous word.

Table 5 shows the results for the two groups of participants, one of which got feedback information over the previous word and the other which did not. Getting feedback on the previous word did not result in transcribing more items, but it did lead to significant higher scores (t' (156.848) = 2.58, p = .012). This is to be expected: participants who get this feedback, learn what they have to do to get a good score and can adapt their strategy in this direction. In practice,

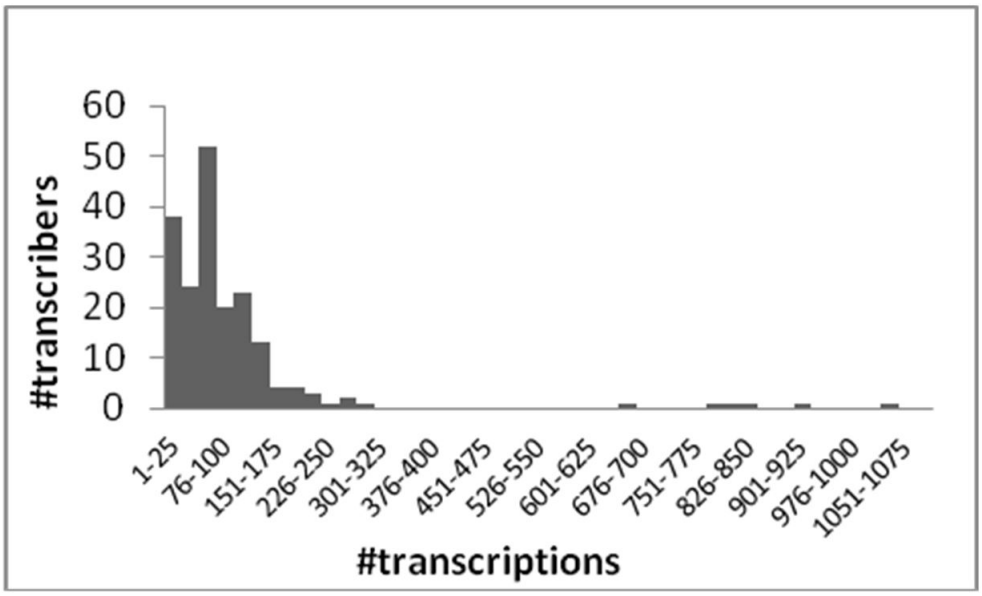

Figure 4: Distribution of number of transcribed items.

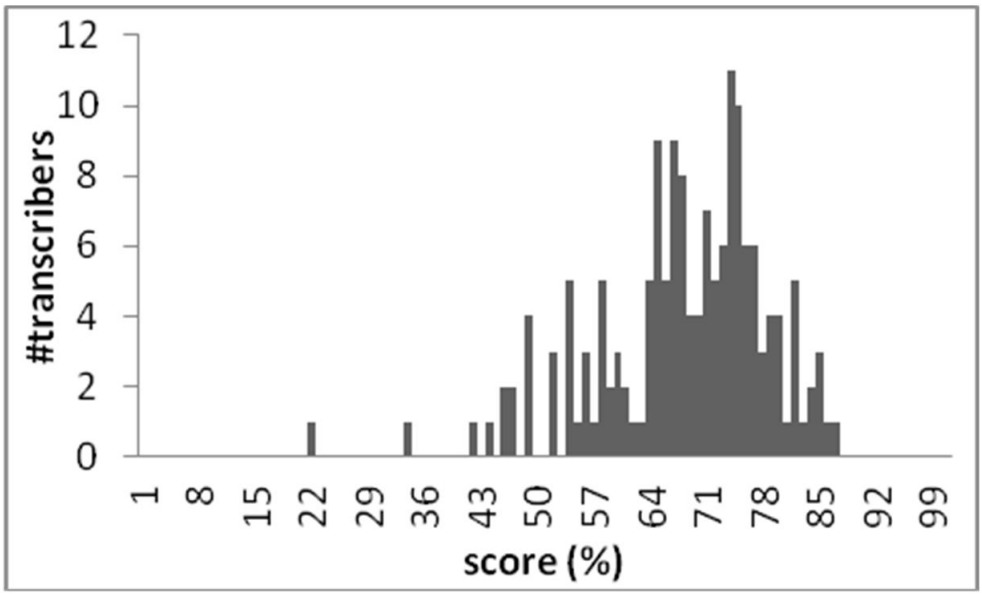

Figure 5: Distribution of scores.

this means they will transcribe more often what they think a participant intended to say, probably neglecting (small) pronunciation errors.

feedback	#participants	#words	average #words	%correct words
todo	48	7077	147	70
done	40	3557	88	61
none	71	6900	97	69

Table 6: Results for different feedback on number of words.

Table 6 shows the results with respect to the feedback about the number of words in the transcription set. Half of the participants got the number that had to be transcribed in total. Half of them got the number that still had to be done (todo) and the other half got the number that had already been done (done). The other half of the participants got none of this information (none). The participants that received to do information transcribed on average far more items than the other two groups. Because of the high standard deviation in the number of transcribed items this is not significant ($F(2,156) = 2.289$, $p = .105$), but the tendency is clear.

This feedback might be an incentive to continue. The group that received done information scored lower than the other two groups, but the differences between the groups were not significant ($F(2,156) = 2.614$, $p = .077$).

4. Discussion and Conclusions

When we started the crowdsourcing experiment we recruited participants in our direct social network, but eventually many of the participants are unknown to the authors, which means that the Facebookshare method worked. With 159 useful participants and over 20 transcriptions per item on average, this crowdsourcing method is definitely a satisfactory result for L2 speech research.

An unexpected result of the crowdsourcing was the large proportion of women that participated. We do not have a direct explanation for this. The recruitment started in environments with equal numbers of both genders and we do not know whether the transcription task was more appealing to women than to men. We have no indication that this difference might have influenced our results, since men and women scored almost equally. Giving the user information about how many items were to be transcribed in total seemed to be an incentive to do more items, but strangely only in the case when presented with the number of items still to be done and not with the number of items that had been done.

It turned out that in an application in which participants do not get any monetary remuneration (Cooke et al., 2013), adding a score to the application made it more attractive to do the transcription task. We got feedback from participants indicating that the score indeed did stimulate them to go further with the task, for example to beat their friends score. It has the disadvantage though that the participants main goal might not be to transcribe precisely what they hear, but what they think will give them a higher score. Users appeared to be confused when they transcribed what they heard, but the correct transcription appeared to be something else. Sometimes they adapted their strategy by transcribing what they thought was meant to be said to get a higher score. This gives a bias towards the canonical transcription. However, clear pronunciation errors still got the noncanonical transcription in the majority vote, which indicates that serious pronunciation errors were penalized anyway, while less serious errors were not noted down because they are probably considered not to hamper communication.

Looking back at the goals of the present study, 1) to evaluate the transcription system designed and its parameters, and 2) to determine how feedback and reward affect transcribing behavior in the context of crowdsourcing, we can conclude that 1) the overall system worked satisfactorily and produced a considerable amount of interesting data, and that 2) feedback and reward had a positive effect because they motivated the participants to continue as in (Kaufmann et al., 2011), but they did not always have a desirable effect on transcription behavior, which can be considered an important lesson for designing future experiments.

The software that was developed for *Palabras* is reused for another project in which tweets are annotated. The software is open source and can be obtained by contacting the first author of this paper.

5. Bibliographical References

Burgos, P., Cucchiarini, C., Hout, R. V., and Strik, H. (2013). Pronunciation errors by spanish learners of dutch: a data-driven study for asr-based pronunciation training. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 2385–2389.

Burgos, P., Cucchiarini, C., van Hout, R., and Strik, H. (2014). Phonology acquisition in spanish learners of dutch: error patterns in pronunciation. *Language Sciences*, 41, Part B:129 – 142.

Burgos, P., Sanders, E., Cucchiarini, C., Hout, R. V., and Strik, H. (2015). Auris populi: crowdsourced native transcriptions of dutch vowels spoken by adult spanish learners. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 2819–2823.

Cooke, M., Barker, J., and Lecumberri, G. (2013). Crowdsourcing in speech perception. In *In Eskenazi, M. and Levow, G.A. and Meng, H. and Parent, G. and Suendermann, D. (eds.) Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. Wiley publishing, 2013.

Evanini, K., Higgins, D., and Zechner, K. (2010). Using amazon mechanical turk for transcription of non-native speech. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 53–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kaufmann, N., Schulze, T., and Veit, D. J. (2011). More than fun and money. worker motivation in crowdsourcing - a study on mechanical turk. In *Proceedings of the 17th Americas Conference on Information Systems (AMCIS 2011)*, page Paper 340, Atlanta, Ga. AISeL. Online Ressource.

van der Harst, S., van de Velde, H., and van Hout, R. (2014). Variation in standard dutch vowels: The impact of formant measurement methods on identifying the speaker's regional origin. *Language Variation and Change*, 26:247–272, 7.

van der Harst, S. (2011). *The vowel space paradox. A sociophonetic study on Dutch.* Utrecht: LOT Dissertation, The Netherlands.

Wang, H., Qian, X., and Meng, H. (2013). Predicting gradation of L2 english mispronunciations using crowdsourced ratings and phonological rules. In *ISCA International Workshop on Speech and Language Technology in Education, SLaTE 2013, Grenoble, France, August 30 - September 1, 2013*, pages 127–131.