Hyperarticulation of Corrections in Multilingual Dialogue Systems

Ivan Kraljevski, Diane Hirschfeld

voice INTER connect GmbH, Dresden, Germany

{ivan.kraljevski,diane.hirschfeld}@voiceinterconnect.de

Abstract

This present paper aims at answering the question whether there are distinctive cross-linguistic differences associated with hyperarticulated speech in correction dialogue acts. The objective is to assess the effort for adaptation of a multilingual dialogue system in 9 different languages, regarding the recovery strategies, particularly corrections. If the presence of hyperarticulation significantly differs across languages, it will have a significant impact on the dialogue design and recovery strategies.

Index Terms: hyperarticulation, multilingual speech dialogue systems, human-computer interaction

1. Introduction

The development of a multilingual Spoken Dialogue System (SDS) requires significant effort, resources, time and expertise. It would be more feasible if most of the features could be shared across languages without extensive adaptation and providing equal level of performance [1].

The question is, at which extent a multilingual dialogue system actually needs to be adapted to a specific language or culture, while retaining most of its original design. Conducting "Wizard of Oz" (WOz) experiments with native speakers would provide clues about how language specific adaptation of the dialogue system should be performed. The speakers' experience is simulated to prove the design concepts, to collect relevant data and to analyze the behavior and expectations of the participants.

Automatic speech recognition is still a challenging task and recognition errors are ultimately unavoidable, causing communication problems in spoken human-computer interaction. Thus, it is of great importance to facilitate detection and correction of miscommunication. For humans, recognizing and clarifying such situations utilizing lexical and syntactic cues is part of everyday communication. However, a robust dialogue system has to be able to detect miscommunication and to apply appropriate recovery and error strategies.

Many research studies are dealing with the topic of prediction, detection and reduction of miscommunication in SDS. In [2], the data-driven approach for detecting instances of miscommunication is described. The authors in [3] proposed a system which integrates an error correction detection module with a modified dialogue strategy. In the study [4], a machine-learning approach employed automatically derived prosodic features, the speech recognition process, experimental conditions and the dialogue history to identify user corrections of speech recognition errors. An error handling strategy based on dynamically created correction grammars for recognizing correction sentences is described in [5].

Using prosodic features for recognizing and classifying dialogue acts in general was investigated in [6]. In [7] the duration, pause, and pitch features were employed to train a decision tree classifier, which was extended and integrated with recognizer confidence scores for further improvements in detection of corrections [8]. The speaking style changes associated with correction dialogue acts. Hyperarticulation can be used as a clue in order to identify problematic turns.

The authors in [9] observed that human speech during error resolutions shifts to become lengthier and more clearly articulated. A similar study presented in [10] shows that English speaker utterances of correction and non-correction dialogue acts differ prosodically in ways consistent with hyperarticulated speech. They defined it as: "slower and louder speech with wider pitch excursion and more internal silence". This change in speaking style is most likely to provoke recovery "error spirals", consequently ending the machine dialogue unsuccessfully. Some studies have shown this phenomenon is associated with higher recognition error rate [11-12] and that speakers hyperarticulate more often after several errors [13]. In [14] longer average phone duration, significant changes in pitch and fundamental frequency were observed in hyperarticulated German speech data. However other studies reported minor changes [15] or no negative impact [13] on error rate.

In this paper, we are investigating cross-linguistic differences related to hyperarticulated speech in correction dialogue acts. The objective is to assess the effort for adaptation of a multilingual dialogue system from the aspect of recovery strategies. If the presence of hyperarticulation significantly differs across languages, it will consequently have significant influence on the dialogue design as well as the recovery strategies. Although there are many research studies dealing with cross-linguistic prosodic differences [16-19], they are mostly done on a pair of languages and on a limited number of participants. By our best knowledge, there are no research studies dealing with simultaneous investigation of the characteristic features of hyperarticulated speech on several languages in parallel.

2. Speech Database

In order to collect a multilingual database containing audio-visual data, as well as to find the answer to how speakers of different native languages interact with an intelligent smart-home system, we prepared and performed WOz experiments. In a preparatory phase, an online questionnaire with a total of 870 participants was carried out in 16 languages [20]. The results provided suggestions for the dialogue design and revealed general user expectations. In the implementation phase, 19 different user scenarios for usage and control of smart home devices were designed and the corresponding dialogs were created.

The scenarios were carefully designed to elicit spontaneous reactions and to trigger recovery behavior from the participants in case of miscommunication.

The WOz experiments were carried out for the following languages (abbreviation and number of participants in brackets): English (EN:40), German (DE:40), French (FR:23), Spanish (ES:27), Italian (IT:19), Dutch (NL:15), Finnish (FI:7), Norwegian (NO:7), Swedish (SE:6), Danish (DA:8), Russian (RU:20), Turkish (TR:20) and Mandarin Chinese (CN:19). The dialogs were translated and adapted for all languages, while keeping the same meaning and the semantic structure whenever possible.

During the session, the wizard triggered speech dialogue acts and device functions to simulate a perfect dialogue system. Before the session start, the participants were asked to sign a consent waiver and they received reimbursement for their participation. Afterward, the moderator would introduce the system, give examples about the usage of the greeting keywords and demonstrate a simple scenario.

Miscommunication was simulated by introducing embedded error speech prompts, categorized as:

- Substitutions: wrongly recognized parameters;
- Insertions: confirmation of non-uttered sentence;
- Deletions: request to repeat the last sentence.

The participants were asked to use a special keyword in case of misunderstanding to perform correction, for example, English *"Correction"*, German *"Korrektur"* and so on. The explanation for the participants was that the system is more robust in presence of noise when using such special keywords. The maximum number of error prompts (around 20%) in a session was estimated over the number of the required parameters (options, entries) per scenario, including occasional system rejections and repetitions.

The error prompts were not triggered automatically, but manually by the wizard. However, not all the planned error prompts were played since the actual dialogue flow never reached intended states where the errors should be introduced. By our best knowledge, none of the participants realized that he or she was interacting with a human operator and not with a machine. The experiments were staged in a time span of several months. Approximately 4500 scenarios for all languages were fulfilled yielding audio-visual recordings with a duration of 125 hours. The speech was segmented and orthographically transcribed by native speakers (the associates engaged in the WOz), producing detailed descriptions of dialogue flows for each scenario. Having such parallel multilingual corpus is a solid basis to perform investigations on the behavioral patterns of native speakers, from the linguistic as well as paralinguistic aspects.

3. Data and Tools

3.1. Data organization

Common data-sets were compiled for all languages, based on collected corpus and the timestamped logs of the dialogue acts. Nine languages were selected for further analysis according to the number of participants (at least 15): German, English, Spanish, French, Italian, Turkish, Russian, Mandarin Chinese and Dutch. Since the main objective is the analysis of features corresponding to hyperarticulation, we selected pair of utterances of "statement" and "correction" dialogue turns (in total 3026).

3.2. Prosodic features

For the prosodic features analysis of the selected utterance pairs, the following Praat [21] scripts were used.

"Praat Script Syllable Nuclei v2" [22] was used for automatic detection of syllable nuclei in order to estimate the speech rate without the need of manual transcription. Peaks in intensity (dB) that are preceded and followed by dips in intensity are considered as potential syllable nuclei, while the peaks that are not voiced were discarded. The following measures were considered: speech rate (nsyll/speech-duration), articulation rate (nsyll/phonation-time) and average syllable duration (phonation-time/nsyll). Where *nsyll* is the number of syllables detected in either speech duration or phonation time.

"ProsodyPro 6beta" [23] was used for systematic analysis of the data-sets to generate detailed discrete prosodic measurements suitable for statistical analysis: maximal *f0* (Hz), minimal *f0* (Hz), pitch excursion (semitones), averaged *f0* (Hz), averaged intensity (dB) and maximum *f0* velocity (semitone/s).

In order to simulate real conditions in a smart-home environment during experiments, there was a frequent and significant presence of stationary and non-stationary noise. The transcribers were instructed in such cases to label longer leading and trailing silences as a requirement for later speech recognition evaluations. However, such noisy segments would clearly disturb the prosody analysis and render it unreliable. Therefore the first tool, aside from measuring speech rate, was also used to segment silence and speech as an input for the ProsodyPro tool. The same script settings were used for all audio recordings providing consistency for the cost of some erroneous segmentation which were identified and corrected by a human expert prior to analysis with ProsodyPro.

At the end, the measured values for of the "statement" turns were subtracted from those of the corresponding "correction" turn, producing a data-set representing quantitative changes in prosodic features. Delta values (Δ) for the "correction" features are better suited for analysis since they are already paired with the preceding "statement", in the same time compensating speaker and environment specific influences.

4. Results

4.1. Exploratory statistics

The total distribution of the introduced errors on all languages is: deletions 35.85%, insertions 8.20% and substitutions 55.95%. Table 1 presents the introduced error distribution across languages. For some languages there are differences in the count of insertions errors, because often the speakers were quite confused providing no answer that could be paired with the previous statement.

Table 1: *Error category distribution.*

%	CN	DE	EN	ES	FR	IT	NL	RU	TR
DEL	38.57	29.00	36.69	39.15	36.27	41.06	28.29	43.24	34.71
INS	2.69	20.38	3.35	3.44	9.15	2.85	9.87	5.74	5.50
SUB	58.74	50.63	59.96	57.41	54.58	56.10	61.84	51.01	59.79

Gender distribution of the participants on all languages was: female 56.74% vs male 43.36% (Table 2).

Table 2: *Gender distribution.*

%	CN	DE	EN	ES	FR	IT	NL	RU	TR
F	70.85	58.31	42.21	51.59	48.47	53.66	53.95	77.03	66.32
M	29.15	41.69	57.79	48.41	51.53	46.34	46.05	22.97	33.68

The participants were instructed to use the appropriate correction keyword to solve the miscommunication. However in 61.86% of the cases, the speakers did not use it. Table 3 presents the frequency of using the correction keyword across languages. To confirm that the observed differences are significant across languages, for each speaker the frequency of the usage was calculated and Kruskal-Wallis test confirmed the observations ($p<0.001$).

Table 3: *Correction keyword usage.*

%	CN	DE	EN	ES	FR	IT	NL	RU	TR
NO	54.71	61.91	57.99	44.44	53.56	44.72	86.84	77.70	90.38
YES	45.29	38.09	42.01	55.56	46.44	55.28	13.16	22.30	9.62

The transcribed correction responses were further analyzed and categorized according to the utterance content:

- Different, non-matching content (DIFF);
- Full, identical content (FULL);
- Mixed, statement contained in correction (MIX);
- Partial, correction contained in statement (PART).

Assuming non-normal distribution, Kruskal-Wallis test for each response category showed that they are language dependent ($p<0.001$), presented in Figure 1.

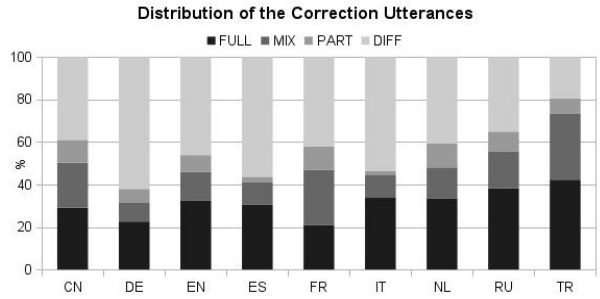

Figure 1: *Distribution of the correction utterances into categories across languages.*

In a case of totally different and mixed responses, the keyword was part of the answer in 65.22% and 49.16% of the cases respectively. The rest of the responses contained arbitrary content difficult to foresee and handle from the aspect of speech technology by using a restricted set of dialogue acts and responses.

4.2. Normality Test

Firstly, the presence of characteristic hyperarticulated speech for correction turns has to be confirmed, in [9-10] it is defined, with reference to prosody, as "slower and louder speech with wider pitch excursion and more internal silence".

Shapiro Wilk normality test on the delta features, regardless of the language, showed that they are not represented with normal distribution, as is usually true for large sample count real data. Non-parametric Wilcoxon test applied on all the delta features, regardless of the language, showed that the minimum *f0* is significantly higher for the correction turns ($p<0.001$), there are less pitch excursions ($p<0.001$), higher maximal velocity ($p<0.001$), longer speech duration with unchanged phonation time ($p<0.001$), lower speech rate ($p<0.001$), lower articulation rate ($p<0.001$) and longer average syllable duration ($p<0.001$). This confirmed the presence of distinctive prosodic features particularly related with slower speech, indicating hyperarticulation. Kruskal-Wallis rank sum test on the delta features with language as a factor showed ($p<0.01$) that except for minimum *f0* (floor of the speaker's physiological *f0* range), all other features belong to non-identical populations.

4.3. Linear Mixed Models

More elaborate analysis was performed to show whether some prosodic and speech rate features are more significant in different languages. Since the prosody features were measured on speech recorded in a complex experimental setup with speakers in many languages, better insight about the effects of the factors could be given by employing Linear Mixed Models [24].

The respective measures were taken as dependent variables, speaker and the scenarios as random factors, the native language, gender and the introduced error type (insertions, deletions, substitutions) as fixed factors, as well as all their possible interactions were included in the model.

For each delta feature, we constructed an initial model as described and then performed automatic backward elimination of non-significant effects. The p-values for the fixed effects are calculated from F test based on Satterthwaite's approximation. The gender alone had a significant effect on the delta values of mean *f0* ($F[1,377] = 4.5251$, $p<0.05$), female speakers lowered their *f0* more than males in the correction turn. Gender had a significant effect on the maximal *f0* velocity ($F[1,204] = 7.5214$, $p<0.01$), where female speakers had higher maximal pitch velocity.

The gender in interaction with language had significant effect for maximal *f0* ($F[8,199] = 2.3244$, $p<0.05$). The interaction of the factor gender with factor error type was significant for: mean *f0* ($F[2,2953] = 7.1530$, $p<0.001$), mean intensity ($F[2,2959] = 7.7792$, $p<0.001$), articulation rate ($F[2,2957] = 5.0150$, $p<0.01$) and average syllable duration ($F[2,2957] = 3.4432$, $p<0.05$).

Table 4: *Analysis of variance table.*

Feature (Δ)	DenDf	F	p
Speaker language (NumDf=8)			
mean intensity	534.62	2.9547	0.01
max *f0* velocity	636.03	3.4290	0.001
speech rate	614.97	3.8572	0.001
articulation rate	594.41	3.7968	0.001
avg. syll. duration	569.10	2.9333	0.01
Introduced error (NumDf=2)			
max *f0*	2954.54	4.1352	0.05
mean *f0*	2935.74	15.1808	0.001
mean intensity	2943.11	6.7247	0.01
speech rate	281.32	8.8997	0.001

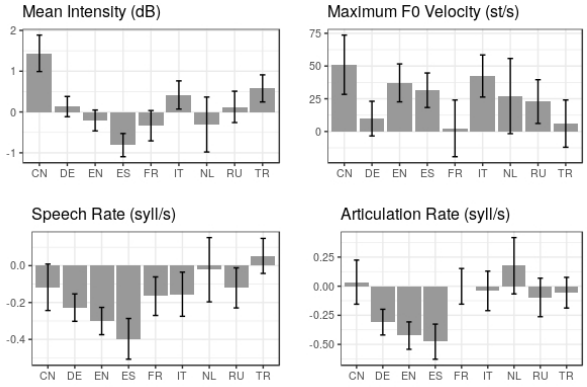

Figure 2. *Mean values with 95% CI of the features with language as significant factor.*

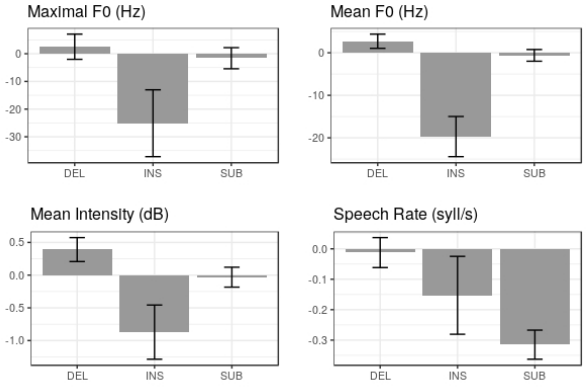

Figure 3. *Mean values with 95% CI of the features with introduced error type as significant factor.*

Table 4 presents the analysis of variance table of type III with Satterthwaite approximation with the language and the type as factors. Only the dependent variables where the language or the error type were significant factors are presented. The statistical analysis further revealed significant interactions between language and the type of the introduced error ($p<0.001$) for all measured dependent variables.

5. Discussion

Firstly, there are differences how speakers reacted in the case of introduced errors. The suggestion to the speakers to use the keyword to recover from miscommunication was differently accepted, ranging from 55.3% cases in Italian down to only 9.6% for Turkish. More specific, in 61.9% of cases of not using the correction keyword, significant differences were observed among the Dutch, Russian and Turkish speakers.

Regarding the response types, in case of full repetitions, Turkish speakers used to repeat the statement sentence in correction turn more often. German and Spanish speakers had above average frequency of responses with different content, while Turkish was significantly below the average. In partial responses, Italian and Spanish tend to have lower than average frequency than other languages. For mixed correction responses, German and Spanish had notably lower frequency.

The Linear Mixed Model analysis showed that there is a significant influence of the fixed factors language and type of the introduced error, as well as their interaction on the delta features related to hyperarticulated speech. The observed means with the 95% CI are presented in Figure 2 and 3.

Least-squares means analysis of the model showed that language was a significant factor for mean intensity where Spanish speakers had lower intensity in corrections than the others. Maximum pitch velocity was significantly higher for Italian ($p<0.001$), followed by English ($p<0.001$), Spanish ($p<0.01$) and French ($p<0.01$) speakers.

The speech rate was significantly different, the Spanish ($p<0.001$) were talking slower in corrections than all others, followed by Italian ($p<0.05$), French ($p<0.01$), English ($p<0.05$) and German ($p<0.05$).

Similar behavior was observed also for articulation rate, which was lower for: Spanish ($p<0.001$), English ($p<0.05$), German ($p<0.01$), and higher for the Dutch ($p<0.05$) and Mandarin Chinese ($p<0.05$). Average syllable duration was (in order): shorter for Dutch ($p<0.05$) and longer for German ($p<0.05$), Spanish ($p<0.05$) and English ($p<0.01$).

Regarding the introduced error type (Figure 3), the maximal *f0* was lower in the case of insertions ($p<0.05$), mean *f0* was lower in insertions ($p<0.001$) and higher in deletions ($p<0.01$). Similarly for mean intensity, it was lower for insertions ($p<0.05$) and higher for deletions ($p<0.01$), while the speech rate was significantly slower in the case of substitutions error ($p<0.001$).

In general the speakers raised their voice (pitch and intensity) in the case of reacting on the request to repeat the last utterance (deletions) but they did the opposite in the case of insertions, mostly confused by the sudden and unexpected system confirmation. The speech rate (including the pauses and hesitations) was slower in misrecognition clarifications (substitutions). While for the language and error type there are clear findings in significant differences, in the case of their interactions things are far more complex what makes the problem of defining optimal recovery strategies even more difficult. All those observations should be taken with care in the development process of a multilingual spoken dialogue system.

6. Conclusions and Future Work

From the results it could be clearly seen that there are distinctive prosodic features across languages associated with hyperarticulated speech in correction dialogue acts. This is a potential issue in developing or adapting existing SDS to new languages.

Many solutions are possible for language dependent adaptation of a SDS (some presented in Section 1) to address the hyperarticulated speech. Training classifiers of prosodic features to detect hyperarticulated speech, employing special dialogue configurations for the recognition (confidence thresholds, grammars [5]) or different strategies to cope with uncertainty in corrections, acoustic model adaptation (as in [14]) or modifying feature extraction settings (frame rate and size, voice activity detection, etc).

Our future work will be focused to more elaborate analysis of significant interactions of the gender, language and error type factors expressed by prosodic features that are already certain indicators of hyperarticulated speech and to suggest methods to detect and handle such occasions.

7. Acknowledgments

The authors are thankful to Maria Paola Bissiri and Rüdiger Hoffman for their comments on an earlier version of this paper.

8. References

[1] R. Jonson, "Multilingual NLP Methods for Multilingual Dialogue Systems". *http://stp.lingfil.uu.se/~nivre/gslt/RebeccaNLP.pdf*, Dec. 2002, [retrieved on Dec 20, 2016]

[2] R. Meena, G. Skantze, and J. Gustafson, "Automatic detection of miscommunication in spoken dialogue systems," *in Proc. of SIGdial*, 2015.

[3] I. Bulyko, K. Kirchhoff, M. Ostendorf, and J. Goldberg, "Error-correction detection and response generation in a spoken dialogue system," *Speech Communication,* vol. 45, no. 3, pp. 271–288, 2005.

[4] J. Hirschberg, D. Litman, and M. Swerts, "Characterizing and predicting corrections in spoken dialogue systems," *Comput. Linguist*, vol. 32, pp. 417–438, 2006.

[5] H. Sagawa, T. Mitamura, and E. Nyberg, "Correction grammars for error handling in a speech dialog system". *HLT/NAACL*, *Boston,* 2004

[6] E. Shriberg et al, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?", *Language and Speech*, 41:439–487, 1998

[7] G.-A. Levow, "Characterizing and recognizing spoken corrections in human-computer dialogue." *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 1998.

[8] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures". *Speech Communication, 43(1-2):155–175,* 2004

[9] S. Oviatt, "Modeling hyperarticulate speech during human-computer error resolution". *In Proceedings of the International Conference on Spoken Language Processing*, pp 797–800, 1996.

[10] M. Swerts, D.J. Litman, and J. Hirschberg, "Corrections in spoken dialogue systems". *In INTERSPEECH* (pp. 615-618), 2000, October

[11] L. Bell and J. Gustafson, "Repetition and its phonetic realizations: Investigating a swedish database of spontaneous computer directed speech". *In Proceedings of ICPhS*, pages 1221–1224, 1999.

[12] E. Shriberg, E. Wade, and P. Price, "Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction." *In Proceedings of the DARPA Speech and Natural Language Workshop*, pages 49–54, 1992.

[13] A. J. Stent, M. K. Huffman, and S. E. Brennan, "Adapting speaking after evidence of misrecognition: Local and global hyperarticulation," *Speech Communication,* vol. 50, (pp. 163-178), 2008.

[14] H. Soltau, A. Waibel, "Specialized acoustic models for hyperarticulated speech." *In Acoustics, Speech, and Signal Processing,* 2000. (Vol. 3, pp. 1779-1782).

[15] K. Vertanen, "Speech and speech recognition during dictation corrections". In *Interspeech,* 2006, paper 1094-Wed2CaP.10

[16] B. Andreeva, G. Demenko, B. Möbius, F. Zimmerer, J. Jügler and M. Oleskowicz-Popiel. "Differences of pitch profiles in Germanic and slavic languages". *In INTERSPEECH* (pp. 1307-1311), September, 2014

[17] I. Mennen, F. Schaeffler, G. Docherty, "Cross-language differences in fundamental frequency range: A comparison of English and German" *The Journal of the Acoustical Society of America*. 2012 Mar;131(3):2249-60.

[18] E. P. Altenberg and C. T. Ferrand, "Fundamental frequency in monolingual English, bilingual English-Russian, and bilingual English-Cantonese young adult women." *J. Voice* 20(1), 89–96. 2006

[19] P. Keating and G. Kuo. "Comparison of speaking fundamental frequency in English and Mandarin." *UCLA Work. Papers Phonetics* 108, 164–187, 2010

[20] I. Wendler, A. Jatho, I. Kraljevski, M. Wenzel, "Nutzerzentrierter Entwurf von Multimodalen Bedienkonzepten", *28. Konferenz Elektronische Sprachsignalverarbeitung 2017,* Universität des Saarlandes, Saarbrücken, 15.–17. März 2017

[21] P. Boersma. "Praat, a system for doing phonetics by computer." *Glot International* 5:9/10, 341-345, 2001

[22] N.H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically". *Behavior research methods*, 41 (2), 385 – 390, 2009

[23] Y. Xu, "ProsodyPro – A Tool for Large-scale Systematic Prosody Analysis". *In Proceedings of Tools and Resources for the Analysis of Speech Prosody* (TRASP 2013), Aix-en-Provence, France. 7-10, 2013

[24] D. Bates, M. Maechler, B. Bolker, S. Walker, "Fitting Linear Mixed-Effects Models Using lme4". *Journal of Statistical Software*, 67(1), 1-48, 2015