Signalizing and Predicting Turn-Taking in Multilingual Contexts: Using Data from Transcribed International Spoken Journalistic Texts in Human-Robot Interaction

Christina Alexandris
National University of Athens
calexandris@gs.uoa.gr

Abstract
Data from transcribed spoken journalistic texts from international news networks is employed in the signalization and prediction of turn-taking in Human-Computer Interaction and Human-Robot Interaction in multilingual contexts, taking into account the verbal and non-verbal behavior of international speakers.

The International Public as a User Group: System Initiative and Controlled Turn-Taking
The International Public as a User Group is characterized by diverse linguistic and diverse cultural backgrounds influencing User behavior and interaction, especially in applications intended for the general public. Language and culture-specific parameters may concern the question of the necessity of Machine Translation (MT) and, also in the case of monolingual applications intended for international Users, the question of what type of language and behavior is expected during interaction. Behavior of international Users is dependent on the general social and cultural context in which the interaction between Users and Robots takes place (Rehm 2013), User expectations from the System (Wiegers 2005), as well as User behavior towards the System reflecting the degree of familiarity of the society with HCI Systems, including Robots. The latter may be related to a User or System Initiative (Jurafsky and Martin, 2008) approach, where the System or Robot may “guide” less experienced Users (Jurafsky and Martin, 2008) (Lewis 2009). In the case of Human-Robot Interaction (HRI) targeted to an international public, for example, in applications involving interactions in the fields of health care, construction, transportation or international travel, culture-specific factors may also define turn taking. This includes differences in non-verbal behavior (Novick and Gris 2013) and the question of whether all participants are defined by symmetric or asymmetric roles, whether turn-taking in two-party or multiparty interaction takes place according to a specific “protocol” monitored by a controlled System-Initiative approach or whether turn-taking is spontaneous.

Controlled turn-taking is proposed for multilingual HCI applications and involves a System-initiative and System monitoring approaches. For example, in multilingual routine business meetings with an agenda (Alexandris 2013), the System creates a form of interaction similar to that of a Directed Dialog using existing MT tools for written texts (and spoken texts, such as Interlinguas generated simultaneously with text messages chosen by the User). In this case, where the interaction involves Skype meetings with (or without) translated subtitles and a standard, controlled conversation, interaction is controlled by the template-agenda of the System-mediator (Alexandris 2013), designed in a similar manner to that of templates used in Dialog Systems for registering and controlling User input (Nottas et al. 2007). The template-agenda, containing topics covered during interaction, allows the System to automatically intervene, asking additional questions, until the issue in question is addressed. The flow of the conversation, including topics covered, can be checked by the System and User, with the System’s intervening messages appearing on the interface.

However, if controlled turn-taking and moderated questions are not possible in applications concerning the interaction of two or more international speakers, different strategies should be employed. Data from journalistic texts indicates that non-verbal and paralinguistic features related to turn-taking is often unpredictable, culture-specific and dependent on a Speaker’s individual characteristics.

The present data and observations are based on transcriptions from two-party or multiple party discussions of spoken journalistic texts, which are also compared to older
Data and Observations from Transcribed Data of Spoken Journalistic Texts

The data was collected from graduate courses for journalists, in particular, the Program M.A in Quality Journalism and Digital Technologies, Danube University at Krems (Austria), Athena- Research and Innovation Center in Information, Communication and Knowledge Technologies, Athens, Institution of Promotion of Journalism Ath.Vas. Botsi, Athens. The collected data concerned assignments in transcription and linguistic analysis from two groups of 18-20 professional journalists for the academic years 2012-2013 and 2013-2014. Each journalist worked with 4 interviews/panel discussions and performed a rigorous transcription of 15-20 minutes of each discussion, two in English (or any other natural language, for example, German) and two in Greek. The processed files included transcriptions from the following news broadcast networks: BBC World News, CNN, AL Jazeera, Russia Today (RT) ("Crosstalk"), ZDF and ARD (Germany), available transcriptions from the German Parliament and Greek TV channels (ERT-NERIT (public television), the Parliament Channel, Mega TV, SKAI TV, ALPHA). Speakers-Participants (S-Ps) were either non-native speakers of English, or native speakers of (British or American) English, German (Standard or Dialect) and Greek. Non-native S-Ps included countries as varied as India, Russia, Eastern Europe, the Middle East (Arab-speaking countries and Israel), Iran, Pakistan, P.R. China, Nigeria and Kenya.

Journalists were assigned the task to transcribe two-party or multiparty discussions from national and international news networks, to record, among other features, turn-taking, time assigned to each participant, interruptions, gestures and other paralinguistic features and to provide an outline of the discourse structure of the two-party or multiparty discussions.

The main objective was not to signalize turn-taking but was focused on S-Ps behavior and socio-linguistic features (role of the S-Ps, success of the dialog-interaction within the recorded discourse structure): instances of turn-taking were registered to support conclusions concerning the extent and nature of interruptions in an interview or panel discussion. Transcriptions were compared to the actual video files.

Limitations to be considered in the above-described process involve the signalization of turn-taking as (a) an additional outcome (not as a main target) of analysis performed (b) by professional journalists and not by linguists in light of a specific turn-taking model, (c) a process often concerning non-native speakers transcribing the speech of non-native speakers of English where typical verbal and/or prosodic cues in English did not always apply for the international S-Ps and (d) without the use of a specific (or same) transcription or annotation tool, outside a laboratory environment.

However, the empirical data is partially compatible (i) with the turn-taking model of Sacks et al. 1974 as well as Wilson and Wilson 2005, where each participant (un-)consciously keeps track of whose option it is to speak and the listener/speaker selects self (ii) and in some cases comparable to turn-taking in “push-to-talk conversations” (Taboada 2006). For the S-Ps, turn-taking and interruptions were all determined by three basic factors, namely (1) the presence or absence of predefined “protocol”, (2) S-Ps’ individual style and personality and (3) language-specific parameters and general socio-cultural norms. The third factor plays an important role in non-verbal and paralinguistic features such as gaze, gestures and prosody.

We note that in some transcriptions, subtle non-verbal and paralinguistic features related to turn-taking were not recorded by many observers, in others some Speakers did not even gaze towards the person they were addressing and/or retained the same facial expression and/or barely moved their hands or changed posture (observation 1).

Additionally, it is noted that typical paralinguistic indicators for turn-taking involving gaze and gesture were absent in some multimedia files, for instance in a cases where the speaker (for, example, a correspondent) is not visible (observation 2).

In regard to prosodic features, more subtle features in languages such as Standard German were not always registered by non-native speakers of German (observation 3).

The processing of the transcribed data included the following features occurring with monitored (controlled by moderator) or spontaneous turn-taking:

- Turn-taking signalized by gaze (Level 1.a), by facial expression (Level 1.b) (pursing of lips, raising eyebrows etc.), by movement of head (Level 1.c) (for example, a nod), by gesture (Level 1.d) (raising a hand, waving hands etc.), by changes in posture (Level 1.e)
- signalized by a combination of some or all of the above
- signalized by change in prosody only (Level 2.a) (Sentences uttered with language-specific or language-independent prosodic models of questions or sentence boundaries, triggering turn-taking signals of other S-Ps)
- signalized by an exclamatory only (Level 2.b)
- signalized by an exclamatory word or phrase only (linguistic features) (Level 3)
- signalized by a combination of prosodic and linguistic features (Level 2 and Level 3) (prosodic emphasis on specific words, hesitation markers)
- signalized by a combination of any of the above paralinguistic, prosodic and linguistic features (all Levels)

However, it should be noted that only about 20% of the journalists marked features occurring with spontaneous turn-taking in Level 1.a, 1.b and Level 2.a, in contrary,
gestures (Level 1.d) and exclamations (Level 2.b and Level 3 features) were more commonly registered.

We note that the evaluation of detailed quantitative data requires using the same transcription tool and a native speaker control-group.

**Signalizing Linguistic Features and Turn-taking Attempts: Employing Practices from Dialog Systems and HRI**

In contrast to non-verbal and paralinguistic features related to observations 1, 2 and 3, it was noted that in many cases certain types of words, expressions and exclamations uttered by S-Ps were occurring before or during turn-taking and interruptions, demonstrating a broader usage across various international participant groups, regardless of their native tongue or fluency in English. In particular, exclamations such as “No, no, no”, “Ja, ja” (German) some discourse particles (“Please”, “OK”, “Ma” (“But”- Greek), “Aber” (German)) and deictic expressions, (“you”, “they”) were commonly occurring before or during S-Ps turn-taking and interruptions. These utterances may be referred to as “Level 3.a features” and are sublanguage-independent. In addition, particular types of words and/or utterances triggering turn-taking attempts of other S-Ps include sublanguage-specific expressions in politics and economy or proper names triggering controversial reaction and non-neutral “marked” words containing connotative features (for instance, “bad” or “mistake”), often accompanied by prosodic emphasis or change of tone. These words and/or expressions may be referred to as “Level 3.b features” and are mostly sublanguage-specific.

Level 3.a features, which may constitute a database, may be recognized and processed by the Robot’s Speech Processing Component in a similar manner to the sublanguage-specific utterances recognized and processed in Spoken Dialog Systems in cases in which the User interrupts the System to correct spoken input or to change options with keywords and natural commands within a Directed Dialog context (Lewis 2009). We note that, in a similar manner to natural commands, Level 3.a features can be subjected to MT.

Topic-shift signalized by uttered words and expressions is processed within a network-framework, namely the framework of the relation of words-hyperlinks of a Wikipedia (Jokinen and Wilcock 2013). A similar network-framework strategy may be used by the Robot’s Speech Processing Component for the second category of Level 3.b features, as “turn-taking triggers”, many of which are sublanguage-independent “connotatively marked” words and expressions based on the flouting of the Gricean Cooperative Principle, especially in regard to the violation of the Maxims of Quality and Quantity (Grice 1989) (Hatim, 1997) (Alexandris 2010). These sets of expressions may be grouped into a finite set based on word stems or suffix type. Recognition on a word-stem or a suffix basis involves the detection of word-classes such as adjectives and adverbials or types of verbs, containing specific semantic features such as malignant/benign action or emotional/ethical gravity (some of them are sublanguage-independent), among others, accessible with Wordnets and/or Selectional Restrictions (Alexandris 2010). The set of Level 3.b features also includes a sublanguage-specific set of expressions triggering controversial reaction, according to application domain and type. These features may also be processed by an MT module.

The transcriptions of the spoken journalistic texts showed that, at least for English spoken both by native and non-native international speakers, for German and for Greek, Level 3.a and 3.b types of words, expressions and exclamations uttered by S-Ps occurring before or during spontaneous turn-taking and interruptions constitute safer indicators for turn-taking than non-verbal and paralinguistic features, as noted in observations 1- 3. Level 3.a and Level 3.b features labeled as “turn-taking triggers” may therefore be related to higher confidence levels in predicting or signalizing turn-taking.

The confidence score is highest if Level 3.a and 3.b features occur in combination with non-verbal and paralinguistic features such as gestures or prosodic emphasis (Levels 1 and/or 2).

**Confidence Levels and the Role of the Speech Processing Component for the Signalization and Prediction of Turn-taking in HRI**

The Speech Recognition Module of the Robot’s Speech Processing Component can process the finite set of Level 3.a and Level 3.b features marked as “turn-taking triggers”, alerting the Robot for the possibility of additional non-verbal-paralinguistic and prosodic types of indicators for turn-taking, either co-occurring or following the “turn-taking trigger”. Thus, in applications concerning the international public the processing of turn-taking signals is proposed to include combining features from three levels: the Paralinguistic Level 1, the Prosodic Level 2 and the Linguistic Level 3. Both Levels 2 and 3 are managed by the Speech Processing Component of the Robot.

Level 3.a and 3.b word categories as “turn-taking triggers” may coincide with prosodic emphasis or other prosodic phenomena in the speech signal (wav file) and/or with any non-verbal features, such as gestures signalized in the respective temporal points. In other words, the partially sublanguage-specific (I) Level 3.a and 3.b “turn-taking triggers” of the Linguistic Level 3 (II) detected by the Speech Processing Component (III) are proposed to play the basic role of (a) predicting and (b) signalizing turn-taking attempts, providing the highest confidence level for turn-taking. Thus, a turn-taking attempt is alerted on the
Linguistic Level -possibly with a partial recognition of words (Baumann and Schlangen 2012), and reinforced by any subsequent signal and feature detected on the Paralinguistic Level and Prosodic Level. Therefore, any subsequent signal and feature detected on the Prosodic and/or Paralinguistic Levels has a higher confidence score if it co-occurs with or follows a “turn-taking trigger” detected on the Linguistic Level. The confidence scores of the Prosodic and the Paralinguistic Levels are dependent on the parameters of the natural language involved and the S-Ps native tongue.

It should be additionally stressed that the linguistic features in Level 3 can be processed by an MT module, if necessary. Machine Translation may involve an Interlingua and/or a framework such as Universal Words (Uchida et al., 2005), especially for Level 3b words. Signalizing and predicting turn-taking in HRI concerning multilingual contexts, therefore, involves the processing of (I) partially sublanguage-specific (II) linguistic (III) “turn-taking triggers” by the Robot’s Speech Processing Component, in addition to non-verbal and paralinguistic features. The proposed strategy can be summarized as following:


Conclusions and Further Research

Data from transcriptions of spoken journalistic texts involving international Speakers provides an insight on the variety of turn-taking signals and linguistic features observed to trigger turn-taking. Linguistic features, labeled as “Level 3” features and “turn-taking triggers” are proposed to be used as safer indicators signalizing or predicting turn-taking for an international user group. Recognized and processed by the Robot’s Speech Recognition Module “Level 3” features are proposed to be related a higher confidence level as “turn-taking triggers” in applications for the International Public and may also be subjected to Machine Translation. Full implementation in a multilingual/international HRI context may provide concrete results of the proposed strategy’s success level and an insight to resolving any further issues.

References


