An Integrated Representation of Linguistic and Social Functions of Code-Switching

Silvana Hartmann, Monojit Choudhury, Kalika Bali
Microsoft Research Lab India
Bangalore, India
nlp@silvanahartmann.de,{monojitc,kalikab}@microsoft.com

Abstract

We present an integrated representation of code-switching (CS) functions, i.e., a representation that includes various CS phenomena (intra-/inter-sentential) and modalities (written/spoken), and aims to derive CS functions from local and global properties of the code-switched discourse. By applying it to several English/Hindi CS datasets, we show that our model contributes i) to the standardization and re-use of CS data collections by creating a resource footprint, and ii) to the study of CS functions by creating a systematic description and hierarchy of reported functions together with the (local and social) properties that may affect them. At the same time, the model provides a flexible framework to add emerging functions, supporting theoretical studies as well as the automatic detection of CS functions.

Keywords: Code-Switching, Multilingual, Bilingual, Pragmatics, Discourse

1. Introduction

Code-Switching (CS), or alternating between two or more languages in a single conversation, is a marked feature of multilingual communities. Linguists have studied this phenomenon in great detail and recently, with the rise of social media, the processing and generating of CS content has gained attention in the NLP community as well. The amount and type of code-switching depends on a number of structural, functional and social factors (Begum et al., 2016). Linguistic studies in the past have focused on two aspects that may be summarized as the "how" and the "why" of code-switching (Poplack, 2015). The "how" aims to explain the grammatical principles that underlie code-switching performance; the "why" aims to explain the function of code-switching in discourse, defining social, pragmatic, and discourse functions of code-switching, for instance *addressee specification*, *emphasis*, or *marking quotations*. These studies are mostly based on limited recordings of conversations, concentrating on a small subset of functions, either linguistic, based on local properties of the discourse, or social and other global aspects. There are no large empirical studies on the interaction of linguistic and social aspects of CS functions. Further, there is no unified theory on the different CS functions and how they are expressed in discourse. Previous theoretical work identifies lists of functions (see e.g., Abdul-Zahra (2010)), and recent work in NLP picks specific functions and aims to identify them in large collections of CS texts (Begum et al., 2016; Rudra et al., 2016). Thus, such studies use different levels of analysis, and often confuse observations (such as observing code-switched *reiteration* or *translation*) with CS functions (*reiteration* is often used for *emphasis*), as in the following example from Begum et al. (2016) that uses the translation of an utterance for emphasis:

> dimaag mein bhoosa bhara hai [gloss:*up in their heads with fodder*]. up in their heads with fodder.

The task of *defining* a comprehensive set of functions is very difficult, if not elusive. The same switching phenomenon, in our example *translation*, is used for emphasis in some communities and for weakening request in others (Poplack, 1988). In such cases it is crucial to distinguish between surface form and (community-specific) function of a switching event. In this work, we present a new integrated framework for representing code-switching functions that builds the foundations for studying the functions of code-switching empirically at a large scale, particularly the interaction between the linguistic and social CS functions. Instead of defining a static set of functions, our framework provides the tools to define functions according to community-specific analyses. We define a flexible set of code-switching functions by first outlining a number of properties that characterize code-switched discourse, and then using these properties to characterize different switching phenomena and associate them to their functions. The proportion of properties and phenomena in a dataset can also be used to create a footprint and compare different CS styles, for instance across language pairs or communities. We apply our framework to several English/Hindi code-switched datasets, i) creating footprints of CS corpora for comparison and standardization (adding a wider context to the statistics introduced by Guzmán et al. (2017)) and ii) systematically describing and deriving CS functions for code-switched conversations.

2. Related Work

Early work on the functions of CS distinguishes between high-level *social functions* and lower-level linguistic functions that depend on the surface form of a specific code-switched utterance. Gumperz (1982) distinguishes between situational and metaphorical/conversational switching. Situational switching operates on the social level and is performed to accommodate different discourse partners, metaphorical switching happens on the utterance level. Later work stresses the discourse-structuring function of CS (Auer, 2013): specific functions are considered to emerge dynamically from the discourse. These functions are strongly tied to a specific code-switched utterance and its context. Following this theory, creating exhaustive lists of CS functions is problematic, since new functions can emerge any time.

Previous work identified a large number of functions of CS, ranging from utterance-level functions (marking switches from narrative to evaluative language) to conversation-level functions (marking an addressee in a conversation), or even higher-level social functions (showing an affilliation to a certain community). Gumperz (1982) lists six functions: *quotation*, *addressee specification*, *interjection*, *reiteration*, *message qualification*, and the semantic function of *personalization versus objectivization*. Crystal (1987) reports the functions *lack of competence*, *solidarity with a group*, and *communicating a certain attitude*. Reyes (2004) lists functions such as *reported speech*, *imitation quotation*, *turn accommodation*, *topic shift*, *situation switch* (e.g., from academic to non-academic), *insistence non-command* (by reiteration), *emphasis on command*, *clarification/persuasion*, *person specification*, *question shift*, *discourse marker*, and *other*. Abdul-Zahra (2010) adds *quotation* and *addressee qualification* to Crystal (1987)'s list.

Recently, large empirical studies have started to use Twitter as a data source to explore the functions of code-switching: Begum et al. (2016) analyze a set of English/Hindi Tweets and distinguish between several structural and pragmatic functions. Rudra et al. (2016) discover tendencies of bilingual speakers to express negative sentiment in (what is perceived as) their primary language. Rijhwani et al. (2017) is one of the first works to explicitly explore diverse CS communities on Twitter: they study different CS communities based on different language pairs and different localities. Due to the broadcasting nature of Twitter most of this work is focused on linguistic functions, ignoring functions that require global knowledge of the conversation setting.

Quantitative studies have proposed metrics on the distribution of languages in CS corpora. Gambäck and Das (2016) introduce the Code-Mixing Index, and Guzmán et al. (2017) present a set of six metrics to quantify CS statistics, including the equality and burstiness of the language distribution, language entropy, and switching probability. These metrics consider the distribution of the different languages in the code-switched discourse. These metrics can thus, provide detailed statistics on the granularity of code-switching as a means to systematically study CS functions.

Mapping the above-mentioned function lists to a unified model is very difficult for several reasons: i) they cover social and local functions to different degrees, ii) small terminological differences are not clear (the terms *reiteration*, *repetition*, and *translation* are used for similar phenomena), and iii) they show varying degrees of granularity (e.g., whether or not to distinguish between *quotation*, *imitation quotation*, and *reported speech*). Thus, confusing properties of a CS utterance with its function: an observed *translation* can have the function of *emphasis* or *de-emphasis* depending on the community. In general, the relation between surface forms and potential functions available for a surface form is not well-defined. In addition, the lists are often incomplete and tailored to community-specific observations. A comprehensive model should provide descriptive capabilities for representing conflicting functions across communities. We aim to create a framework that represents the properties of CS and uses them to define CS functions dynamically, in an attempt to resolve these problems.

3. New Representation

In this section we introduce our new representation of CS functions that aims to fill the research gap discussed above. Our goal is to organize the functions observed in the literature into a systematic hierarchy that considers the individual functions and the functional levels on which they apply. This new representation of code-switching functions lays the foundation for the study of the interaction between different functional levels of code-switching in the future.

Requirements. Our representation should be a) the basis for analyzing the relationship between social and local CS functions, b) the means to distinguish surface properties of CS phenomena from their functions, c) the means to define functions based on their properties, and d) extensible by other functions and properties, for instance prosody and pauses in speech. We now define code-switching and the analysis dimensions used in our model.

Definition of code-switching. We define code-switching, short CS, as the use of at least two languages by the same speaker that is fluent in these languages within a single conversation. We call the alternating sequences in the different languages **CS segments**. As a prerequisite to defining code-switching functions, we define the unit of analysis for CS functions: a **CS pair** $p({s_1}^a, {s_2}^b)$ consists of two **CS segments** s_1 and s_2 where s_1 is in language l_a and s_2 is in language l_b. The **switch point** marks the boundary between s_1 and s_2. The segments s_1 and s_2 can be of variable length (number of tokens); s_2 extends to the next switch point or the end of the utterance.

3.1. Analysis Dimensions

We define five analysis dimensions relevant to the identification of CS functions, four global dimensions (Granularity, Modality, Discourse, and Social dimension), and the **Local** dimension that considers properties of the specific CS pair. **Granularity** considers properties on the level of CS segments. Its properties are represented by basic CS statistics and the metrics from Guzmán et al. (2017) and Gambäck and Das (2016). **Modality** defines different modes of recording CS language, such as written, spoken, multi-modal (including visual information). **Discourse** relates to the type of communication that is studied, e.g., monolog, dialog, multi-party dialog or broadcasting situations. Certain CS functions such as turn-taking have only been observed for dialogs. The **Social** dimension includes the social aspects of CS discussed earlier, representing information on the relationship and hierarchy between the conversation partners, including familiarity and politeness, but also whether the language community uses certain functions, e.g., repetition for emphasis. In the following paragraphs we provide more detail on the properties we measure for each analysis dimension.

Modality dimension The modality dimension defines different modes of producing code-switched language, such as written, speech-only, and speech with visual information, see the first line in Fig. 1. The categories in this dimension differ in the way an utterance is transmitted and in the presence or absence of linguistic and meta-linguistic information relevant to communication such as prosody, facial

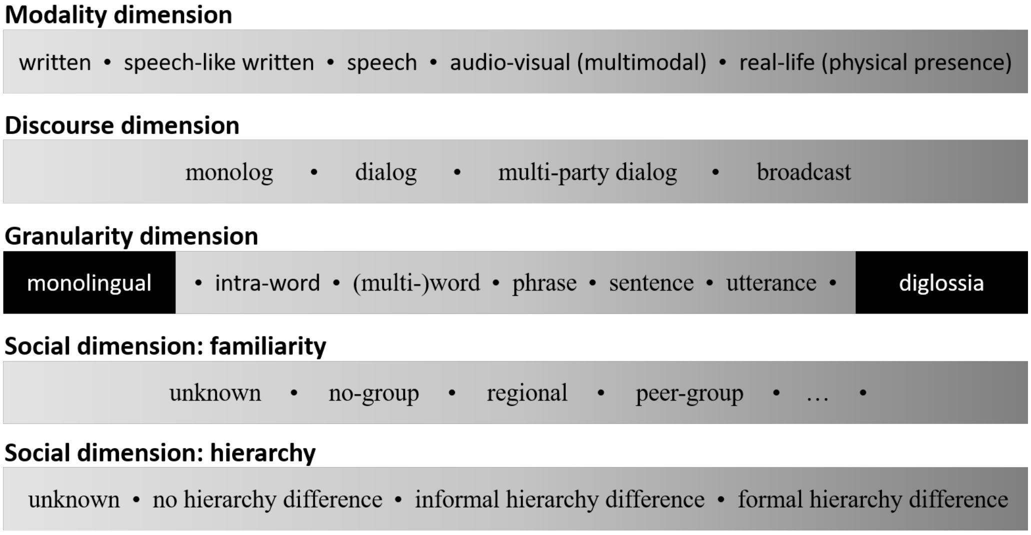
Modality dimension
written • speech-like written • speech • audio-visual (multimodal) • real-life (physical presence)
Discourse dimension
monolog • dialog • multi-party dialog • broadcast
Granularity dimension
monolingual
• intra-word • (multi-)word • phrase • sentence • utterance •
diglossia
Social dimension: familiarity
unknown • no-group • regional • peer-group • … •
Social dimension: hierarchy
unknown • no hierarchy difference • informal hierarchy difference • formal hierarchy difference

Figure 1: Global analysis dimensions for code-switched discourse.

expression, posture, gaze, gestures, etc. The modality of CS datasets depends on the recording medium. Previous work mostly uses speech recordings (and their transcripts), as well as written texts. It is thus focused on their inherent properties, such as emphasis and hesitations to structure utterances in spoken conversation, and punctuation to structure them in written text. Socio-linguistic studies of CS however frequently refer to discourse-structuring functions of CS, and other ways to express the same functions, for instance gestures when addressing a speaker.[1] We propose the following categories for the modality dimension, of which at least written or audio are typically present: *written, informal speech-like written, speech*, and *audio-visual*.

Discourse dimension The discourse dimension relates to the type of communication that is studied, e.g., whether we are studying at a monolog, dialog, multi-party-dialog or a broadcasting situation (which can be assumed for some Twitter messages), see the second line in Fig. 1. Certain functions that have been observed for CS contributions, such as turn-taking, are only applicable to dialog. The discourse dimension is closely related to the modality dimension: turn-taking is more important in immediate dialog, in particular in speech settings (be it present or via phone call), but may also occur in computer-mediated communication. We propose the following categories for the discourse dimension: *monolog, dialog, multi-party dialog, broadcast*.

Granularity dimension The granularity dimension describes the code-switched segments based on their size and syntactic type. It is illustrated in the third line of Fig. 1. The segment sizes can range from parts of words, single words and multi-word-expressions, via phrases and sentences to full utterances (up to sequences of sentences). One of the most frequently studied phenomena in CS research is intra-sentential switching, which maps to the segments up to the phrase level.

The granularity dimension helps to demarcate the boundaries of switching: we do not consider the use of loan-words by a predominantly monolingual speaker part of CS. The other end of the scale (as shown in Fig. 1) shows diglossia, where one speaker is capable of speaking two languages, but does not mix them in the same conversation, also outside of our definition of CS. Therefore, they are shown with a black background in Fig. 1.

Segments of different type will, of course, occur in the same conversation and in the language of the same speaker or community. The distribution over segment types, or dominant segment size can be used to characterize a specific CS dataset. Syntactic analysis, identifying sentences and phrases, might not be available for the studied language pair. In this case, the number of tokens in a segment can be used as a proxy for these categories. Previous work on CS in Twitter messages exploits quantitative statistics on segment size, studying CS with segment sizes of at least 3 words (Rudra et al., 2016). We propose the following list of categories for the granularity dimension: *sub-word, single word, phrase, sentence, utterance*.

Social dimension Code-switching speakers are known to modulate their use of switching to specific situations, depending on the participants involved. This information is captured in the social dimension. For a particular speaker, different social groups and situations may evoke different degrees of CS and affect their use of CS.

The social dimension includes aspects relevant to the meta-data associated with the conversation partner(s) or the addressed audience, such as formality and politeness, familiarity, and the existence of certain group-internal (cultural) conventions of expressions. Details of the social dimension are typically recorded when gathering code-switched conversations, but they are mostly unknown when studying large Twitter datasets. As a consequence, most work on CS on Twitter did not consider the social dimension.

We propose the following two sub-categories for the social dimension, each with a list of tentative values that are subject

[1] In how far emoticons and other frequent phenomena in written computer-mediated communication, such as expressive lengthening, emulate other modalities is an interesting research question in this context.

to further development: a) familiarity (*unknown*, *no-group* (no common group membership), *regional* (group defined by geographic properties), *peer-group* (various social factors, with further specification of sub-groups)) and b) hierarchy levels (*unknown, no hierarchy difference, informal hierarchy difference, formal hierarchy difference*).

3.2. Local Dimension

Since they play a large role in our representation model, we introduce local properties in detail. We distinguish between *simple atomic properties* on the CS segment level and *compound properties* that build upon atomic properties. Both are relevant to identifying local CS functions. Relevant properties include syntactic, semantic, and pragmatic properties of the CS segment:

Syntactic properties

- $syntactic(s_i) = x \in$ {sentence, clause, phrase, word, sub-word}
- $quotation(s_i) = x \in \{0,1\}$, $x = 1$ if s_i is highlighted by quotation marks[2]
- $continuation(s_i) = x \in \{0,1\}$, $x = 1$ if s_i starts a new sentence
- $tag(s_i) = x \in \{0,1\}, x = 1$ if s_i is a tag (a fixed phrase used for greeting, etc.
- $ne(s_i) = x \in \{0,1\}, x = 1$ if s_i is a named-entity

Semantic properties

- $topic(s_i) = x$ in a set of topics or in a distribution over topics
- $content(s_i) = x$ where x is a semantic representation of s_i (e.g., vector representation, logic form)

Sentiment properties

- $sentiment(s_i) = x \in$ SENT = {-1,0,+1}, where -1 stands for negative, 0 for neutral, and +1 for positive sentiment, cf. Nakov et al. (2013).

Pragmatic properties

- $speechact(s_i) = x \in SACS =$ {question, request, command,...}, cf. Searle (1969).
- $dm(s_i) = x \in \{0,1\}$, where $x = 1$ if s_i contains a discourse marker, cf. Prasad et al. (2014).

Compound properties (on the *CS pair* level) Just like s_i and s_{i+1} are combined to build a CS pair $p(s_i, s_{i+1})$, properties of s_i and s_{i+1} can be combined to build compound properties. Some of the possible combinations are relevant with respect to the analysis of CS functions, for instance discovering the function of sentiment change.

In general, we create compound properties by a) pairing arbitrary properties of s_i and s_{i+1} and b) by comparing the property value of s_i and s_{i+1} for the same property. For the $sentiment$ property, a) leads to a sentiment-pair(s_i, s_{i+1}) = $\langle x, y\rangle$ with 4^2 values for $\langle x, y\rangle$, e.g., $\langle x, y\rangle$ $\in$ SENT×SENT=$\{\langle 0,0\rangle, \langle 0,-1\rangle, \langle 0,+1\rangle, \ldots, \langle +1,+1\rangle\}$. The latter, b) can be used to define topic change: topic-change$(s_i, s_{i+1}) = x \in \{0,1\}, x = 1$ if $topic(s_i) \neq topic(s_{i+1})$.

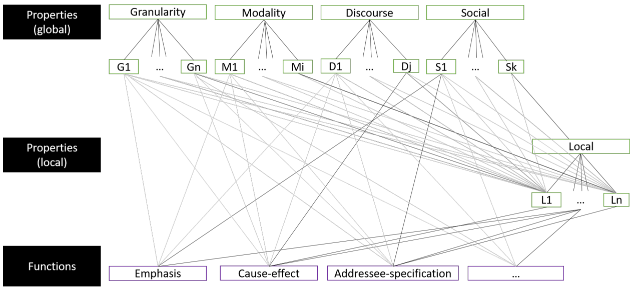

Figure 2: Dependency between CS properties and functions.

In the following, we define a number compound properties that are relevant for the definition of CS functions:

- topic-change$(s_i, s_{i+1}) = x \in \{0,1\}, x = 1$ if $topic(s_i) \neq topic(s_{i+1})$
- translation$(s_i, s_{i+1}) = x \in \{0,1\}, x = 1$ if $content(s_i) = content(s_{i+1})$
- sentiment-pair$(s_i, s_{i+1}) = \langle x, y\rangle$ with 4^2 values for $\langle x, y\rangle \in$ SENT×SENT
- speechact-pair$(s_i, s_{i+1}) = \langle x, y\rangle$ with $|SACT|^2$ values for x, e.g., $\langle x, y\rangle \in$ SACT×SACT
- discourse-rel(s_i, s_{i+1}) = x $\in$ a set of discourse relations SREL that connect s_i and s_{i+1}

These properties can be used to describe a CS phenomenon independently of its function in discourse. We use these properties to define several common CS functions below.

3.3. Deriving CS Functions from Properties

We use simple and compound properties of a CS pair to create operationalized definitions of common CS functions. We show three examples of common CS functions and property derivations here.

Example 1: the *narrative-evaluative* function that can be identified based on the discourse relation between the two CS segments and the sentiment involved, changing from neutral to positive or negative sentiment.

Example 2: *negative reinforcement*: if the sentiment-pair(s_i, s_{i+1}) consists of two sentiment expressions ⟨a,b⟩ and b = -1, this indicates negative reinforcement.

Example 3: *marking quotation*: if the second CS segment is highlighted by quotation marks (quotation(s_{i+1})=1), this indicates the use of CS to introduce a quotation or reported speech.

Example 4: the definition of Riloff et al. (2013) can be used to identify *sarcasm* in CS tweets as tweets with positive sentiment in one segment and a negative situation (determined by topic) in the other CS segment.

Example 5: the function of *topic shift* is observed when topic-change(s_i, s_{i+1})=1.

[2] In our introduction of local properties, we focus on their identification in written texts and speech transcripts. For the modalities *audio* and*visual*, quotation marks can be replaced by gestures or prosodic markers of quotation.

The integration of local and global properties in a joint representation of CS functions is a core motivation of this paper: besides local properties, CS functions are often also dependent on higher-level global properties (e.g., translation for *emphasis* is community-specific and depends on social contexts, social and discourse dimensions are relevant for *addressee specification*). These dependencies are illustrated in an abstract fashion (i.e., without relating to specific properties) in Fig. 2. The black lines mark observed dependencies between types of properties, grey lines show hypothetical dependencies that may exist between two property hierarchies or between a property and a function. One goal of this work is to provide a representation framework for observed and expected dependencies for a given CS corpus.

Example 6: a CS function that incorporates local and global properties is *addressee qualification*: to identify addressee qualification, we need some change in the switching behavior (e.g., monolingual to code switching), some deictic aspect (e.g., a greeting tag or a gesture addressing the speaker) and information on the relation between the speaker and addressee, e.g., belonging to a peer group.

Example 7: an instance of the local function *topic switch* that occurs in a certain social setting, for instance a conversation among colleagues in a professional setting, may be a marker of a *situation switch*. The *situation switch* can only be identified if the social factors of the interaction are known.

Summary: this section shows several examples on how local properties indicate local CS functions, and how local and global properties, for instance social properties, can be jointly used to identify CS functions. The local and global properties provide necessary, but not sufficient information to identify CS functions. Still, they help to characterize CS phenomena and provide prerequisites for the large-scale corpus-based analysis of CS functions. Many local properties can be determined using automatic methods and corpus meta-data capture the smaller set of global properties to a certain degree. We believe that a stronger formalization and standardization of corpus meta-data along the lines proposed in this work can support corpus-based analysis of the interaction of local and global CS properties in the future. In the next sections, we show two applications of our integrated representation of CS functions.

4. Application Scenario 1: Footprinting Corpora

To demonstrate its use, we apply our model to several English/Hindi CS datasets. The first dataset D1 is based on speech transcripts of informal conversations, the other datasets M1 to M6 are based on dialogs from six Indian movies titled *Pink*, *Kapoor and Sons*, *Neerja*, *Talvar*, *Ek Main Aur Ek Tu*, and *D-day*. These movies contain code-switched dialog in Hindi and English, whereby the majority of the tokens was labeled as Hindi for all datasets. Transcripts of the movie dialogs are available for non-commercial purposes from `https://moifightclub.com/category/scripts`. We performed automatic language labeling and annotated all datasets with the properties introduced above, providing elaborate annotation guidelines and an extensive training for our three annotators. Automatically created language tags were corrected as needed during the manual annotation of local properties. We plan to publish the annotations and datasets subject to permissions. We present two applications of our representation: footprinting CS discourse according to properties and functions in this section, and supporting automatic functions derivations in Section 5.

Footprinting CS discourse We can use the property values defined above to create *footprints* of a specific code-switched dataset. This is done by accumulating statistics for the different analysis dimensions as shown in Table 2. Fine-grained Granularity properties are provided by the CS metrics from Guzmán et al. (2017) and Gambäck and Das (2016). To save space, this is shown exemplary for the corpora D1 and M1 in Table 1. We refer to the original publications for detailed explanations of the different metrics.

Collecting the property values leads to a feature vector that can be compared across datasets to categorize and compare different manifestations of CS. Table 2 contrasts a subset of the global and local properties for D1 and M1 to M6. We were able to access D1 as speech transcript and speech recording, while the movie datasets are based on written transcript, but could in theory be accessed in their original audio-visual format. Conversations in D1 are generally based on two speakers, but also include a single monolog, while the movie datasets contain multi-party dialog with up to 13 speakers. There are no explicit hierarchy levels in D1, and the speakers do not know each other, while social properties are varied in the movie dialogs. The local properties in the lower part of the table are split up by the language of the CS segments, for Compound properties, the language column corresponds to the language of the second CS segment s_2, contrasting the two switching directions. The information in Table 2 goes beyond the detailed analysis of language distribution in Table 1, that shows similar Span Entropy, but different CMI, M-Metric, Burstiness, and Language Entropy for the two corpora D1 and M1. Relevant for studying CS functions is the more even distribution of languages in M1 (indicating more instances of switching and less borrowing) and its larger variety of social properties (e.g., speakers with different relationship and hierarchy levels). To contrast CS for different social aspects, statistics on interactions between specific pairs of speakers can be analyzed as sub-corpora of M1, as proposed in (Pratapa and Choudhury, 2017). The local properties also show a larger proportion of positive and negative sentiment, sentiment change, and discourse relations in M1 compared to D1, which indicates that it is a more promising source for studying CS functions related to these properties. One exception are Hindi discourse relations that seem to be more prevalent in D1 compared to the movie datasets.

The other movie datasets, M2 to M6 are fairly similar to M1. The proportion of word-level switches is lower in all movie datasets compared to D1. In contrast to previous work on Twitter datasets (Rudra et al., 2016), there is no strong prevalence for expressing negative sentiment in Hindi or to switch to Hindi for expressing negative sentiment, but a slight tendency to express positive sentiment in English for M2 and M5. CS at the sentence level and cross-speaker switches are more prevalent when the second CS segment s_2

is Hindi. Tag-switching on the other hand is more prevalent for switches to English in all datasets, which supports the notion that Hindi is the main language in the movies. The use of English discourse relations at the beginning of s_2 is more prevalent in the movie datasets compared to the speech dataset D1.

We observe that many property combinations associated with CS functions are fairly rare in our datasets. An example is translation for emphasis. This function has been widely reported for English/Hindi code-switching, but there are only 13 instances even in the largest dataset D1. Finding evidence of CS functions in corpus data is difficult, which highlights the need for large corpus-based analyses.

Source	Metric	D1	M1
Guzmán et al. (2017)	M-Metric	0.275	0.772
	I-Metric	0.153	0.222
	Burstiness	0.293	0.141
	Memory	-0.174	-0.076
	Language Entropy	0.538	0.05
	Span Entropy	3.527	3.245
Gambäck and Das (2016)	CMI Index (C_c)	45.51	71.87
	% EN	12.23	27.44
	% HI	67.20	58.14

Table 1: CS metrics for detailed Granularity properties.

5. Application Scenario 2: Function Derivation

For the second application, deriving functions from properties, we focus on the movie dataset M1 that is based on the movie *Pink*, a court-room drama centering around a group of three girlfriends. One of the friends is accused of attempted murder, but claims she injured the victim, a young man, in self-defense. The interactions among the three friends and dialogs with the judge and attorney in the court room show different social settings, thus providing a good background for our analysis of the interaction of local and social functions of CS.

We first show examples for deriving local functions from properties, then discuss the interaction of local and global properties and their effects on deriving CS functions.

Deriving local functions from local properties The following example from M1 shows the *negative reinforcement* function:

> [This is b***s**t]s_1 [Bakwaas hai poori ki poori]s_2 - jhooth bol rahein hain yeh ladke aur aap bhi
>
> **Gloss**: [This is nonsense]s_1 [This is all rubbish.]s_2 - These boys are lying and so are you.

The English segment s_1 and the Hindi segment s_2 both have negative sentiment (sentiment(s_2)=sentiment(s_1) $= -1$), which together with the knowledge of a social property, i.e. the tendency of the English/Hindi-speaking community to use switching to emphasize negative emotions (Rudra et al., 2016) indicates the function of *negative reinforcement*. The same applies to the following example from M1:

> [Aur Minal ko tumhari ek nahi dono aankhein phhodni chahiye thi]s_1 [and you think you can scare us.]s_2
>
> **Gloss**: [And Minal should have taken out not one but both your eyes ...]s_1 [and you think you can scare us.]s_2

The following example from M1 shows the *narrative-evaluative* function:

> [Chaar din ho gaye - kuchh hua nahin na ...]s_1 [So why are we tense ?]s_2
>
> **Gloss**: [It's been four days, nothing has happened ...]s_1 [So why are we tense ?]s_2

In this example, a neutral statement is followed by an evaluation that shows negative sentiment. In the following example from M1, the *narrative-evaluative* function appears in inverse order: the evaluative s_1, thats shows negative sentiment, is followed by the statement s_2.

> [Look roz subah-subah tense hone se kya faayda !]s_1 [We'll go mad ...]s_2
>
> **Gloss**: [Look, what is the use of getting tense every day from the morning ...]s_1 [We'll go mad ...]s_2

The interaction of local and global properties The following example from *Kapoor and Sons* (M2) shows a social function of CS, namely the assertion of Neetu Chachi's identity through the use of English. Neetu Chachi has come to visit from New York. The mother, who has a higher status in age and relationship, continues using Hindi, even though she does know English. Here the switch to English by Neetu Chachi (in s_2 and s_4), who also knows Hindi perfectly well, is used to denote her identity:

> **Neetu Chachi:** [And look what I found in ...]s_1
>
> **Mother:** [Arre ... haan ...!Ye toh pata nahin kitni purani hain. Yeh *taste* kar.]s_1
>
> **Gloss**: [Oh yes! God knows how old this is! Taste this.]s_1
>
> **Neetu Chachi:** [It's delicious!]s_2 [Timmy ko bolo na mujhe kuch mutton recipes bheje]s_3 [Sharic just loves this stuff!]s_4
>
> **Gloss**: [It's delicious!]s_2 [Tell Timmy to send me some mutton recipes ...]s_3 [Sharic just loves this stuff!]s_4
>
> **Mother:** [Haan usse bolti hoon *email* karne ke liye.]s_5
>
> **Gloss**: [Ok, I will tell her to email them to you.]s_5

The examples shown in the previous paragraphs are based on informal conversation among friends and acquaintances. The following example from the movie *Pink* (M1) shows an interaction in a formal court-room setting. In this exchange, CS is aligned with situational changes based on social factors. The following statement is made by Deepak, the attorney defending Minal in the trial:

> [Minal Honourable judge sahab aapko baithne ke liye keh rahein hain]s_1 [and please be quiet]s_2
>
> **Gloss**: [Minal, the honourable judge is asking you to sit down ...]s_1 [and please be quiet]s_2

Dimension	Property	D1		M1		M2		M3		M4		M5		M6	
		Speech		Pink		Kapoor …		Neerja		Talvar		Ek main …		D-day	
Global Properties (of corpora)															
Granularity	languages	EN, HI		EN, HI		EN, HI		EN, HI		EN, HI		EN, HI		EN, HI	
	# tokens	38,624		10,541		8,227		3,438		4,404		5,386		5,420	
	# switch points	6,561		2,790		2,421		1,204		1,534		2,052		1,364	
	more Granularity: see Table 1														
Modality	written, formal	-		-		-		-		-		-		-	
	written, informal	+		+		+		+		+		+		+	
	speech	+		(+)		(+)		(+)		(+)		(+)		(+)	
	audio-visual	-		(+)		(+)		(+)		(+)		(+)		(+)	
Discourse	monolog	+		-		-		-		-		-		-	
	dialog	+		+		+		+		+		+		+	
	multi-party	-		+		+		+		+		+		+	
	max # speakers	2		13		11		9		7		15		7	
Social	hierarchy levels	-		-, +		-, +		-, +		-, +		-, +		-, +	
	familiarity	-		-, +		-, +		-, +		-, +		-, +		-, +	
	repetition/emphasis	+		+		+		+		+		+		+	
Local Properties (of CS segments and CS pairs)															
		EN	HI	EN	HI	EN	HI	EN	HI	EN	HI	EN	HI	EN	HI
Syntactic	sentence	0.014	0.078	0.105	0.073	0.064	0.096	0.080	0.096	0.056	0.063	0.076	0.067	0.007	0.069
	phrase	0.098	0.579	0.197	0.568	0.025	0.578	0.252	0.578	0.266	0.553	0.315	0.543	0.251	0.608
	word	0.853	0.229	0.664	0.315	0.666	0.248	0.666	0.252	0.668	0.357	0.581	0.331	0.702	0.294
	tag	0.025	0.010	0.066	0.019	0.197	0.047	0.140	0.030	0.087	0.050	0.144	0.048	0.034	0.026
	named-entity	0.107	0.010	0.043	0.071	0.043	0.057	0.041	0.096	0.050	0.087	0.054	0.045	0.083	0.069
	continuation=0	0.096	0.254	0.147	0.163	0.164	0.219	0.184	0.248	0.171	0.238	0.191	0.214	0.124	0.231
	continuation=1	0.386	0.240	0.349	0.335	0.334	0.280	0.319	0.266	0.325	0.263	0.309	0.286	0.380	0.283
Sentiment	sentiment=-1	0.167	0.171	0.371	0.378	0.401	0.401	0.120	0.144	0.157	0.158	0.352	0.359	0.158	0.165
	sentiment=0	0.690	0.699	0.571	0.571	0.851	0.463	0.871	0.875	0.821	0.829	0.546	0.577	0.829	0.843
	sentiment=+1	0.117	0.121	0.048	0.043	0.162	0.138	0.023	0.014	0.016	0.015	0.103	0.067	0.022	0.018
Pragmatic	has-DM	0.037	0.438	0.061	0.253	0.030	0.206	0.022	0.144	0.018	0.196	0.047	0.212	0.014	0.213
Compound	speaker-change=1	0.027	0.041	0.046	0.076	0.059	0.102	0.061	0.098	0.059	0.082	0.056	0.087	0.032	0.067
	speaker-change=0	0.460	0.455	0.454	0.425	0.441	0.399	0.446	0.419	0.441	0.421	0.445	0.414	0.473	0.446
	topic-change	0.024	0.019	0.022	0.027	0.027	0.025	0.015	0.017	0.016	0.020	0.024	0.027	0.010	0.021
	translation	0.001	0.001	0.001	0.001	0.000	0.001	0.000	0.003	0.000	0.002	0.000	0.001	0.001	0.001
	senti-pair $\langle -1, 1\rangle$	0.000	0.001	0.001	0.004	0.003	0.010	0.000	0.000	0.000	0.000	0.001	0.005	0.000	0.002
	senti-pair $\langle 1, -1\rangle$	0.001	0.002	0.001	0.003	0.005	0.005	0.000	0.000	0.000	0.001	0.004	0.002	0.001	0.001
	senti-pair $\langle 1, 0\rangle$	0.002	0.005	0.003	0.004	0.013	0.009	0.007	0.001	0.003	0.001	0.017	0.004	0.001	0.001
	senti-pair $\langle 0, 1\rangle$	0.003	0.005	0.001	0.006	0.004	0.019	0.002	0.005	0.001	0.002	0.003	0.020	0.000	0.002
	senti-pair $\langle -1, 0\rangle$	0.004	0.006	0.013	0.032	0.014	0.028	0.006	0.018	0.011	0.018	0.018	0.023	0.006	0.018
	senti-pair $\langle 0, -1\rangle$	0.004	0.005	0.015	0.030	0.014	0.033	0.018	0.012	0.007	0.014	0.017	0.022	0.007	0.014
	discourse relation†	0.010	0.040	0.032	0.014	0.025	0.009	0.012	0.003	0.021	0.003	0.022	0.007	0.015	0.007

Table 2: Footprinting CS corpora; DM = Discourse marker; + and - stand for boolean true and false; † marks property labels created automatically. Statistics in the lower half of the table are relative to the number of switch points in each dataset.

English is the marked language in the court-room setting, and switching to English emphasizes and reinforces the strength of the request. The emphasis is further enhanced by a local function, namely starting the English segment with discourse marker *and*. Deepak in his role as the attorney uses code-switching to English to emphasize the seriousness of his request, which is licensed by the hierarchy induced by the formal court-room setting. The characters Deepak and Minal are not familiar figures, and they are not separated by a hierarchy outside of the courtroom setting, so it would be rude for Deepak to switch to English to reinforce his request in this way in an informal setting.

A third interesting example of the interaction of social and global functions of code-switching occurs in the movie *Neerja* (M3). In this movie, a group of terrorists kidnaps a plane. They are opposed by a group of police officials. Depending on who has the upper hand in the negotiations, each of the groups changes between formal, polite language (i.e. talking up when being lower in the hierarchy), or rude, impolite language (i.e. talking down when being higher in the hierarchy). One main characteristic of the impolite language is the stronger use of code-switching. These changes between up-talk and down-talk show an interaction between social hierarchy and use of code-switching in discourse.

Summary The above examples show an application of our representation to the identification and derivation of CS functions. To further validate our strategy for the identification of CS functions based on properties, we will contrast the statistics on functions automatically derived from properties to manually annotated functions for our datasets. This will help to identify property configurations that are *sufficient* for the (automatic) identification of CS functions in future work.

6. Conclusion

We present an integrated representation of code-switching functions to facilitate their systematic empirical study, particularly the interaction between local and social aspects of the CS functions. The proposed representation is language-independent and can be extended by additional properties (for instance gestures for multi-modal corpora) and emerging functions.

Comparative and systematic corpus-based study of CS is desired (Gullberg et al., 2009; Myslín and Levy, 2015) and facilitated by an increasing number of available corpora (Diab et al., 2014; Çetinoğlu, 2017; Rudra et al., 2016); With our framework we aim to contribute to the comparative corpus-based study of code-switching and to foster the discussion of the interaction between local and social functions. Besides extending and improving the proposed representation in our future work, we plan further applications of the framework to exemplify in which ways it can be used for CS research. Moreover, we will study the automatic derivation of functions in more detail, with a particular focus on discourse relations.

Acknowledgments

This paper is built upon two annotation studies. We gratefully acknowledge the support by Prof. Niloy Ganguly, Koustav Rudra, and Bidisha Samanta during the course of the two studies. Adithya Pratapa kindly provided the language tags and CMI metrics. Last, but not least, we acknowledge the contributions of our annotators Sanjana Arora, Shivangi Shukla, and Jyothsna Chakravarthy.

7. Bibliographical References

Abdul-Zahra, S. (2010). Code-switching in language: An applied study. *Journal Of College Of Education For Women*, 21(1):283–296.

Auer, P. (2013). *Code-switching in conversation: Language, interaction and identity*. Routledge.

Begum, R., Bali, K., Choudhury, M., Rudra, K., and Ganguly, N. (2016). Functions of Code-Switching in Tweets: An Annotation Framework and Some Initial Experiments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1644–1650, Portorož, Slovenia.

Çetinoğlu, Ö. (2017). A code-switching corpus of turkish-german conversations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 34–40, Valencia, Spain.

Crystal, D. (1987). *The Cambridge Encyclopedia of Language*. Cambridge University Press, Cambridge, U.K.

Diab, M., Hirschberg, J., Fung, P., and Solorio, T. (2014). *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Doha, Qatar.

Gambäck, B. and Das, A. (2016). Comparing the Level of Code-Switching in Corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1850–1855.

Gullberg, M., Indefrey, P., and Muysken, P., (2009). *Research techniques for the study of code-switching*, pages 21–39. Cambridge University Press, Cambridge.

Gumperz, J. (1982). *Discourse Strategies*. Cambridge University Press, Cambridge.

Guzmán, G., Ricard, J., Serigos, J., Bullock, B. E., and Toribio, A. J. (2017). Metrics for modeling code-switching across corpora. In *Proceedings of the Interspeech 2017 Special Session on Code Switching*, Stockholm, Sweden.

Myslín, M. and Levy, R. (2015). Code-switching and predictability of meaning in discourse. *Language*, 91(4):871–905.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). SemEval-2013 Task 2: Sentiment Analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, GA, USA.

Poplack, S. (1988). Contrasting patterns of code-switching in two communities. *Codeswitching: Anthropological and sociolinguistic perspectives*, 48:215–44.

Poplack, S. (2015). Code switching: Linguistic. In *International Encyclopedia of the Social and Behavioral Sciences, 2nd edition*, pages 918–925. Elsevier Ltd.

Prasad, R., Webber, B., and Joshi, A. (2014). Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics*, 40(4):921–950.

Pratapa, A. and Choudhury, M. (2017). Quantitative characterization of code switching patterns in complex multi-party conversations: A case study on hindi movie scripts. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 75–84, Kolkata, India, December. NLP Association of India.

Reyes, I. (2004). Functions of code switching in school children's conversations. *Bilingual Research Journal*, 28:83–96.

Rijhwani, S., Sequiera, R., Choudhury, M., Bali, K., and Maddila, C. S. (2017). An Exploratory Study of Structural and Functional Aspects of English-Hindi Code Switching in Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada.

Riloff, E., Qadir, A., Surve, P., De Silva, A., Gilbert, N., and Huang, R. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, WA, USA.

Rudra, K., Rijhwani, S., Begum, R., Bali, K., Choudhury, M., and Ganguly, N. (2016). Understanding Language Preference for Expression of Opinion and Sentiment: What do Hindi-English Speakers do on Twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141, Austin, TX, USA.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge University Press, Cambridge, U.K.