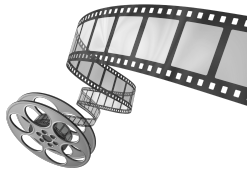

The AMC TRANSCRIPTION CONVENTIONS

How to quote:

Forchini, Pierfranca. 2022. *The AMC Transcription Conventions.* www.americanmoviecorpus.net

The AMC contains the ***orthographic transcriptions*** of dialogs from movies produced in the United States of America from 1959 to 2019. These movie dialogs were transcribed by proficient speakers of English who received specific training. The transcriptions were checked at several stages by different transcribers to make sure that they were free from misunderstandings and typographical errors. The conventions illustrated here are the most recent update of the ones illustrated in Forchini (2012) and (2021). Starting from the assumption that orthographic transcriptions are imperfect written approximations of speech, possible misrepresentations may still be present in the dialogs: if found, further updates will be issued and will be illustrated here.

1. FILE FORMAT and CONTENT

The AMC is stored in two electronic formats:

1) The ***.xlsx format*** contains both the ***speaker identifications*** and the ***dialogs*** (see Example 1);

2) The ***.txt format*** (in unicode 8) contains the ***dialogs*** only (see Example 2).

Example 1:

FORREST	what's my destiny mama?
MRS. GUMP	you're gonna have to figure that out for yourself // life is a box of chocolates forrest you never know what you're gonna get
FORREST	mama always had a way of explaining things so i could understand them

Example 2:

```
what's my destiny mama?
you're gonna have to figure that out for yourself // life is a box of chocolates forrest you never
know what you're gonna get
mama always had a way of explaining things so i could understand them
```

1.1 SPEAKER IDENTIFICATIONS

The *Speaker Identification* is present only in the ***.xlsx format*** and is in **UPPERCASE** (see Example 1 above). At the beginning of each turn, the speaker is given a unique identifier which belongs to one of these categories:

a) a **proper name**, if the name is known (see Examples 1, 3 , 4 and 5);

b) a **general identifier**, if the name is not present (in this case the speaker's identifier is numbered and when relevant/possible, the gender is also indicated - see Example 3);

c) the **unidentified male/female speaker** label, if the speaker is not present and the speaker cannot be identified (when relevant/possible, the gender is also indicated - see Example 4).

Example 3:

MEDICAL STUDENT 01	i have one question doctor frankenstein
DR. FRANKENSTEIN	that's frankensteen
MEDICAL STUDENT 01	i beg your pardon?
DR. FRANKENSTEIN	my name is pronounced frankensteen

Example 4:

COLONEL CHESTER PHILLIPS	grenade
GILMORE HODGE	move move move
STEVE ROGERS	get away get back
UNIDENTIFIED MALE SPEAKER	it was a dummy grenade all clear back in formation
STEVE ROGERS	is this a test?

Further Notes:
- A **change of row** in the *.xlsx* files containing the same speaker indicates that the speaker's interlocutor has changed and/or that the speaker is talking in a new/different scene (see Example 5);
- The **setting of the scenes** is not provided.

Example 5:

LINUS	these things are gonna hold us right?
DANNY	they should huh
DANNY	livingston we're set
RUSTY	livingston we're set
LIVINGSTON	basher we're set
BASHER	hang on a minute chief

1.2 DIALOGS

The movie dialogs are present both in the ***.xlsx*** and ***.txt format*** and are transcribed in **lower case** to highlight their distance from written language and emulate the *International Phonetic Alphabet* which does not use capital letters (see Examples 1 and 2). They are spelled in ***American English*** even when they represent speech from other varieties of English.

Further Notes:
- **Abbreviations** are not used in dialogs when they are part of personal titles such as in *mr.*, *ms.*, *mrs., dr.*, *jr.*, *st.* (which are, respectively, transcribed as follows: *mister*, *ms*, *missus, doctor, junior, saint*); they are, instead, used in speaker identifications. **Contracted forms** such as *wanna, gonna, gotta, ma'am, ain't, y'all, 'cause* are transcribed as such if this is how they are pronounced in the dialogs;
- **Songs** and **movies** are not included in the dialogic transcriptions, whereas **television** and **radio talks** are transcribed when relevant to the dialogs;
- **Overlaps** are not indicated and when they occur, the priority is given to the utterance that can be heard first;
- **Punctuation** is kept to a minimum (see Table 1);
- **Inserts** have specific functions and/or connotations (see Table 2).

Table 1. PUNCTUATION and SYMBOLS

//		**marks a point in the dialog where a pause lasting two seconds or more occurs:**
	E.g.	andrea **//** runway is a fashion magazine so an interest in fashion is crucial
?		**marks a point in the text where a question occurs:**
	E.g.	what makes you think i'm not interested in fashion**?**
-		**marks a word that is *incomplete* [at the beginning (a) or at the end (b) of a word] or *hyphenation* [in the middle of a word (c)]:**
	E.g:	(a) okay you got it? now don't let it touch the sides **-ides -ide**s when you're coming out (b) he's the most beautiful thing i've ever seen // but uh // is is he **s-** smart or can he (c) **smash-and-grab** job huh?
< >		**marks that a foreign language (i.e. not English) is spoken:**
	E.g.	oh victor victor we have done it // i'm going to set you free would you like that **<german>**?
<unintelligible>		**marks that the speech/utterance is either not clear or can't be heard:**
	E.g:	ma'am you dropped your book ma'am **<unintelligible>**
:		**separates hours from minutes:**
	E.g:	please bore someone else with your questions and make sure we have pier 59 at **08:00** a.m. tomorrow and remind jocelyn i need to see a few of those satchels that marc is doing in the pony and then tell simone i'll take jackie if maggie isn't available did demarchelier confirm?
.		**marks Latin ante and post meridiem in the 12-hour time convention:**
	E.g:	please bore someone else with your questions and make sure we have pier 59 at 08:00 **a.m.** tomorrow and remind jocelyn i need to see a few of those satchels that marc is doing in the pony and then tell simone i'll take jackie if maggie isn't available did demarchelier confirm?

Table 2. INSERTS

Interjections	**ah** (see NOTE 1), **aargh** (see NOTE 1), **eh**, **ho**, **mm** (see NOTE 1), **oh**, **oof** (see NOTE 2), **ooh** (see NOTE 2), **oops**, **ouch**, **ow** (see NOTE 2), **ugh** (see NOTE 1), **uhu** (see NOTE 2), **uh-uh** (see NOTE 2), **whoa** (see NOTE 2), **whoop**, **whoops**, **whoopee**, **woo** (see NOTE 2), **woo-hoo**, **wow** (see NOTE 2), **ugh** (see NOTE 1), **yuck**. **NOTE 1:** **ah** = only used for *positive connotation* (e.g. for surprise, pleasure, sympathy and realization); **aargh** = only used for *negative connotation* (e.g for pain, anguish, horror, rage or any other strong emotion); **mm** = a unique identifier used either as an interjection (**hm**) or a response form (**mm** and/or **mhm**) to express pleasure, agreement, uncertainty or reflection; **ugh** = used to express disgust or distaste and covers various pronunciations: /ʌg//ʌx/ /ɜ:h/. **NOTE 2:** **oof** = pronounced/u:f/; **ooh** = pronounced /u:/; **ow** = pronounced /aʊ/; **uhu** = pronounced/'ju:hu:/ /'u:hu:/; **uh-uh** = pronounced/'ʔʌʔˌʔʌ/; **whoa** = pronounced/hwoʊ/; **woo** = pronounced/wu:/; **wow** = pronounced/waʊ/.
Attention signals	**hey**, **yo**, **psst**
Response elicitors	**all right? huh? okay/ok? see?**
Response forms	**huh-uh** (see NOTE 3), **mm** (see NOTE 3), **no**, **nope**, **okay**, **ok**, **sure**, **uh-huh**, **yeah**, **yep**, **yes** **NOTE 3:** ***huh-uh*** = pronounced /hʌ'hʌ/ /'hʌhʌ/ **mm** = a *unique identifier* used either as an interjection (**hm**) or a response form (**mm** and/or **mhm**) to express pleasure, agreement, uncertainty or reflection; **mhm** = not used: it is replaced by mm which represents a unique identifier used either as an interjection or a response form to express pleasure, agreement, uncertainty or reflection; **uh-huh** = pronounced/ʌ'hʌ/ /'ʌhʌ/.
Hesitators	**uh** (see NOTE 4), **um** **NOTE 4:** **uh** = only used for *hesitation.*
Others	**mwah** (see NOTE 5), **shh** **NOTE 5:** **mwah** = transcribed only if uttered as such and not for actual kissing.
FURTHER NOTES:	When there are more than six *inserts* (especially *interjections and vocalizations*) in the same utterance by the same speaker, they have not been necessarily transcribed: when they are represented like this *ow ow ow ow ow ow ow* or like this *ouch oh ouch aargh oh wow aargh oh oh mm oh*, it implies that there can be more interjections than those actually represented in the transcribed text (this also means that if the number of interjections is lower than six they all have to be transcribed). Non-dialogic vocalizations (e.g. those occurring in fight scenes) have not been transcribed. This is a good example of the mentioned imperfect written approximation of the speech event and justifies possible misrepresentations of the transcribed dialogs.