

The Challenge

Dubbing companies and subtitling houses often need to extract information from scripts that come in a huge variety of formats. Isolating a speaker list, dialogue and timings is far from easy even when Word macros are used - and can be very time-consuming.

Bull Nerd:	05.16	Dat blauwe joch zit straks vol met blauwe plekken als ik hem te pakken krijg. (lache)	01:06:24:04	JACK	A piece of the steering column. All right, let's go. Keep that collar steady. (to intern) Page my father. (collar : orthopedic brace to support an injured neck) (page : contact or call on mobile message device, i.e. pager, beeper)
Penny:	05.21	(schrikzucht) Snap!	01:06:29:06	MALE INTERN	Already did, but--
Kinderen:	INT. FBI OFFICE				Tell me.
Snap:	MYLES: Whereupon they proceed to tell us we need more and incriminate themselves to their heart's content.				Her tire blew, carjumped the divider, went head-on with an SUV. (tire blew : 'the car wheel's rubber tire exploded') (divider : cement curb or partition separating traffic moving in opposite directions)
Bull Nerd:	0:03:31:20	Seonaidh:	ALEX walks in, numb.		
Snap:	0:03:40:10	Leigidh sinn sios chun a tùir e, glacaidh esain ar tùr.	GENE I expect you to do your - (clocking ALEX) Where the bloody hell have you been? Where'd you hide it?		
	0:03:51:20	Ajay:	MUSIC IN M11 10:29:09		
	0:03:53:20	Seonaidh:	ALEX What? ALEX freezes.		
	0:04:01:00	Ach, tha m'uaireadair na thàmh.	GENE The note. Doctor Zhivago's precious bloody poem. Where is it? ALEX I don't, I'm not-- GENE I thought I'd take another look, get a proper translation. Only I find it's disappeared. Gone. Vanished.		

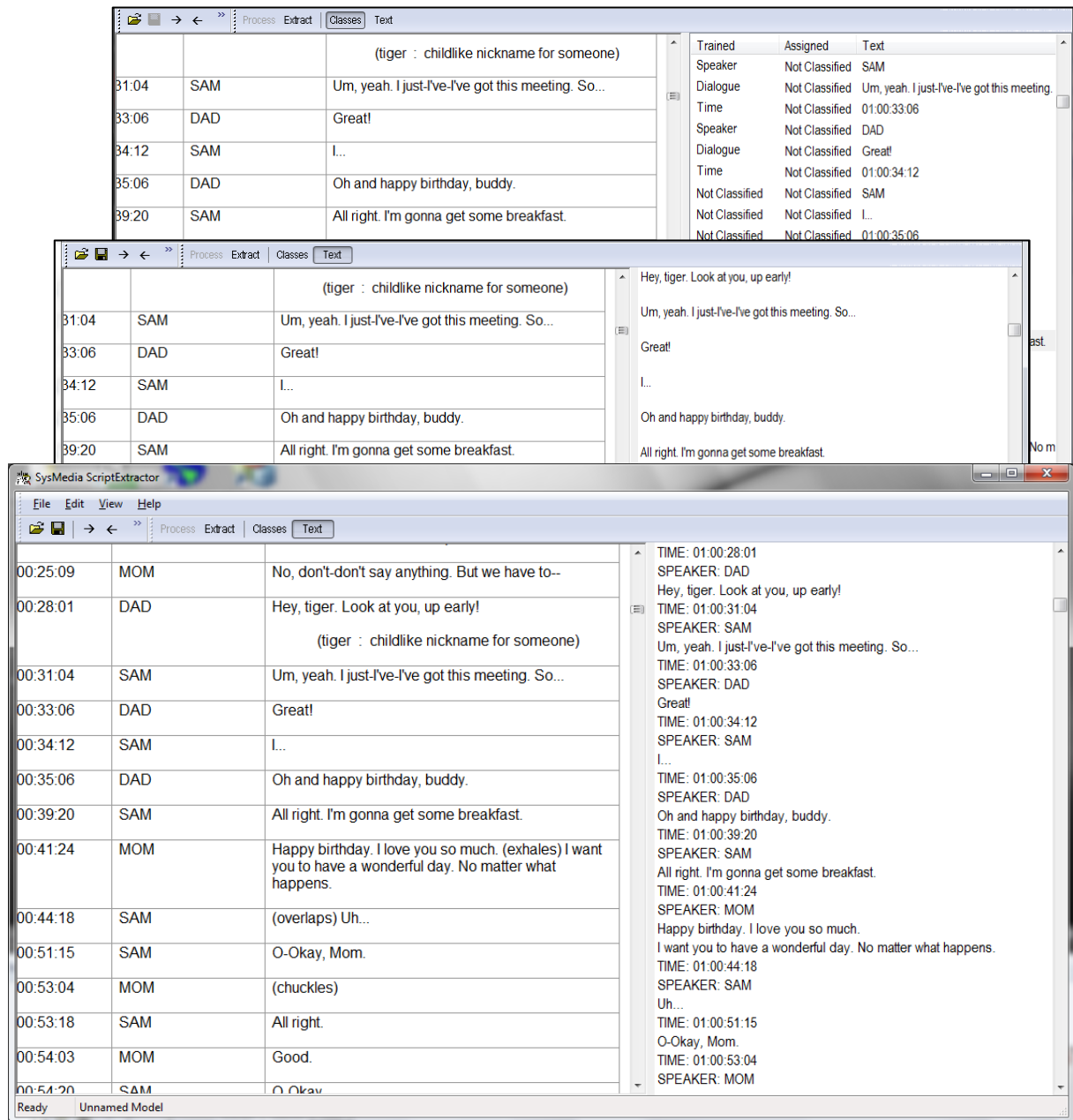
The Answer

SysMedia's Script Extractor utility – a new system for isolating the useful content from scripts based on pattern matching technology. The user just loads a Word document in any self-consistent layout, and then simply gives some example classifications for the different constituent elements. Script Extractor then uses these clues to extract the useful content and present it on screen for checking and export into a consistent layout as TXT or XML.

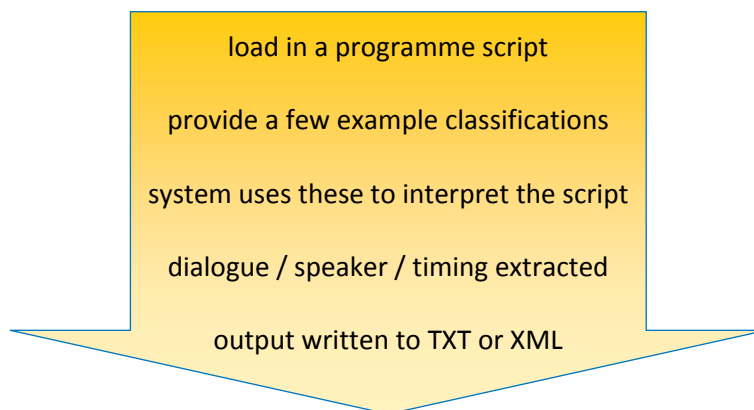
The extracted items can be loaded for subtitling for the deaf and hard of hearing, or used for scoping of dubbing operations or as the starting point for a translation for dubbing or subtitling.

Tests with a wide range of real scripts have shown that with just a few minutes of effort on a new layout, the system delivers very reliable results. The training pattern for a given layout can be saved for re-use, making the process even faster.

SCRIPT EXTRACTOR



The screenshots illustrate the workflow of the Script Extractor. The top two windows show the 'Process' and 'Extract' stages, with a 'Classes' pane on the right listing categories like 'Trained', 'Assigned', and 'Text'. The bottom window shows the 'Text' pane with a detailed list of extracted dialogue segments, including timestamps, speaker names (e.g., MOM, DAD, SAM), and the actual text content.



huge time-saving compared to manual script processing