Subtitle reading speeds in different languages: the case of *Lethal Weapon*

José Luis Martí Ferriol

Universitat Jaume I. Departament de Traducció i Comunicació. Traducció i Interpretació
Av. de Vicent Sos Baynat, s/n. 12071 Castelló de la Plana. Spain
martij@uji.es

Abstract

This paper presents results of subtitle reading speeds in five languages (Spanish, English, German, French and Italian), for subtitles both ripped from a commercial DVD and downloaded from an Internet site of the film *Lethal Weapon* (Richard Donner, 1987). The reading speed calculations are shown in the parameters most widely used in the field: CPS (characters per second) and WPM (words per minute). The research questions posed in this descriptive and quantitative study have been two-fold. In the first place it is intended to find out which of the two above mentioned parameters is best suited to express reading speeds which are language-independent. Secondly, the possible difference, as far as reading speed parameters is concerned, between more commercial DVD-ripped and more «fan-oriented» and downloaded from Internet subtitles is investigated as well. The reading speed calculations were carried out by a new application developed by the author and some other researchers. A total amount of about 5000 subtitles were statistically evaluated.

Keywords: subtitles; reading speed; CPS; WPM; *Lethal Weapon*.

Resum

Aquest article presenta el resultats de velocitats de lectura de subtítols en cinc llengües diferents (espanyol, anglès, alemany, francès i italià), tant per a subtítols extrets d'un DVD comercial com per a subtítols descarregats de portals d'Internet, i obtinguts de la pel·lícula *Lethal Weapon* (Richard Donner, 1987). Els valors de les velocitats de lectura es presenten mitjançant els paràmetres que s'empren majoritàriament en aquest camp: CPS (caracters per segon) i WPM (paraules per minut, del terme anglès *words per minute*). Les preguntes de recerca que es plantegen en aquest estudi descriptiu i quantitatiu són dues. D'una banda, es pretén descobrir quin dels dos paràmetres mencionats més amunt és més adient per a expressar velocitats de lectura de subtítols, amb una magnitud que siga independent de la llengua. D'altra banda, també s'investiga la possible diferència, pel que fa a velocitat de lectura, entre els subtítols més comercials i extrets d'un DVD i els descarregats d'Internet, en moltes ocasions desenvolupats per aficionats (*fan subs*). Els càlculs de velocitats de lectura es van dur a terme emprant una nova aplicació desenvolupada per l'autor amb la col·laboració d'altres investigadores. Tot plegat, es van avaluar estadísticament uns 5.000 subtítols.

Paraules clau: subtítols; velocitat de lectura; CPS; WPM; *Arma letal*.

ISSN 1138-5790 (imprès), ISSN 2014-9735 (en línia)

Summary

1. Introduction

Subtitling, together with dubbing, is the audiovisual translation mode most widely used worldwide. Spain is among the countries (most of them European) with a solid dubbing tradition. However, the introduction of DVD in the audiovisual market at the end of the previous decade, together with digitalization of television broadcast and the rise of accessibility policies in Spain have helped consolidate subtitling as an almost necessary tool to watch audiovisual content in our country.

Many scholars have devoted their research efforts to subtitling, having most of them started back in the 90s. It is not intended to make here a detailed review of all their contributions, since some of them like Díaz Cintas (2003) and Chaume (2004) have carried out this task with rigor and completeness. Instead, this paper is intended to focus on the viewer as the key element of subtitling reception, and more specifically, on the viewer's ability to read subtitles, which is usually expressed by means of two different parameters: characters per second (CPS) and words per minute (WPM).

A more thorough review of the bibliography devoted to subtitle reading speed and their expression by means of the two above mentioned parameters can be found in Martí Ferriol (2012), a publication which introduced the reading speed application (or tool) which has also been used to obtain the empirical and quantitative results presented in this article. Such a review includes scholars like Karamitroglou (1998), Mayoral (2001), Díaz Cintas (2008), Toda (2009) and Romero Fresco (2009). All of them present reading speed values for subtitles which could be termed as «conventional», in the sense that they include up to two lines of about 35 characters per line. Experimental research data has shown that an average reader needs about 6 seconds to read comfortably a subtitle of those characteristics. Based on those results, the so-called «6 second rule» has been developed, which is commonly accepted as a standard as of nowadays, not only in the subtitling industry, but also in academic and teaching environments. Romero Fresco (2009) elaborates on the empirical prove from which the rule stems: an experiment carried out by D'Ydewalle *et al.* in 1987 by using «eye tracking» technology; a technology widespread used nowadays to investigate reception of audiovisual contents, as well as information relevance.

> Using eye-tracking technology, he tested three different presentation times for subtitles: two lines of 32 characters in 4 seconds (approximately 192 wpm), 6 seconds (130 wpm) and 8 seconds (96 wpm) respectively. The object of this study was to ascertain if the six-second rule (a full two-line subtitle displayed on screen for 6 seconds and shorter subtitles scheduled proportionally), accepted as com-

> mon practice in most subtitling countries, could be validated by empirical research on reading speed. His results leave little room for doubt, the six-second rule being identified as setting the appropriate reading speed for the participants. This rule has later on been supported by other scholars such as Díaz Cintas (2003), who applies it to longer lines than the ones referred to by D'Ydewalle (72 characters instead of 64), thus setting the recommended speed at 144 wpm (12 cps). (Romero Fresco, 2009: 114)

It seems that the «6 second rule» is most widely accepted by viewers. If those reading speed values are translated into the parameters employed in our empirical study, values of 12 CPS and 144 WPM are obtained.

2. Research questions and hypotheses

The process of subtitling involves two fundamental stages: «spotting» (or text segmentation including time assignment), and subtitle creation in itself. By spotting it is usually understood the definition of the cue-in and cue-out times for a given subtitle. This means that it is necessary to specify the moment when the dialogue in the original source text starts, so that the corresponding subtitle appears on the screen. Likewise, the subtitle must disappear from the screen when the corresponding fragment of the source text dialogue covered by the subtitle finishes. The difference between the cue-out and the cue-in times sets the subtitle duration. It must be taken into account that, if the spotting is carried out correctly, it allows the viewer for the reading of the subtitle without problems or difficulties. As a general rule, subtitle duration on screen spans between 2 and 6 seconds.

The main constraint in the subtitle creation phase in itself, which also accounts for its implicit difficulty, lies on the fact that generally only two lines (of about 35 characters each) are available to translate that particular piece of dialog. As a consequence, translation for subtitling is intrinsically characterized by the need to synthesize (or condense) the information included in the source language.

The process of subtitle development can be undertaken in a manual fashion, or by means of commercial subtitling programs. A series of those are available on the market: some of them are free software, others can be downloaded from specific sites on the Internet as a «demo» version for a limited period of time, and others require the purchase of a license, which can be acquired by paying in some cases significant prices. A list of the most common subtitling software programs can also be found in Martí Ferriol (2012). Some of those also provide the user with reading speed values expressed by means of the most common parameters (CPS and/or WPM): in some cases, only one of the two is offered, although some programs generate values for both. Oddly enough, those values do not always coincide.

Such inconsistencies in reading speed parameters as provided by conventional subtitling programs became the motivation to develop a new tool to fulfill this functionality, i.e. to provide solid reading speed values calculated in a simple and robust fashion. As stated above, details on the application operation and its advantages are thoroughly detailed in the article by the same author already mentioned. This application (based on proprietary design and development) has been the one used to carry out CPS and WPM calculations for the subtitles in five languages of the film selected as object of study.

As far as research questions are concerned, two of them have already been mentioned based on intuitive, common sense practice at the very beginning of this paper. If they are rephrased in terms of hypotheses, they may read as the following:

- CPS should be a better-suited parameter to express reading speed versus WPM, since it could be language-independent (word length may vary across languages, see Mayoral, 2001).
- Subtitles ripped from commercial DVD or downloaded from Internet sites should not present significant differences, as far as reading speed considerations are concerned.

As it has been explained above, only common sense practice and prior knowledge on the field have been used to formulate these hypotheses in the way it is done, and not in the reverse one. The following section deals with the corpus selection and its justification.

3. Corpus description

Making the appropriate corpus selection for an empirical and descriptive research work usually involves a decision-making process which is not always straightforward. For this particular case, since reading speed figures in several languages are to be calculated and both subtitles from a DVD and Internet need to be analyzed, the selection should entail a film which has been famous worldwide and not too recent, because the availability of subtitles in different languages (and from both sources: DVD and Internet) would then probably be secured.

Based on these criteria, the film *Lethal Weapon* (Richard Doner, 1987) was selected. The film is the first of a well-know series of four (as a matter of fact, a fifth one is scheduled for 2012) where the actors Mel Gibson and Danny Glover form a peculiar pair of very different policemen, both in their personal lives and in their way to approach the job. The combination of these strong personalities with an action plot consisting of a wild rhythm and unexpected turns, made the script a best candidate to achieve success (see http://www.imdb.com/title/tt0093409/awards for a list of awards of this film).

A commercial copy of the DVD was purchased in Spain, a copy with subtitles in as many as 18 languages. Apart from these, subtitles for the deaf and hard of hearing were also offered in two languages: German and English. These subtitles for five of the languages, once extracted from the DVD, were the ones to be compared with the corresponding sets (in terms of languages) of subtitles downloaded from Internet sites.

The best procedures and programs to be used for extracting subtitles from DVD have been documented, for example, in Martí Ferriol (2009). There are many potential applications which can help the user do that. For this particular research, the combination of the program «DVD Decrypter» and the standard «SubRip» (version 1.50) were utilized. Both of them are free available software from the Internet.

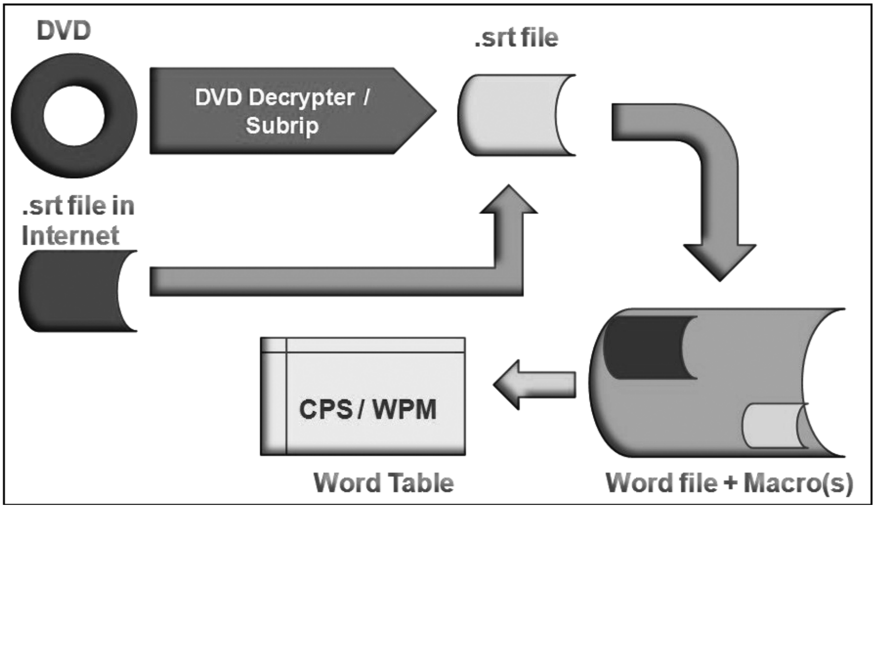
Another alternative way to get subtitles is to download them from specific sites in Internet. There are several websites which specialize on this. One of the most famous ones is «Allsubs» (www.allsubs.org), from where the sets in five languages requested for this research were obtained. As an example, the address for the subtitles in Italian for this film would be: http://www.allsubs.org/subtitulos/busqueda-subtitulos/lethal+weapon++it/20.

4. Methodology

There follows a short explanation of the working procedure which has been carried out in order to obtain the results presented in the next section. In the first place, the DVD subtitles in the five selected languages were ripped and the corresponding files were generated. As far as the Internet subtitle files in the same five languages, they were located and downloaded from the above mentioned website. Each of the ten files (5 languages and 2 sources) contains about 500 subtitles. These numbers oscillate up to 10%, as it will be shown in the tables in next section. However, they all cover the same exact period of time of the film, of around the first 40 minutes.

The contents of each of the ten files were copied each to a different Word file which included the reading speed application tool (a macro in Visual Basic). A slight formatting adjustment for each subtitle was also necessary for the macro to make the calculations properly. Then, the macro was run for each file, and a table which contained the reading speed values expressed in CPS and WPM was generated.

The complete process is shown in the following diagram:



The reading speed results in both parameters were subsequently transferred from the table (Word file) to a program to carry out the statistical handling of the data (Minitab for Windows), where the necessary tests and calculations discussed in the following section have been performed.

5. Results and discussion

The following two tables show the results obtained with the macro for subtitles of the film in five languages, and two sources (DVD and Internet). Table 1 shows the results in CPS, and Table 2 the ones in WPM.

The statistical values shown in the tables will be explained first. Then, their implications on the two stated hypotheses will be further analyzed.

As it is shown, the overall number of studied subtitles is 5020, and the reader can see that the numbers change depending on the language and the source. These differences can to a certain extent be explained by the fact that the distribution of subtitles with just one line (also called «one-liners») and two lines («two-liners») can change, depending on the person who does the spotting. This practice, like translation itself, is not an exact science, and more often than not, more than one solution is possible.

The reading speed figures were analyzed by using some general descriptive statistical calculations. To be more precise, for each set of subtitles expressed in CPS and WPM, values for means, standard deviations, medians and Q1 and Q3 (first and third quartile) were obtained. Each of the two tables shows the results. Based

Table 1. Reading speed results expressed in CPS

CPS VALUES (RIP vs. WWW) FOR 5 LANGUAGES *(LETHAL WEAPON)*						
Language	**Source**	**# OF subtitles**	**Mean**	**Std. Dev.**	**Median**	**Q1 - Q3**
DE	CPS_RIP	530	11,5	4,1	12	9 - 14
DE	CPS_WWW	486	11,8	4,3	12	9 - 14
ES	CPS_RIP	530	11,5	3,9	12	9 - 14
ES	CPS_WWW	451	11,8	3,7	12	9 - 14
EN	CPS_RIP	531	12,2	3,9	13	10 - 15
EN	CPS_WWW	540	12,6	4	13	10 - 15
FR	CPS_RIP	490	12,6	4,1	13	10 - 15
FR	CPS_WWW	443	12,1	3,7	12	9 - 15
IT	CPS_RIP	533	12,6	4,3	13	9 - 16
IT	CPS_WWW	486	12,8	4,3	13	10-16
Total / Avg.		**5020**	**12,2**	**4,03**	**12,5**	
Max.			**12,8**	**4,3**	**13**	**16**
Min.			**11,5**	**3,7**	**12**	**9**
% Variation			***11***	**15**	***8***	**56**

Table 2. Reading speed results expressed in WPM

WPM VALUES (RIP vs. WWW) FOR 5 LANGUAGES (*LETHAL WEAPON*)						
Language	**Source**	**# OF Subtitles**	**Mean**	**Std. Dev.**	**Median**	**Q1 - Q3**
DE	WPM_RIP	530	124	50	124	88 - 152
DE	WPM_WWW	486	124	52	125	87 - 154
ES	WPM_RIP	530	128	55	126	88 - 159
ES	WPM_WWW	451	124	46	124	90 - 151
EN	WPM_RIP	531	146	51	151	110 - 183
EN	WPM_WWW	540	153	57	155	111 - 191
FR	WPM_RIP	490	156	59	151	115 - 195
FR	WPM_WWW	443	150	54	148	110 - 183
IT	WPM_RIP	533	136	56	136	95 - 175
IT	WPM_WWW	486	138	55	137	97 - 180
Total / Avg.		**5020**	**138**	**53,5**	**138**	
Max.			**156**	**59**	**155**	**195**
Min.			**124**	**46**	**124**	**87**
% Variation			***23***	**24**	***23***	**78**

on these results, maximum and minimum values were computed, with the idea to choose the parameter with the lowest variation percentage as the best-suited to convey reading speed values. The following conclusions can thus be drawn:

- Mean and median values for both parameters (12.2 and 12.5 for CPS and 138 and 138 for WPM) are very close to each other, or even the same. In statistical terms, this fact can be considered a symptom of «normal» behavior, or an indication that the values are the result of a non-manipulated process.
- The percent variation for the 10 series of data expressed in CPS is 11%, while the same for values in WPM is twice as big (23%). This simple calculation proves that expressing reading speed in CPS for different languages can be more consistent.
- The same conclusion can be obtained if values for variation in standard deviations (15% for CPS and 24% for WPM) and medians (8% for CPS and 23% for WPM) are considered.
- Finally, comparisons of percentage variation for Q1 and Q3 indicate the same. For CPS, limit values of 9 an 16 are presented for a median of 12.2 (56% variation), whereas the values for WPM are 197 and 87 for a median of 138 (78% variation).
- As far as reading speeds expressed in WPM for different languages is concerned, they seem to be grouped in three levels (which have been shown in the table by different shades): Spanish and German appear to be the ones with the lowest values (120's), while English and French are the ones with highest

values (150's). Italian lies somewhere in between the other two groups, with values closer to the 140's.

- The reading speed results for a corpus of more than 5000 subtitles confirm the famous «6 second rule», which translates into values of 12 CPS and 144 WPM, very close to the ones obtained in this research (12.2 and 138, respectively).

The data presented above seems to validate the first of the hypothesis formulated for this particular case study: CPS should be a better parameter to express reading speed across languages than WPM, since it is more constant. The higher variation in reading speed expressed in WPM tends to generate groups of languages, whose behavior as far as this parameter is concerned, is more alike. All these conclusions are, of course, only valid for the present case study, which covers a particular film, its subtitles in five languages from two different sources, and an overall number of about 5000 subtitles.

In order to validate the second hypothesis, meaning that subtitles from the two sources behave in the same way (or belong to the same population, to put it in statistical terms) regarding reading speeds, two populations need to be created first from the Table 2 above. The first one would consist of the WPM values for subtitles ripped from the DVD, namely: 124, 128, 146, 156 and 136. The second one, with the corresponding values for the WPM values from Internet files would be: 124, 124, 153, 150 and 138. Reading speeds expressed in WPM rather than in CPS have been selected for this particular exercise, according to the data shown in Tables 1 and 2. This decision has been willingly made with the intention to present a «worst-case scenario», due to the fact that the sets of values expressed by means of this parameter present a larger variability (as documented above), thus making the demonstration of the «null hypothesis» (there is no variation between the two sets) more difficult.

Then, a hypothesis test («Two-sample T-test») will be conducted, to try to prove if both series are different or not (the latter would be the above mentioned «null hypothesis», in statistical terms). The results of the test are shown below, as well as on the illustrative graph (see figure 1).

The test results clearly indicate that the two series belong to the same population. Firstly, the means are almost exactly the same (138 vs. 137.8 for ripped subtitles and Internet files, respectively). The same happens with the standard deviations (13,1 vs. 13,8). However, the most conclusive result is the «p» value (0.98), which would have to be lower than 0.05, so that we could conclude with a 95% certainty that both series had average values which make them belong to different populations. The box plot graph depicts in a graphic way the same conclusion: the boxes generated by the two series almost overlap perfectly. When $p<0.05$, this overlapping does not exist.

The above hypothesis test helps us confirm our second hypothesis: for this particular corpus of 5000 subtitles for a single film, no differences were observed, as far as reading speed expressed in WPM is concerned, between subtitles ripped from DVD and downloaded from Internet sites. It may be argued, of course, that

```
Two sample T for WPM_RIP vs WPM_WWW
              N      Mean     StDev    SE Mean
WPM_RIP       5      138,0    13,1       5,9
WPM_WWW       5      137,8    13,8       6,2

      95% CI for mu WPM_RIP - mu WPM_WWW: (-19,9; 20,3)
T-Test mu WPM_RIP = mu WPM_WWW (vs not =): T = 0,02   P = 0,98   DF = 7
```

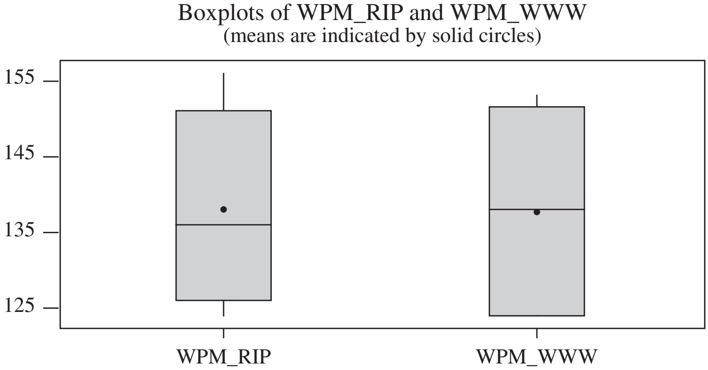


Figure 1. Two Sample T-Test and Confidence Interval.

the subtitles downloaded from the Internet site were originally ripped from other DVD's, before being posted on the web. This may turn out to be true (or not). In any case, these DVDs may have been published by other companies in different countries, and they might as well have been manipulated (or not) by the people who ripped them before uploading them on the web. This is a variable over which the researcher had no control. However, the empirical tests conducted in the study indicate that the degree of manipulation of DVD subtitles before being posted on Internet websites may have been low, due to the consistent values obtained for reading speed parameters in the five different languages.

In order to conclude this study, it should be pointed out that the reading speed calculation tool has proved very robust and user-friendly, and that being the case, it is bound to produce some new and interesting results for new corpora and subsequent case studies.

Bibliography

Chaume, F. (2004). *Cine y traducción*. Madrid: Cátedra.

Díaz Cintas, J. (2003). *Teoría y práctica de la subtitulación inglés-español*. Barcelona: Ariel.

— (2008). «Teaching and learning to subtitle in an academic environment». In: Díaz Cintas, J. (ed.). *The didactics of Audiovisual Translation*. Amsterdam/Philadelphia: John Benjamins, p. 89-105.

Karamitroglou, F. (1998). «A Proposed Set of Subtitling Standards in Europe». *Translation Journal, vol. 2* (http://accurapid.com/journal/04stndrd.html).

Mayoral, R. (2001). «El espectador y la traducción audiovisual». In: Chaume, F. and Agost, R. (eds). *La traducción en los medios audiovisuales*. Castelló: Universitat Jaume I, p. 33-48.

Martí Ferriol, J.L. (2009). «Herramientas informáticas disponibles para la automatización de la traducción audiovisual ("revoicing")». *Meta 54 (3)*, p. 622-630.

— (forthcoming 2012). «Subtitle reading speed: a new tool for its estimation». *Babel*.

Romero Fresco, P. (2009). «More haste less speed: Edited versus verbatim respoken Subtitles». *Vial vol. 6* (http://webs.uvigo.es/vialjournal/pdf/Vial-2009-Article6.pdf).

Toda, F.; González Iglesias, D. (2009). «Spoken language and ICC: Managing Cultural Diversity in Dubbing and Subtitling in Spain». Boras NIC 2009 (http://nic.hb.se/index.php?id=12).