Applied Psycholinguistics **28** (2007), 661–677
Printed in the United States of America
DOI: 10.1017/S014271640707035X

The use of film subtitles to estimate word frequencies

BORIS NEW
Université Paris Descartes and CNRS

MARC BRYSBAERT
Royal Holloway, University of London

JEAN VERONIS
Université de Provence

CHRISTOPHE PALLIER
CNRS, INSERM, and Service Hospitalier Frédéric Joliot

Received: April 3, 2006 Accepted for publication: January 18, 2007

ADDRESS FOR CORRESPONDENCE
Boris New, 71 Avenue Edouard Vaillant, Boulogne-Billancourt F-92100, France.
E-mail: boris.new@univ-paris5.fr

ABSTRACT
We examine the use of film subtitles as an approximation of word frequencies in human interactions. Because subtitle files are widely available on the Internet, they may present a fast and easy way to obtain word frequency measures in language registers other than text writing. We compiled a corpus of 52 million French words, coming from a variety of films. Frequency measures based on this corpus compared well to other spoken and written frequency measures, and explained variance in lexical decision times in addition to what is accounted for by the available French written frequency measures.

The availability of digitally stored texts on the Internet has opened a completely new avenue for linguists and psycholinguists to gain access to large corpora of written language. For instance, Blair, Urland, and Ma (2002) and New, Pallier, Brysbaert, and Ferrand (2004) showed that word frequency estimates obtained with Internet search engines correlate highly with those from well-established sources such as Celex for English (Baayen, Piepenbrock, & Gulikers, 1995) and Lexique for French (New, Pallier, Ferrand, & Brysbaert, 2004). This opens the possibility to obtain frequency estimates for words in languages without an existing frequency list. Similarly, Grondelaers, Deygers, Van Aken, Van Den Heede, and Speelman (2000) showed how Internet sources can be used to get access to texts from different language registers. They downloaded materials from newspapers, discussion groups, and chat channels, and showed how the presence of a particular word ("er" in Dutch, a word meaning something like "there" and in many instances

facultative) varied systematically between these different language registers (see also Desmet, De Baecke, Drieghe, Brysbaert, & Vonk, 2006, for another use of this particular corpus).

A much bigger problem is to find spoken word frequencies. The method used thus far consisted of registering dialogues (e.g., from the radio or from "spontaneous" interactions) and transcribing them. Unfortunately, much of the transcription still has to be done by hand, as current programs are not good enough to yield an acceptable error rate. The estimated transcription costs amount to some 40 hr per 1 hr of spoken input. For this reason, the availability of spoken word frequencies is very limited, both in terms of the magnitude of the corpus on which they are based and in terms of the languages for which they are available. Still, it is generally accepted that spoken word frequencies are urgently needed, because there is a feeling that written word frequencies seriously underestimate the frequency with which words are encountered in everyday life (e.g., words related to eating, clothing, furniture, casual social interactions, etc.).

The ideal spoken corpus would be to record everything some people listen to and say during everyday life. However, as mentioned previously, making such a corpus would be very costly.

There is, however, one source of transcribed spoken text widely available on the Internet: subtitles of films and television programs. This type of corpus has two potentially interesting features. First, it deals with spoken interactions between people in a visible setting. Second, for many people films and television programs comprise a substantial part of their language input, given that current estimates of television watching easily reach an average of 3–4 hr per day. Below we discuss the method we used and the results we obtained for the French language. We expect very similar findings for other languages.

COLLECTING A CORPUS OF SUBTITLES

The raw materials

Digital movies allow users to watch films with and without subtitles. This is done by using two different files: one with the original movie and one with subtitles and codes to synchronize the presentation of the subtitles with the movie. Thousands of subtitle files are freely available on the Internet, and their number is constantly increasing. In French we saw the number double in 2 years. First we searched the net for Web sites providing good subtitles in French using Google. Once the Web site was found, we used a Web crawler named Wget to download subtitles for 9,474 movies and television series. The films came from four different categories[1]:

1. subtitled French films for a total of 1.9 million words (e.g., *Camille Claudel*, *C'est arrive près de chez vous*),
2. subtitled English and American movies for a total of 26.5 million words (e.g., *Arizona Dream*, *Schindler's List*),
3. subtitled English and American television series for a total of 19.5 million words (e.g., *Friends*, *Ally Mc Beal*), and
4. subtitled non-English-language European films for a total of 2.5 million words (e.g., *Cria Cuervos*, *Good Bye Lenin!*).

Most of the materials movies were from the English language, in line with the Anglo-Saxon dominance in the film industry. We made a special effort, however, to include as many French materials as we could find. Most of them were French films that had been subtitled for the hearing impaired.

Once the files had been downloaded, they needed to be cleaned for optical character recognition (OCR) mistakes. Most of those subtitles files have been scanned from DVD with an OCR system to extract the subtitles, and sometimes the OCR software confuses two letters such as "I" and "l." We also needed to get rid of the time indications and other nonfilm-related materials (like the names of the actors and the director). This is the only part of the whole process that has to be done manually and it can be done in less than 2 min per movie. This is an example of the type of materials that remains after this cleaning process:

C'est ton ami!
Elle n'est plus aussi jolie qu'à 29 ans.
Mlle Green aimerait fixer quelques principes avant de sortir.
Veuillez ne pas employer les mots "vieux" . . . "sur le déclin" ou "toujours verts pour leur âge." Ils collent bien!
Amène-toi!
Monica a préparé le petit-déj.
Des pancakes au chocolat!
On a des cadeaux!
Des bien?
Tous issus de la liste que tu nous avais filée.
Je peux garder les cadeaux et avoir encore 29 ans?
Le cap des 30 ans, c'est pas si méchant que ça.
Tu t'es dit ça, le jour où tu les as eus?
Pourquoi, Seigneur?
Pourquoi?
On avait un deal. Tu laissais les autres vieillir, pas moi! ll n'y a que moi qui le prenne aussi mal?
Le jour de mes 30 ans, je me fendais pas la poire non plus.
Et maintenant, Chandler!
On prend tous un coup de vieux!

In the end, our corpus consisted of more than 50 million words, which is considerably larger than any other source available for spoken French language.

Calculating word frequencies

On the basis of the raw materials there are two ways to calculate word frequencies. The first consists of simply calculating the frequency of all different word forms that are encountered in the corpus. This is the easiest option, but also the least informative, as the following example in English illustrates. The word "play" can be both a verb form and a noun; the same is true for "plays." Thus, knowing the frequencies of the word forms "play" and "plays" (and "played") does not allow us to have an idea of the frequency of the word play as a verb or the word play as

a noun. Given that the processing of singular nouns is influenced by the frequency of its plural (New, Brysbaert, Segui, Ferrand, & Rastle, 2004), this is important information we are missing.

The second option is to parse the sentences, so that we know which syntactic role each word has (this is called a tagged corpus). Currently, there are many good parsers available. For our research, we opted for Cordial Analyseur 8.13, which is, to our knowledge, the best tagger for French[2] at the moment.

On the basis of the tagged corpus, we obtained a list of 313,656 entries, including compounds, first names, punctuations, and so forth. To clean this list, we used the spelling checker Aspell 0.50.3.3, the dictionary Le Grand Robert (Robert, 1996), the databases Morphalou 1.01 (Romary, Salmon-Alt, & Francopoulo, 2004), and Lexique 2.62 (New et al., 2004). The outcome of this filtering is available on the Internet as part of our project on French word characteristics (www.lexique.org).

On the basis of extensive testing, it seemed to us that the best frequency measure to derive from the subtitle corpus was one in which we gave equal weight to each of the four subcorpora (French films, English films, English television, and non-English films). In this way, the frequency estimates were based on the largest possible corpus, and we avoided that they were overly dependent on (American) movies. Therefore, we first calculated the frequency per million words for the French films, the English films, the English television series, and the non-English films. Then, the average was taken of these four measures.

THE VALIDITY OF THE NEW CORPUS AND THE NEW FREQUENCIES

There may be some concerns about the validity of the subtitle measure. After all, subtitles usually consist of a shortened and edited form of what is said. They lack all the hesitations and pronunciation errors common to spoken language usage. In addition, the topics covered in movies and television series are biased to certain topics. For instance, they more often deal with adultery and contacts with the police than is true for the average participant of a psycholinguistic experiment (although many participants watch a considerable number of these movies every week and hence are quite familiar with the topics).

We used two ways to test whether these are real concerns. The first is to see how the subtitle frequencies compare to those of existing sources (in test research, this is called congruent validity). The second is to see how well the new frequencies predict word processing times (called the criterion validity).

Congruent validity with another database of spoken frequencies

A first comparison we made was between the subtitle frequencies and the frequencies from a classical French spoken corpus the "Corpus de Référence du Français Parlé" (CRFP; Equipe DELIC, 2004). The CRFP consists of a series of interviews lasting between 10 and 30 min that took place in 40 French towns. Interviews have been directed and corrected by a senior researcher from the DELIC team. Their questions were mainly related to the participant's life or work. It consists of 1 million words based on 36 hr of speech. The interviews were held in real-life situations (at home, at work, in a shop, on the radio, etc.).

There were 5,206 entries common to our corpus and the CRFP. Because we only had access to the word form frequencies (i.e., play[noun + verb], plays[noun + verb]) from the CRFP, we calculated the corresponding frequencies for our corpus. All frequencies were coded as frequency per million words. The correlation between the subtitle and the CRFP frequencies (both log transformed) was .73, which is respectable.

To get a better idea of the origins of the discrepancies between the two lists, we looked at the entries that had a much higher or much lower frequency in one of the lists. We used the ratio of the subtitle frequency/CRFP frequency to select them. Table 1 presents the words for which the subtitle frequency was much higher than the CRFP frequency.

Two types of entries seem to pop out. The first category consists of words that are related to police matters (*tuer* [to kill], *prison* [jail], *police* [police], *armes* [weapons], *balle* [bullet]), which is in line with the fact that police-related issues figure more dominantly in movies and television series than in everyday life of most people (although many of these people watch the films and television series from our database and so do get quite a bit of exposure to these words). Finally, typical spoken expressions seem to be more frequent in the subtitle corpus than in the CRFP (*dieu* [god], *salut* [hi], *désolé* [sorry], *laissez* [let], *papa* [daddy], *docteur* [doctor], *vérité* [truth], *con* [dumb], *minute* [minute], *devrais* [should], *dormir* [to sleep], etc.). This is easily explained by the composition of the two corpora: the subtitle corpus is mostly made of people interacting in conversations, whereas the CRFP mainly comprises monologs from participants. Also notice that these words are words that are of a reasonable frequency in both lists.

The second question we wanted to ask was to know if our subtitle corpus would not miss some big lexical field compared to the more classical CRFP corpus. To do that we looked at Table 2, which shows the reverse situation, where the frequency in the CRFP corpus was much higher than the frequency in the subtitle one.

There seem to be five main categories of words that have a higher frequency in CRFP than in the subtitle corpus. The first category consists of words that are used in particular in some regions of France only, such as *pétanque* [bowls], *lyonnaise* [of Lyons], *provençal* [of Provence], *Roquefort* [Roquefort], *calandre* [a kind of Mediterranean bird], and *tarot* [tarot]. The second category consists of words related to French administrations, such as *administrations*, *municipalité* [municipality], *collectivités* [local authorities], and *spécification* [specification], and probably represent the questions asked to participants such as "What is your work?" The third category consists of onomatopoeias that are typical for spontaneous spoken language (*euh, bé, mh, hum*). The fourth category contains entries that form part of fixed expressions (*parce, abord*). These frequencies are an artefact because of differences in tokenization used in the two corpora. Finally, there is a subcategory of words that seem to be typically French and that do not figure in many of our films (*viticole* [wine producing], *charcutier* [butcher], *viticulture* [vine growing]). These would be the only words that are seriously underestimated in our list. The numbers are underrepresented because they are more represented as Arabic than Roman in the subtitle corpus. Notice, however, that many high ratios were because of very low frequencies in the subtitle corpus (e.g., *omnisports* [sports center] got a ratio of 800, because there were only 0.01 words per million in the subtitle corpus against 8 words per million in the CRFP).

Table 1. *Words for which the subtitle frequency per million words is much higher than the CRFP frequency*

Word	Word Translation	Frequencies		Ratio	Word	Word Translation	Frequencies		Ratio
		Subtitles	CRFP				Subtitles	CRFP	
Dieu	God	842.49	5	169	Arrête	Stop	453.25	23	20
Salut	Safety	486.19	4	122	Feu	Fire	234.88	12	20
Papa	Daddy	478.21	4	120	Taxi	Taxi	58.69	3	20
Tué	Killed	263.82	3	88	Tom	Tom	58.58	3	20
Tuer	Kill	342.3	6	57	Mort	Death	735.86	38	19
Désolé	Sorry	382.49	7	55	Balle	Ball	77.19	4	19
Docteur	Doctor	220.91	5	44	Emmène	Take	77.11	4	19
Laissez	Leave	262.78	6	44	Amoureux	Lover	76.74	4	19
T’	T’	4289.77	100	43	Marie	Marie	76.23	4	19
Dormir	Sleep	158.72	4	40	Excusez	Excuse	228.64	12	19
Vérité	Truth	187.93	5	38	Suivez	Follow	57.13	3	19
Ira	Will come	148.18	4	37	Attendez	Wait	228.41	12	19
Con	Idiot	145.34	4	36	Demain	Tomorrow	470.48	25	19
Prison	Prison	141.27	4	35	Secret	Secret	111.87	6	19
Fous	Madmen	203.88	6	34	Amour	Love	446.95	24	19
Ta	Your	1250.15	39	32	Hier	Yesterday	221.03	12	18
Police	Police	272.26	9	30	Allons	Let us go	495.44	27	18
Viens	Come	934.18	32	29	Bientôt	Soon	182.66	10	18
Devrais	Should	233.11	8	29	Faim	Hungry	125.86	7	18
Devoir	Duty	115.71	4	29	Fric	Cash	107.62	6	18
Minute	Minute	144.09	5	29	Te	You	3956.13	221	18
Es	Are	2359.39	85	28	Sang	Blood	300.68	17	18
Merci	Thank you	1298.82	47	28	Heureuse	Happy	87.77	5	18
Venez	Come	300.65	11	27	Viendra	Will come	52.27	3	17
Dirait	Would say	188.38	7	27	Déjeuner	Lunch	69.51	4	17
Dois	Must	884.38	33	27	Mange	Eat	103.02	6	17
Bonsoir	Good evening	159.18	6	27	Calme	Peace	255.07	15	17
Silence	Silence	104.16	4	26	Clé	Key	67.89	4	17
Folle	Mad	101.96	4	25	Pire	Worse	134.69	8	17
Maman	Mom	530.85	21	25	Colère	Anger	67.11	4	17
Toi	You	2488.11	99	25	Sexe	Sex	50.03	3	17
Visage	Face	123.97	5	25	Yeux	Eyes	312.02	19	16
Ton	Your	1755.24	71	25	Voix	Voice	129.19	8	16
Tue	Kill	122.31	5	24	Croyais	Believed	160.7	10	16
Appelez	Call	94.91	4	24	Ferai	Shall make	144.39	9	16
Mec	Fellow	250.16	11	23	Sois	Be	252.6	16	16
Coucher	Bedtime	89.19	4	22	Aurai	Shall have	110.45	7	16
Prie	Pray	244.47	11	22	Attends	Wait	473.11	30	16
Homme	Man	771.48	35	22	Serez	Will be	78.14	5	16
Fut	Was	87.72	4	22	Ferais	Would make	109.05	7	16
Victime	Victim	65.72	3	22	Sors	Go out	154.41	10	15
Bébé	Baby	171.87	8	21	Ne	Not	13314.15	863	15

Table 1 (*cont.*)

Word	Word Translation	Frequencies		Ratio	Word	Word Translation	Frequencies		Ratio
		Subtitles	CRFP				Subtitles	CRFP	
Voyons	Let us see	126.61	6	21	Sérieux	Serious	107.27	7	15
Armes	Weapons	105.06	5	21	Triste	Sad	91.86	6	15
Honneur	Honor	125.2	6	21	Ennuis	Troubles	61.18	4	15
Roi	King	164.68	8	21	Paie	Pay	60.87	4	15
Penses	Think	184.75	9	21	Cacher	Hide	60.44	4	15
Sale	Salt	121.85	6	20	Morte	Dead woman	135.79	9	15
Jolie	Beautiful	100.29	5	20	Garçon	Boy	193.14	13	15
Tes	Your	681.89	34	20	Donnez	Give	117.82	8	15

Note: CRFP, Corpus de Référence du Français Parlé (Equipe DELIC, 2004). Words are ranked as a function of the ratio of subtitle frequency/CRFP frequency (frequencies/million words).

Congruent validity with written frequencies

Another question that we can ask concerning this new corpus is to what extent it is similar to written language. To address this problem, we also compared the subtitle frequencies with written frequencies based on a corpus of 14.8 million words (New et al., 2004). These frequencies are based on 220 novels published between 1950 and 2000. Because this corpus has been tagged, we could make use of the lemma frequencies (i.e., the frequency of play[noun]), which consists of the summed frequencies of play[noun] + plays[noun]; or the frequency of play[verb], which consists of the summed frequencies of play[verb] + plays[verb] + played[verb].

We also analyzed the discrepancies for the surface frequencies but they showed essentially that the past tense is more frequent in written language than in spoken language. That's why we decided to use lemmas frequencies here.

There were 28,598 lemmas in common with a frequency larger than 0 per million. The correlation between the written and the spoken frequencies for these lemmas was .85. To get a better idea of the discrepancies, we again looked at the most extreme cases. Table 3 shows the lemmas for which the subtitle frequencies were much higher than the written frequencies.

Two types of words again seemed to be prominent. The first are words that are typical for the spoken language in everyday life (*ok*, *désolé* [sorry], *super* [great], *info* [information], *petit-déjeuner* [breakfast], *baby-sitter*, *cappuccino*, *stress*, *shampooing* [shampoo], etc.). The second are words related to (American) film themes (*astéroïde* [asteroid], *capitole* [capitol], *missile* [missile], and *federal* [fede ral]).

Table 4 lists the extremes at the other end, with much higher frequencies in the written corpus than in the subtitle corpus. A look at the words in the table indicates that none of them seem frequently used in everyday language.

Table 2. *Words for which the CRFP frequency per million words is much higher than the subtitle frequency*

Word	Word Translation	Frequencies		Ratio	Word	Word Translation	Frequencies		Ratio
		Subtitles	CRFP				Subtitles	CRFP	
Cépages	Vines	0.01	29	2900	Romane	Romanic	0.03	4	133
Lyonnaise	Of Lyons	0.01	14	1400	Velum	Awning	0.03	4	133
Embut	Coated	0.01	8	800	Spécificité	Specificity	0.07	9	129
Mygales	Mygales spiders	0.01	8	800	Approximations	Estimates	0.04	5	125
Omnisports	Sports center	0.01	8	800	Destinataires	Addressees	0.04	5	125
Hectolitres	Hectoliters	0.01	7	700	Enduits	Fillers	0.04	5	125
Quatre-vingt-dix-huit	Ninety-eight	0.02	14	700	Multimédia	Multimedia	0.08	10	125
Départementaux	Local	0.01	6	600	Soignante	Medical	0.04	5	125
Collectivités	Communities	0.01	5	500	Agglomération	Conglomeration	0.1	12	120
Piétonnes	Pedestrians	0.01	5	500	Annotations	Notes	0.05	6	120
Tandis	Whereas	0.12	52	433	Levures	Yeasts	0.05	6	120
Apposition	Apposition	0.01	4	400	Bas-relief	Bas-relief	0.07	7	100
Cloisonnement	Subdivision	0.01	4	400	Bourguignonne	Burgundian	0.04	4	100
Provençal	Provencial	0.01	4	400	Litho	Lithograph	0.04	4	100
Soumissionner	Tender	0.01	4	400	Soixante-quatorze	Seventy-four	0.03	3	100
Vernaculaire	Vernacular	0.01	4	400	Soixante-quatre	Sixty-four	0.04	4	100
Cépage	Vine	0.04	15	375	Euh	Euh	107.69	10761	100
Parce	Because	5.33	1944	365	Administrations	Administrations	0.15	13	87
Modo	Roughly	0.02	7	350	Péjoratif	Pejorative	0.07	6	86
Municipalités	Municipalities	0.03	10	333	Lamelle	Small strip	0.13	11	85
Dotations	Endowments	0.02	5	250	Feuillet	Leaf	0.06	5	83
Glacis	Glacis	0.02	5	250	Commercialisation	Marketing	0.15	12	80
Plait	Pleases	0.02	5	250	Cyclable	Cycle	0.05	4	80
Mygale	Trap-door spider	0.04	9	225	Sélectives	Selective	0.05	4	80
Faite	Made	0.15	31	207	Roquefort	Roquefort	0.24	19	79

Animations	Animations	0.07	14	200	Calandre	Calender	0.09	7	78
Asthénie	Asthenia	0.02	4	200	Mh	Mh	0.12	9	75
Départemental	Local	0.02	4	200	Taille-crayon	Pencil sharpener	0.04	3	75
Désherbants	Weedkillers	0.02	4	200	Rocade	Bypass	0.26	19	73
Deuils	Bereavements	0.03	6	200	Quatre-vingt-cinq	Eighty-five	0.07	5	71
Quatre-vingt-huit	Eighty-eight	0.02	4	200	Quatre-vingt-sept	Eighty-seven	0.07	5	71
Satiriques	Satiric	0.02	4	200	Relationnel	Relational	0.14	10	71
Sonorisation	Sound system	0.03	6	200	Dégradations	Damages	0.1	7	70
Spécification	Specification	0.02	4	200	Quatre-vingt-seize	Ninety-six	0.1	7	70
Beh	Beh	0.04	8	200	Hum	Hem	33.2	2281	69
Endogène	Endogenous	0.03	6	200	Faites	Make	1.68	104	62
Viticulture	Vine growing	0.03	6	200	Associative	Associative	0.1	6	60
Bé	Bé	0.84	166	198	Imprimeurs	Printers	0.1	6	60
Pétanque	Bowls	0.17	33	194	Visu	Display device	0.05	3	60
Arcane	Mystery	0.05	9	180	Salariale	Wage	0.17	10	59
Mouflon	Mouflon	0.04	7	175	Quatre-vingt-dix	Ninety	0.41	24	59
Plupart	Most	0.29	48	166	Dictionnaires	Dictionaries	0.31	18	58
Pédagogiques	Educational	0.05	8	160	Brocantes	Secondhand trades	0.07	4	57
Gypaète	Lammergeyer	0.07	11	157	Râteaux	Rakes	0.07	4	57
Viticole	Wine-producing	0.04	6	150	Fiscalité	Tax system	0.09	5	56
Filières	Fields of study	0.05	7	140	Polypes	Polyps	0.09	5	56
Abord	Access	0.92	123	134	Tarot	Tarot	0.36	20	56
Charcutier	Butcher	0.03	4	133	Coraux	Corals	0.38	21	55
Flûtistes	Flutists	0.03	4	133	Dix-septième	Seventeenth	0.26	14	54
Hebdos	Weekly newspapers	0.03	4	133	Solfège	Music theory	0.15	8	53

Note: CRFP, Corpus du Référence du Français Parlé (Equipe DELIC, 2004). Words are ranked as a function of the ratio CRFP frequency/subtitle frequency (frequencies/million words).

Table 3. *Words for which the subtitle frequency per million words is much higher than the written frequency*

Word	Word Translation	Part of Speech	Frequencies		Ratio	Word	Word Translation	Part of Speech	Frequencies		Ratio
			Subtitles	Books					Subtitles	Books	
Sorcière	Witch	NOM	14.36	0.07	205	Bizut	Rookie	NOM	2.29	0.07	33
Ok	Ok	ADJ	232.84	1.15	202	Toxine	Toxin	NOM	2.27	0.07	32
Thérapie	Therapy	NOM	13.48	0.07	193	Astéroïde	Asteroid	NOM	2.26	0.07	32
Petit-déjeuner	Breakfast	NOM	13.4	0.07	191	Technologie	Technology	NOM	17.39	0.54	32
Ana	Ana	NOM	26.26	0.14	188	Activation	Activation	NOM	2.25	0.07	32
Cookie	Cookie	NOM	8.19	0.07	117	Vidéo	Video	ADJ	23.44	0.74	32
Media	Media	NOM	8.06	0.07	115	Nietzschéen	Nietzschian	NOM	2.2	0.07	31
Ok	Ok	ADV	135.05	1.22	111	House	House	NOM	8.27	0.27	31
Crash	Crash	NOM	6.66	0.07	95	Sous-titrer	To subtitle	VER	6.03	0.2	30
Synchro	Synchronization	ADJ	12.93	0.14	92	Fédéral	Federal	NOM	4.21	0.14	30
Gay	Gay	NOM	11.56	0.14	83	Détecteur	Detector	NOM	7.97	0.27	30
Relax	Relaxed	NOM	5.69	0.07	81	Paranormal	Paranormal	ADJ	2.02	0.07	29
Karma	Karma	NOM	10.6	0.14	76	Capitole	Capitole	NOM	2.01	0.07	29
Colocataire	Cotenant	NOM	4.83	0.07	69	Gnocchi	Gnocchi	NOM	1.99	0.07	28
Loser	Loser	NOM	4.73	0.07	68	Mutant	Mutant	ADJ	1.99	0.07	28
Psychopathe	Psychopath	NOM	9.27	0.14	66	Cappuccino	Cappuccino	NOM	1.97	0.07	28
Bingo	Bingo	NOM	9.01	0.14	64	Superviseur	Superintendent	NOM	1.97	0.07	28
Cortex	Cortex	NOM	8.65	0.14	62	Surfer	To surf	VER	9.39	0.34	28
Scanner	Scanner	NOM	8.53	0.14	61	Maintenance	Maintenance	NOM	3.86	0.14	28
Burger	Burger	NOM	4.24	0.07	61	Junior	Junior	NOM	14.66	0.54	27
Gay	Gay	ADJ	20.17	0.34	59	Électromagnétique	Electromagnetic	ADJ	1.9	0.07	27
Portable	Mobile	ADJ	35.42	0.61	58	Propulseur	Propeller	NOM	1.88	0.07	27
Pacificateur	Peacemaker	NOM	3.87	0.07	55	Super	Great	NOM	72.78	2.77	26
Info	Info	NOM	25.5	0.47	54	Stress	Stress	NOM	10.73	0.41	26
Thérapeute	Therapist	NOM	3.63	0.07	52	Sainte	Saint	NOM	12.24	0.47	26
Vidéo	Video	NOM	21.11	0.41	51	Générateur	Generator	NOM	8.84	0.34	26

Master	Master	NOM	3.53	0.07	50	Informatique	Data processing	ADJ	5.2	0.2	26
Mémo	Memo	NOM	3.37	0.07	48	Timing	Timing	NOM	3.64	0.14	26
Jésus	Jesus	NOM	51.46	1.08	48	Logiciel	Software	NOM	3.58	0.14	26
Rap	Rap	NOM	3.29	0.07	47	Country	Country	ADJ	1.78	0.07	25
Fun	Fun	NOM	3.21	0.07	46	Homicide	Manslaughter	NOM	11.93	0.47	25
Hockey	Hockey	NOM	6.37	0.14	46	Joker	Joker	NOM	3.5	0.14	25
Vortex	Whirlpool	NOM	6.09	0.14	44	Gémeau	Gémeau	NOM	1.73	0.07	25
Conteneur	Container	NOM	2.89	0.07	41	Penny	Penny	NOM	3.46	0.14	25
Coréen	Korean	ADJ	2.83	0.07	40	Jacuzzi	Jacuzzi	NOM	3.43	0.14	25
Faxer	To fax	VER	2.83	0.07	40	Pentagone	Pentagon	NOM	4.86	0.2	24
Fax	Fax	NOM	5.52	0.14	39	Passe-la-moi	Cross it to me	NOM	1.69	0.07	24
Baby-sitter	Babysitter	NOM	7.76	0.2	39	Sonar	Sonar	NOM	1.69	0.07	24
Réessayer	Retry	VER	5.38	0.14	38	Immatriculé	Registered	ADJ	1.66	0.07	24
Investisseur	Investor	NOM	2.61	0.07	37	Tequila	Tequila	NOM	4.73	0.2	24
Pissou	Pee	NOM	5.2	0.14	37	Braiment	Braiment	NOM	7.92	0.34	23
Accro	Addict	NOM	2.54	0.07	36	Favela	Favela	NOM	1.59	0.07	23
Activé	Activated	ADJ	2.54	0.07	36	Inapproprié	Inappropriate	ADJ	1.58	0.07	23
Implant	Implant	NOM	5.08	0.14	36	Hot-dog	Hot dog	NOM	6.05	0.27	22
Cash	Cash	NOM	2.53	0.07	36	Stresser	Put under stress	VER	7.6	0.34	22
Shérif	Sheriff	NOM	46.13	1.28	36	Missile	Missile	NOM	16.52	0.74	22
Lesbienne	Lesbian	NOM	2.51	0.07	36	Échographie	Scan	NOM	1.55	0.07	22
Skate	Skate	NOM	2.47	0.07	35	Éradiquer	Eradicate	VER	1.55	0.07	22
Cutter	Cutter	NOM	2.42	0.07	35	Shampoing	Shampoo	NOM	1.55	0.07	22
C	C	NOM	67.71	1.96	35	Désolé	Sorry	ADJ	273.47	12.43	22

Note: NOM, nominative; ADJ, adjective; ADV, adverb; VER, verb. Words are ranked as a function of the ratio of subtitle frequency/written frequency (frequencies/million words).

Table 4. *Words for which the written frequency per million words is much higher than the subtitle frequency*

Word	Word Translation	Part of Speech	Frequencies		Ratio	Word	Word Translation	Part of Speech	Frequencies		Ratio
			Subtitles	Books					Subtitles	Books	
Manivelle	Crank	NOM	0.01	31.96	3196	Futaie	Forest	NOM	0.02	5.27	264
Ébrouer	Snort	VER	0.01	8.11	811	Coudrier	Hazel (tree)	NOM	0.04	10.41	260
Drifter	A kind of boat	NOM	0.05	37.5	750	Dîneur	Dinner guest	NOM	0.01	2.57	257
Gémellaire	Twin	ADJ	0.01	7.43	743	Mâchefer	Clinker	NOM	0.01	2.57	257
Obscurément	Darkly	ADV	0.01	6.96	696	Ourler	Hem	VER	0.02	5.14	257
Goguenard	Quietly ironic	ADJ	0.01	6.15	615	Auvergnat	Auvergne	NOM	0.01	2.5	250
Saccade	Jerk	NOM	0.01	6.15	615	Épineux	Thorny	NOM	0.01	2.5	250
Sénéchal	Seneschal	NOM	0.01	5.81	581	Moellon	Rubble stone	NOM	0.01	2.5	250
Cow-boy	Cowboy	NOM	0.01	5.47	547	Planchette	Small board	NOM	0.01	2.5	250
Pensivement	Thoughtfully	ADV	0.01	5.2	520	Tombereau	Tipcart	NOM	0.01	2.5	250
Ruissellement	Streaming	NOM	0.01	4.53	453	Claie	Sieve	NOM	0.01	2.43	243
Zef	Wind	NOM	0.01	4.53	453	Décacheter	Unseal	VER	0.01	2.43	243
Serpe	Billhook	NOM	0.01	4.32	432	Gaulliste	Gaullist	ADJ	0.01	2.43	243
Bungalow	Bungalow	NOM	0.03	12.84	428	Buis	Box tree	NOM	0.03	7.23	241
Avant-veille	Two days before	NOM	0.01	3.92	392	Fébrilité	Restlessness	NOM	0.01	2.36	236
Frondaison	Foliage	NOM	0.01	3.92	392	Rembrunir	Darken	VER	0.01	2.36	236
Chewing-gum	Chewing gum	NOM	0.01	3.78	378	Remugle	Stale smell	NOM	0.01	2.36	236
Précautionneusement	Carefully	ADV	0.01	3.72	372	Bruissant	Rustling	ADJ	0.02	4.66	233
Brame	Squall	NOM	0.01	3.58	358	Dépoli	Frosted	ADJ	0.01	2.3	230
Tonnelle	Arbour	NOM	0.01	3.51	351	Saillir	Cover	VER	0.03	6.82	227
Cantonade	Speak off	NOM	0.01	3.45	345	Carrée	Square	NOM	0.02	4.53	227
Confusément	Confusedly	ADV	0.03	10	333	Brigadier-chef	Corporal-leader	NOM	0.01	2.23	223
Moleskine	Imitation leather	NOM	0.01	3.31	331	Ouaté	Cotton	ADJ	0.01	2.23	223
Alsacien	Alsatian	NOM	0.01	3.24	324	Volute	Volute	NOM	0.03	6.69	223
Derechef	Once more	ADV	0.01	3.24	324	Rasséréner	Reassure	VER	0.02	4.39	220
Nervure	Nervure	NOM	0.01	3.24	324	Ahaner	Pant	VER	0.01	2.16	216
Prie-dieu	Prie-dieu	NOM	0.01	3.24	324	Épisodique	Occasional	ADJ	0.01	2.16	216

Casemate	Bunker	NOM	0.01	3.18	318	Négligemment	Untidily	ADV	0.04	8.45	211
Complaisamment	Accommodatingly	ADV	0.01	3.18	318	Charentais	Charentais	NOM	0.01	2.09	209
Voluptueusement	Sensually	ADV	0.01	3.11	311	Nirvâna	Nirvana	NOM	0.01	2.09	209
Bâtardise	Illegitimacy	NOM	0.01	3.04	304	Bonhomie	Gentleness	NOM	0.02	4.12	206
Noirâtre	Blackish	ADJ	0.02	6.08	304	Croisillon	Crosspiece	NOM	0.01	2.03	203
Paresseusement	Lazily	ADV	0.01	2.97	297	Dentellière	Lacemaker	NOM	0.01	2.03	203
Entr'ouvert	Half-opened	ADJ	0.01	2.91	291	Déprendre	Get rid	VER	0.01	2.03	203
Louvet	Dun	ADJ	0.01	2.91	291	Gangue	Gangue	NOM	0.01	2.03	203
Ondoyer	To wave	VER	0.01	2.84	284	Iriser	Make Iridescent	VER	0.01	2.03	203
Cordelier	Cordelier	NOM	0.01	2.77	277	Aménité	Friendliness	NOM	0.01	1.96	196
Commissure	Corner	NOM	0.02	5.41	271	Arbitraire	Arbitrary power	NOM	0.01	1.96	196
Lorgnon	Lorgnette	NOM	0.02	5.41	271	Bruni	Tanned	ADJ	0.01	1.96	196
Claire-voie	Fence	NOM	0.01	2.7	270	Constituant	Constituent	ADJ	0.02	3.92	196
Déférent	Deferential	ADJ	0.01	2.7	270	Effranger	Fringe	VER	0.01	1.96	196
Éberlué	Astounded	ADJ	0.01	2.7	270	Épandre	Spread	VER	0.01	1.96	196
Rigolard	Joker	ADJ	0.01	2.7	270	Fondrière	Rut	NOM	0.01	1.96	196
Zanzi	Dice game	NOM	0.03	8.04	268	Râble	Back	NOM	0.01	1.96	196
Haut-commissaire	High-commissioner	NOM	0.03	7.97	266	Sourcilleux	Punctilious	ADJ	0.01	1.96	196
Cagna	Hot	NOM	0.01	2.64	264	Stridence	Strident	NOM	0.01	1.96	196
De guingois	Askew	ADV	0.01	2.64	264	Dolmen	Dolmen	NOM	0.01	1.89	189
Émaillé	Enameed	ADJ	0.01	2.64	264	Fourrier	Harbinger	NOM	0.01	1.89	189
Goulée	Gulp	NOM	0.01	2.64	264	Graminée	Grass	NOM	0.01	1.89	189
Supplicié	Torture victim	NOM	0.01	2.64	264	Grenu	Grainy	ADJ	0.01	1.89	189

Note: NOM, nominative; VER, verb; ADJ, adjective; ADV, adverb. Words are ranked as a function of the ratio written frequency/subtitle frequency (frequencies/million words).

During these four analyses, we have seen that our subtitle corpus seems to provide quite good estimates of spoken frequencies. It represents frequently heard or produced words that are not well represented in "classical" corpora. Furthermore, it does not seem to neglect very frequent lexical fields.

Criterion validity with lexical decision times

In addition to the descriptive analyses presented above, we wanted to find a more objective test to examine the psychological validity of our corpus. The lexical decision task is a very common task used in psycholinguistics to study word processing. Participants have to decide as fast as possible if a stimulus is word or a nonword. An interesting property of the lexical decision task is that the strongest predictor of the reaction times is the word frequency. We computed the correlation coefficient between several frequency measures and the lexical decision times obtained in two recent experiments. Because the CRFP does not have lemma frequencies, we limited our analyses to word surface frequencies (as has been done in English as well; see Baayen et al., 2006; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004).[3]

The first experiment examined the effects of word frequency and age of acquisition on word processing in French (Bonin, Chalard, Méot, & Fayol, 2001; Experiment 3). In this experiment, 30 participants decided for 468 letter strings whether they formed an existing French word (234 stimuli) or not (234 other stimuli). All words were nouns representing concrete things (e.g., bee, needle). Among the 234 words, only 91 were found in the CRFP.

We used four different frequency measures: the CRFP frequencies, the subtitle frequencies restricted to the French movies, the written corpus described above, and our subtitle frequencies. We added 1 to each frequency and then took log 10. In addition, because the relationship between log frequency and reaction time (RT) is not completely linear (Baayen, Feldman, & Schreuder, 2006), we added the square of the log frequency as a second predictor variable in a multiple regression analysis. The number of syllables and letters were also entered in the multiple regressions as words were varying from 3 to 12 letters and from one to four syllables. We applied the logarithmic transformation to the RT to eliminate most of the skewness of the distribution of reaction times (Baayen et al., 2006).

Table 5 lists the percentage of variance explained in the lexical decision times (adjusted R^2) by each of the frequency measures. From this analysis it is clear that the CRFP did much worse than the other two corpora. This was partly because of the fact that for this corpus the log 10 frequency was 0 for nearly 150 of the stimulus words (because the word was not present in the corpus). Another reason, however, was related to the quality of the frequency measures. When the analysis was limited to the 91 words for which we had a CRFP frequency, the percentage of variance accounted for was still substantially smaller than that accounted for by the book and the subtitle frequencies and now was less than 10%, probably because the range of frequencies was too restricted. The CRFP corpus is much

Table 5. *Effects of different frequencies on Bonin's lexical decision reaction times*

Model	Adjusted R^2
Syllables (.) + letters (*) + log CRFP (***) + (log CRFP)^2 (*ns*)	30.1***
Syllables (.) + letters (*) + log French (***) + (log French)^2 (***)	43.3***
Syllables (*ns*) + letters (**) + log books (***) + (log books)^2 (***)	46.3***
Syllables (*ns*) + letters (.) + log subtitles (***) + (log subtitles)^2 (***)	49.7***

Note: CRFP, Corpus du Référence du Français Parlé (Equipe DELIC, 2004).
$^*p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 6. *Effects of different frequencies on Bonin's lexical decision reaction times*

Model	Adjusted R^2
Syllables + letters (**) + log books (***) + (log books)^2 (***)	46.3***
Syllables + letters (**) + log books (***) + (log books)^2 (***) + log (books/subtitles) (***)	50.2***
Syllables + letters (.) + log subtitles (***) + (log subtitles)^2 (***)	49.7***
Syllables + letters (.) + log subtitles (***) + (log subtitles)^2 (***) + Log (books/subtitles) (*ns*)	49.9***

$^{**}p < .01$. $^{***}p < .001$.

less diversified because the same questions were used in each interview (Tell us about you life, tell us about your work).

To find out how much the subtitle frequencies added to the book frequencies, we entered the variable log(frequency subtitles/log frequency books) as a fifth variable to the regression analyses. This extra variable gives us an idea of how much variance is explained by the relative frequency of the words in the subtitle corpus versus the book corpus (Table 6).

The second lexical decision experiment was a purpose-built experiment in which we presented a random sample of 240 two-syllable nouns with high and low frequencies from the written corpus. Seventeen participants took part. Error responses were discarded from the analysis and response times more than 2 standard deviations above or below the mean were discarded. We removed one item because of an experimental problem (*bistro*).

The analyses presented in Tables 6 and 7 show that the subtitle frequency measure is at least as good as the existing book frequency measure to account for differences in lexical decision times. Further large-scale studies comparable to The English Lexicon Project (Balota et al., in press), in which lexical decision data have been collected for 44,000 English words, are planned for French words. This will enable us to see whether the hint of better performance is confirmed when all French nouns are entered into the regression analyses.

Table 7. *Effects of different frequencies on our lexical decision reaction times*

Model	Adjusted R^2
Log CRFP (***) + (log CRFP)^2 (*)	33.2***
Log French (***) + (log French)^2 (*ns*)	43.9***
Log books (***) + (log books)^2 (*ns*)	44.5***
Log books (***) + (log books)^2 (*ns*) + log (books/subtitles) (***)	47.9***
Log subtitles (***) + (log subtitles)^2 (.)	46***
Log subtitles (***) + (log subtitles)^2 (*ns*) + log (books/subtitles) (**)	48.1***

Note: CRFP, Corpus du Référence du Français Parlé (Equipe DELIC, 2004).
*$p < .05$. **$p < .01$. ***$p < .001$.

CONCLUSIONS

In this article we have described a new way to obtain a corpus of social interactions in a matter of weeks, simply by making use of the availability of files with film subtitles on the Internet. Given the rate with which movies and television series are subtitled today, we foresee that the choice of materials will further increase in the coming years, which will open the possibility to make the sampled materials more representative for the language register aimed at. In the current corpus, we do have a slight bias toward American police-related matters but, as mentioned previously, these are words that people do hear quite often as they watch TV. Even so, the quality of the results surprised us. Apart from the foreseen biases (too much police matters, not enough words that refer to typical French instances), the discrepancies between the subtitle corpus and the other databases we checked intuitively turned out to be in favor of the subtitle corpus. This was confirmed when we correlated the frequencies to lexical decision times obtained in two typical experiments that addressed the word frequency issue.

In summary, the current subtitle frequency measure seems to be a useful addition to the existing spoken and written frequencies (e.g., to match stimulus materials on frequency). There is a huge advantage, in particular, related to spoken frequency measures. This kind of corpus can easily be collected without the need of manual transcription, so that it is feasible for all languages that do not yet have a spoken corpus. The corpus can also regularly be updated and further optimized as new movies are released everyday.

ACKNOWLEDGMENTS

This research was supported by Technolangue. We thank Agnès Bontemps for the idea to use movie subtitles for making a corpus and Magali Boibeux for helping to build and run the lexical decision presented here.

NOTES

1. We removed subtitles coming from Asian countries. They had an abnormally low number of word types compared to the other subcorpora. We suspect that this subcorpora has too many specific movies (e.g., mangas).

2. Despite cordial good performances, some errors remain. We corrected some of them.
3. The variance explained by lemma frequencies is 1–5% higher. This will be covered in future work.

REFERENCES

Baayen, H., Feldman, L., & Schreuder, B. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 55*, 290–313.

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database* (Release 2) [CD-ROM]. Philadelphia, PA: University of Pennsylvania, Linguistic Data Consortium.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*, 283–316.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., et al. (in press). The English Lexicon Project. *Behavior Research Method.*

Blair, I. V., Urland, G. R., & Ma, J. E. (2002). Using Internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, & Computers, 34*, 286–290.

Bonin, P., Chalard, M., Méot, A., & Fayol, M. (2001). Age-of-acquisition and word frequency in the lexical decision task: Further evidence from the French language. *Current Psychology of Cognition, 20*, 401–443.

Desmet, T., De Baecke, C., Drieghe, D., Brysbaert, M., & Vonk, W. (2006). Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Language and Cognitive Processes, 21*, 453–485.

Equipe DELIC. (2004). Présentation du Corpus de référence du Français parlé. *Recherches sur le Français Parlé, 18*, 11–42. Also available at http://www.up.univ-mrs.fr/veronis/pdf/2004-presentation-crfp.pdf

Grondelaers, S., Deygers, K., van Aken, H., van den Heede, V., & Speelman, D. (2000). Het ConDiv-corpus geschreven Nederlands. *Nederlandse Taalkunde, 5*, 356–363.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers, 36*, 516–524.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE, *L'Année Pschologique, 101*, 447–462.

Robert, P. (1996). Le grand Robert électronique [Software]. Havas Interactive. Accessed at http://www.havas.com

Romary, L., Salmon-Alt, S., & Francopoulo, G. (2004). *Standards going concrete: From LMF to Morphalou.* Unpublished manuscript, Coling, Geneva, Switzerland, Workshop on Electronic Dictionaries.