C-PROM.
An Annotated Corpus for French Prominence Study

M. Avanzi[1, 2], *A.C. Simon*[3], *J.-P. Goldman*[3, 4] *& A. Auchlin*[4]

[1]Chaire de linguistique française, Neuchâtel University; [2]MoDyCo, Université Paris Ouest Nanterre
[3]Institut Langage & Communication / Valibel Discours & Variation, Université catholique de Louvain; [4]Department of Linguistic, Geneva University

mathieu.avanzi@unine.ch, anne-catherine.simon@uclouvain.be,
jeanphilippegoldman@gmail.com, antoine.auchlin@unige.ch

Abstract

This paper presents C-PROM, an annotated corpus for French prominence studies. The corpus, including different regional varieties of French (Belgian, Swiss and metropolitan French) and various discourse-genres (from oral reading to spontaneous conversations) for a total duration of 70 minutes, was annotated by two phonetics experts. The two experts in charge of the coding followed a strict protocol, which takes into account both the previous mistakes encountered by prior research into prominence detection in French and elements of the methodology followed by scholars working on other languages. We conclude by discussing the average consistency between the two transcribers. The results obtained are quite encouraging, as the F-measure between the two annotators reaches 82.8%, and the kappa-score 0.86.

Index Terms: corpus, spontaneous French, prominence, discourse genre.

1. Introduction

Over the last decade, many studies have dealt with the prosodic annotation of spontaneous speech. Their goal was not always to train automatic systems; they did not concern only prominence, and, finally, were not specifically dedicated to French. Due to considerations of space, we are unable to present this previous work here in its entirety, and therefore refer the interested reader to [1], [2], [3], [4], [5], [6] and references therein.

The earliest empirical approaches to French prominence started within the PFC (*Phonologie du Français Contemporain*) project [7]. This project aims at building a large annotated database of French as it is spoken all around the world. Transcriptions, aligned with the audio signal with the Praat program [8] were annotated in order to allow phonological studies on French schwa and sandhi phenomena. In 2002, the promoters discussed the need to annotate a wider range of prosodic phenomena. The initial discussions and experiments conducted in this framework are summarized in section 2. Section 3 is devoted to the presentation of the protocol (methods, set of symbols) drawn up for prominence annotation and to the annotator-agreement score between the expert transcribers. Section 4 briefly describes the final design of the annotated corpus. The conclusion discusses some investigations made possible by C-PROM.

2. Previous Work

2.1. The first experiment

At the beginning, certain basic principles rinciples for the constitution of an annotation protocol for prominence in the PFC corpora were laid down ([9] and [10]). The coding procedure (i) had to be independent of any theoretical framework; (ii) should rely on perceptual judgments; (iii) is reproducible by non-experts; (iv) could allow for studies on every domain of prosody (accentuation, intonation, rhythm, and so on). In order to build a draft protocol, a pilot experiment was conducted by [11]. Seven phonetics experts were asked to annotate perceived syllabic prominences in a 3-minute spontaneous speech recording of a male speaker, without other instruction. It was expected that agreement would be fairly encouraging, since prominence has to correspond with accent, and the accentuation rules of French were well-known by the experts. Surprisingly, however, among the 165 syllables uttered, the proportion marked as prominent varied from 19% to 49%, that is to say the inter-rater agreement was poorer than expected This led [12] to conclude that the annotation of prominence in French was “more an art than a scientific practice”.

2.2. Acoustic expertise

[13] used the results of this pilot study to conduct experiments on the data. The aim was to evaluate the robustness of two acoustic parameters (f0 and duration) for automatic prominence detection and to suggest measures to evaluate the annotators’ performances. The authors concluded that melodicity was better correlated with inter-rater agreement (the higher the f0 values, the better the inter-rater agreement is), whereas a similar correlation with duration values was not observed. Thus, over a determined duration threshold (between 175 and 200 ms), the proportion inverts and agreement does not increase, but decreases. This is due to the fact that beyond a certain threshold, lengthening is no longer perceived as a prominence clue, but as a mark of hesitation. This confirmed the fact that humans, even when experts, do not share a definition of prominence.

2.3. Consequences

From these initial experiments, the following lessons have been learnt. Firstly, the low rate of agreement comes from the lack of accuracy in the coding instructions. In order to obtain a better inter-rater agreement, the notion of prominence has to be carefully defined, and not conflated with the notion of "stress" (which is a phonological notion implying linguistic knowledge). Secondly, there is a need for defining a context-window for prominence identification, to avoid ending up with large parts of the sound signal without any prominence detection. Furthermore, the above-mentioned authors agreed that visualization of the signal can be helpful. Finally, the study of the acoustic correlates of perceived prominences showed that while f0 was a good cue for automatic identification, so was duration, provided that hesitation marks had a specific annotation, to avoid biasing the relative duration calculations.

3. A Methodology for Perceptual Prominence Annotation

These early studies conducted within the PFC project were undertaken by the authors of the present paper. They led *in fine* to the construction of a multi-genre and multi-speaker corpus, called C-PROM, annotated for French prominence study by two phonetics experts (see Table 3). The C-PROM corpus has been developed with the specific purpose of building an open data-base to train algorithms for semi-automatic prominence detection.

3.1. Data preparation

A 70-minute corpus, sampled for different genres (see Table 3), was automatically segmented into phones, syllables and orthographic words using Praat [8] and the Easyalign script [14]. All the transcriptions were manually checked.

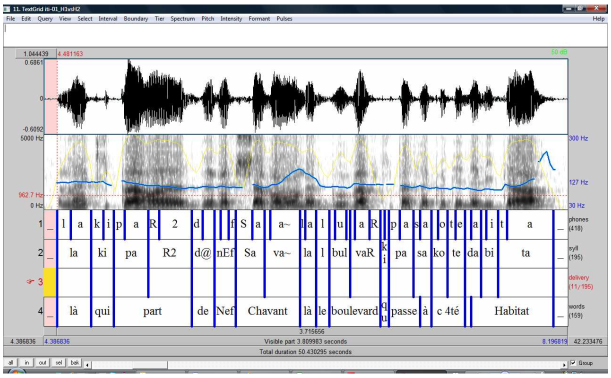
11. TextGrid iti-01_H1vsH2
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
1.044439
4.481163
50 dB
0.6861
0
-0.6092
5000 Hz
962.7 Hz
0 Hz
300 Hz
127 Hz
30 Hz
1 _ l a k i p a R 2 d f S a a~ l a l u a R p a s a o t e a i t a _
2 _ la ki pa R2 d@ nEf Sa va~ la l bul vaR ki pa sa ko te da bi ta _
3
4 _ là qui part de Nef Chavant là le boulevard qu passe à c 4té Habitat _
phones (418)
syll (195)
delivery (11/195)
words (159)
3.715656
4.386836
4.386836
Visible part 3.809983 seconds
8.196819
42.233476
Total duration 50.430295 seconds
all in out sel bak
Group

Figure 1: *Praat screen shot of the utterance: "là qui part de nef Chavant là le boulevard qui passé à côté d'Habitat" [mp-1]. Annotation tiers are, from top to bottom: phones, syllables (both in SAMPA), delivery and word*

Two annotators (among the authors of this paper) annotated the whole corpus following the protocol described in the next section.

3.2. Protocol

The C-PROM coding-protocol takes into account the "errors" encountered in the first PFC studies, and applies some recommendations made by the supervisors of the spoken Dutch corpus coding-protocol [3].

In practice, each annotator starts from an empty annotation tier duplicated from the syllabic tier (the "delivery" tier in Figure 1), and fills every interval with the symbols described in Table 1. Annotation in conducted by listening to a stretch of speech of 3 to 5 seconds, no more than three times (over-listening results in over-coding). As the annotation of prominences relies on auditory perception of salience and not on the visual analysis of acoustic parameters (f0 movements, for example), visualization of the signal was restricted to problematic cases.

The first class of symbols is for annotating prominent syllables. Three symbols can be used (NP, p and P). The distinction between "p" and "P" is heuristic: it forces the transcribers to develop more accurate listening and avoids marking only the strongest prominences. It also avoids the use of an indecision marker such as "?". During the comparison of the two manual annotations, these two categories were merged.

Table 1: *Annotation symbols*

1. Prominence labeling	
P	strongly prominent syllable
p	weakly prominent syllable
NP	non prominent syllable
2. *Delivery* labeling	
z	lengthening connected with a hesitation
@	post-tonic syllabic schwa (as in "c'est dinguE" [sEde~g@])
$	unaccented post-tonic syllables (appendice)
3. Others	
%	junk (noise, laugh, cough, etc.)
*	breath
_	silence

The delivery labeling is used for singling out syllables which have specific properties likely to hamper automatic prominence identification. The "z" symbol is for extra-lengthened syllables marking hesitation. Their length can disturb the calculation of relative duration, as shown by [9]. The marking of hesitation also serves to avoid false automatic detection of prominence. Since hesitations are often followed by a silent pause, and since silent pause is often considered a strong clue for prominence detection (boundaries and prominence being merged in French ([15] and [16]), it could introduce mistakes in the automatic detection. Post-tonic schwa (@) and appendices ($) are considered non prominent, but they are specifically annotated because they introduce irregularity in the final-accent system in French: post-focal syllables being problematic with regards to f0 [17], the status of schwa as a syllabic nucleus is controversial [18]. The number of symbols in this "delivery" class can also be explained by the perspective of a semi-automatic identification of these specific prosodic phenomena. The last category of symbols is for annotating "silence" and the like. It contains silent pauses (resulting from the semi-automatic alignment), audible breaths and "junk", *i.e.* part of the recording that could not be transcribed (noise, laughter, coughing, overlapping, etc.). These could interfere with the automatic processing of the signal.

3.3. Annotation task

It took the two transcribers nearly a year to annotate the whole corpus (the annotation was done between fall 2007 and fall 2008). First, they jointly annotated in a practice session a small stretch of speech (a 1-minute long map task). They then independently annotated the whole corpus, genre by genre. Each time the coding of one sub-corpus (a set of samples from one genre) had been completed, a comparison tier was automatically generated. It revealed the full range of disagreement in coding; the agreed codes were left untouched (see COMPARE-tier in figure 2 below). Then the two experts discussed divergences during a joint session. Due to lack of space, we cannot address here how the coding divergences were resolved. See [6] and [19] for a systematic analysis of the inter-transcriber disagreement cases.

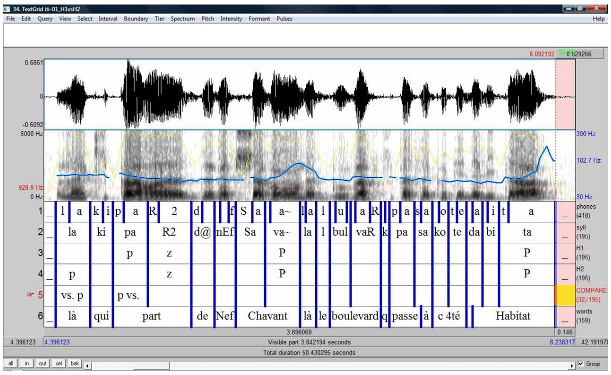
34. TextGrid iti-01_H1vsH2
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
8.092192 0.029266
0.6861
0
-0.6092
5000 Hz
928.9 Hz
0 Hz
300 Hz
182.7 Hz
30 Hz
1 _ l a k i p a R 2 d f S a a~ l a l u a R p a s a o t e a i t a _ phones (418)
2 _ la ki pa R2 d@ nEf Sa va~ la l bul vaR k pa sa ko te da bi ta _ syll (195)
3 _ p z P P _ H1 (195)
4 _ p z P P _ H2 (195)
5 vs. p p vs. COMPARE (32/195)
6 _ là qui part de Nef Chavant là le boulevard q passe à c 4té Habitat _ words (159)
3.696069 0.146
4.396123 4.396123 Visible part 3.842194 seconds 8.238317 42.191978
Total duration 50.430295 seconds
all in out sel bak
Group

Figure 2: *Praat screen shot of the same utterance as in figure 1, with the H1, H2 and COMPARE tiers.*

3.4. Evaluation of the inter-annotator agreement

The COMPARE-tier was used to estimate the degree of agreement between the two annotators. Table 2 gives the inter-transcriber agreement for each recording of the corpus, based on a count of the intervals involving a conflict between a “P/p” symbol versus anything else:

Table 2: *Inter-annotator consistency: name of the file (see Table 3), n00/n11: syllables annotated 0 (non prominent) and p (prominent) by both annotators, n10/n01: syllables annotated p by one transcriber, 0 by the other; R: recall; P: precision; F: F-measure, K: Kappa-score*

file	n00	n11	n10	n01	R	P	F	K
lec-be	495	87	22	5	94.6	79.8	86.6	0.91
lec-ch	390	156	26	33	82.5	85.7	84.1	0.83
lec-fr	469	120	32	2	98.4	78.9	87.6	0.91
pol-be	225	139	21	26	84.2	86.9	85.5	0.8
pol-ch	735	162	20	97	62.5	89	73.5	0.77
pol-fr	519	143	13	71	66.8	91.7	77.3	0.78
jpa-be	924	259	58	29	89.9	81.7	85.6	0.89
jpa-ch	587	172	29	50	77.5	85.6	81.3	0.83
jpa-fr	668	197	25	46	81.1	88.7	84.7	0.87
cnf-be	727	182	26	84	68.4	87.5	76.8	0.8
cnf-ch	586	186	40	59	75.9	82.3	79	0.8
cnf-fr	776	239	39	37	86.6	86	86.3	0.89
int-be	722	224	48	44	83.6	82.4	83	0.85
int-fr	899	262	56	72	78.4	82.4	80.4	0.83
iti-01	124	26	9	2	92.9	74.3	82.5	0.78
iti-02	84	41	1	1	97.6	97.6	97.6	0.95
iti-03	261	83	22	15	84.7	79	81.8	0.81
iti-04	547	154	15	21	88	91.1	89.5	0.92
iti-06	281	87	8	7	92.6	91.6	92.1	0.92
iti-07	94	24	7	1	96	77.4	85.7	0.77
nar-be	582	190	31	36	84.1	86	85	0.86
nar-ch	573	172	16	65	72.6	91.5	80.9	0.82
nar-fr	468	149	25	40	78.8	85.6	82.1	0.83
total	**11736**	**3454**	**589**	**843**	**80.4**	**85.4**	**82.8**	**0.86**

The inter-transcriber agreement for prominence annotation was quantified by means of Cohen’s kappa coefficient [20], and evaluated for the whole corpus at 0.86 (the best performance is for a map task file (0.95 for iti-02), the worst for pol-ch and iti-07, with 0.77). The f-measure calculation (harmonic average between precision and recall [21]) indicates an inter-transcriber agreement of 82.8% (recall: 80.4; precision: 85.4). Similar to [3], we would like to highlight that such a good agreement-score is certainly due to the use of a standard protocol and joint training. We would therefore argue that these good performances are strong evidence against the notion that French prominence transcription is more an art than a scientific practice ([12], [16]). Future experiments involving more transcribers, both experts and non experts, should allow for further confirmation of this fact.

4. Description of the corpus

A consensual annotation emerged from the discussion on the COMPARE-tier, and it was considered as a reference annotation for prominence analysis. Table 3 (next page) details the composition of the C-PROM corpus: it includes 28 speakers (12 females, 16 males) and amounts to 17,778 syllables, from which 805 (4.5%) were excluded via the delivery tier, 4,570 were annotated as prominent (25.7%) and 12,403 (69.7%) were non prominent. The corpus is composed of a set of recordings sampled in seven genres, ranging from high to low degrees of formality: Read Speech (LEC), Political Speeches (POL), Conferences (CNF), News Broadcasts (JPA), Radio Interviews (INT), Map Tasks (ITI) and Life Stories (NAR). Except for the ITI and INT recordings, all the discourse genres collections comprise 3-minute recordings, performed by a native speaker from Belgium (BE), Switzerland (CH) and metropolitan France (FR). All the speakers in the corpus speak a highly standard French.

5. Conclusion

The aim of this paper was to present the C-PROM corpus. We presented previous work which motivated its constitution, and the methodology followed to build it. Although the methodology remains to be tested on larger parts of corpora, including more annotators, this first experiment nevertheless produced encouraging results. Sub-parts of the corpus have already been used to train different automatic prominence detection algorithms ([6], [22], [23], [24]). It also resulted in studies on the automatic estimation of discourse genres based on prosodic features ([25], [26]). We finally hope that our corpus will facilitate comparisons between different studies, which were not previously possible because of the unavailability of shared data. An online version of C-PROM will be available shortly.

6. Acknowledgments

This research was supported by The Swiss National Science Foundation (grants n° PBNEP1-127788 and n°100012-113726/1, Neuchâtel University) and by the Programme Wist2 Convention n°616422, financed by Région wallonne (Belgique), Projet *EXPRESSIVE*.

7. References

[1] Pickering, B., Williams, B. & G. Knowles G., “Analysis of Transcriber Differences in the SEC, *in* Knowles, G. *et al.* (eds), Working with Speech: perspectives on research into the Lancaster/IBM Spoken English Corpus. London/New-York: Longman, 59-105, 1996.

[2] Grabe, E., Post, B., & F. Nolan, “Modelling Intonational Variation in English: The IViE System”, *in* Puppel, S. & G. Demenko (eds), Prosody 2000, Poznan, Adam Mickiewicz University, 51-58, 2001.

[3] Buhmann, J., Caspers, J., van Heuven, V, Hoekstra, H., Martens, J.-P. & M. Swerts, “Annotation of Prominent Words, Prosodic Boundaries and Segmental Lengthening by Non Expert

Transcribers in the Spoken Dutch Corpus", LREC Processing, 779-785, 2002.

[4] Tamburini, F. & C. Caini, "An Automatic System for Detecting Prosodic Prominence in American English Continuous Speech", International Journal of Speech Technology, 8, 33-44, 2005.

[5] Rosenberg, A & J. Hirschberg, "Detecting pitch accent using pitch corrected energy-based predictors," Interspeech'07, 2777–2780, 2007.

[6] Avanzi, M., Goldman, J.-P. Lacheret-Dujour, A. Simon, A.-C. & A. Auchlin, "Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé", Cahiers of French Language Studies, vol. 13, no. 2, pp. 2–30, 2007.

[7] Durand, J., Laks, B. & C. Lyche. "La phonologie du français contemporain: usages, variétés et structure", *in* Pusch, C. & W. Raible (eds.), Romance Corpus Linguistics - Corpora and Spoken Language, Tübigen, Gunter Narr Verlag, 93-106, 2002.

[8] Boersma, P. & Weenink, D. Praat: doing phonetics by computer (Version 5.1). www.praat.org, 2009.

[9] Lacheret-Dujour, A., Lyche, Ch. & M. Morel, "Pour une transcription prosodique normalisée au sein du projet PFC (phonologie du français contemporain): champ d'action et limites", Actes des 25èmes journées d'étude sur la parole, Fès, Maroc, 2004.

[10] Poiré, F., "La codification de la prosodie", Bulletin PFC, 4, 89-98, 2005.

[11] Poiré, P. "La perception des proéminences et le codage prosodique", Bulletin PFC, 6, 69-79, 2006.

[12] Martin, Ph. 2006. « La transcription des proéminences accentuelles : mission impossible ? », Bulletin PFC, 6, 81-87.

[13] Morel M., Lacheret-Dujour, A. Lyche, Ch., Morel, M. & F. Poiré, (2006). "Vous avez dit proéminence?", Actes des 26èmes journées d'étude sur la parole, Dinar, Maroc, 2006.

[14] Goldman, J.-P. "EasyAlign: a semi-automatic phonetic alignment tool under Praat", http://latlcui.unige.ch/phonetique, 2008.

[15] Rossi, M. "Le français, langue sans accent ?", Studia Phonetica, 13-51, 1979.

[16] Vaissière, J. "Rhythm, accentuation and final lenghtening in French", in Sundberg et al. (eds), Music, Language Speech and Brain, Macmillan Press, 108-120, 1991.

[17] Mertens, P. "A Predictive Approach to the Analysis of Intonation in Discourse in French", in Kawaguchi, Y. et al. (eds), Prosody and Syntax. Amsterdam: Benjamins, 64-101, 2006.

[18] Dell, F. Generative Phonology and French Phonology, Cambridge, Cambridge University Press, 1980.

[19] Goldman, J.-P. *et al*. "Prominence perception and accent detection in French. A corpus-based account", Paper submitted to Speech Prosody 2010.

[20] Cohen, J., "A coefficient of agreement for nominal scales", Educational and Psychological Measurement, 20:37-46, 1960.

[21] van Rijsbergen, C.J. Information Retrieval, Butterworths, London, 1979.

[22] Goldman, J.-P.; Avanzi, M.; Lacheret-Dujour, A.; Simon, A.-C.; Auchlin, A.. A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French. In Proceedings of Interspeech'07, Antwerp, Belgium, August 27-31. 98-101, 2007.

[23] Avanzi, M. Lacheret-Dujour, A. & Victorri, B. "ANALOR. A Tool for Semi-Automatic Annotation of French Prosodic Structure", *Proceedings of Speech Prosody'08*, 119-122, 2008.

[24] Obin, N., Goldman, J.-P., Avanzi, M. & Lacheret-Dujour, A. « Comparaison de trois outils de détection semi-automatique des proéminences dans les corpus de français parlé », Actes des 22èmes JEP, Avignon, 2008.

[25] Obin, N., Lacheret-Dujour, A., Veaux, C., Rodet, X & A.C. Simon, "A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features", Interspeech, Brisbane, 2008.

[26] Simon, A.C., Auchlin, A., Avanzi, M. & Goldman, J.-P. "Les phonostyles. Une description prosodique des styles de parole en français", Actes du colloque Les voix du français : usages et représentations, Oxford, 2008.

Table 3. *C-Prom design. With, from left to right: discourse genre, sub-corpus code, speaker (male or female), duration, syllables, Non-Prominent Syllables, Prominent Syllables, Total syll. val., elongations (Z), post-focus syllables ($) and post-tonic schwas (@). Sub-totals are in bold. Total of the whole corpus is given in the last grey line of the table.*

	Genre	File	Speakers	Duration (sec.)	Nb. syll.	N-Prom syll.	Prom syll.	Tot. val._syll.	Z	$	@	Tot. del._syll
+formal	**Read Speech**	LEC-BE	M	114	606	492	111	603	0	0	3	3
		LEC-CH	M	137	606	403	196	599	0	0	7	7
		LEC-FR	M	150	618	462	153	615	0	0	3	3
		total	**3M**	**401**	**1830**	**1357**	**460**	**1817**	**0**	**0**	**13**	**13**
	Political Speech	POL-BE	M	188	420	246	160	406	0	0	14	14
		POL-CH	F	230	1011	753	257	1010	0	0	1	1
		POL-FR	M	217	743	533	209	742	1	0	0	1
		total	**2M/1F**	**635**	**2174**	**1532**	**626**	**2158**	**1**	**0**	**15**	**16**
	News Broadcast	JPA-BE	M	253	1315	963	312	1275	24	0	16	40
		JPA-CH	F	180	879	610	242	852	12	0	15	27
		JPA-FR	M	188	971	683	256	939	19	0	13	32
		total	**2M/1F**	**621**	**3165**	**2256**	**810**	**3066**	**55**	**0**	**44**	**99**
	Conference	CNF-BE	F	244	1066	776	250	1026	35	0	5	40
		CNF-CH	M	219	950	627	260	887	55	7	1	63
		CNF-FR	F	224	1117	798	301	1099	12	1	5	18
		total	**1M/2F**	**687**	**3133**	**2201**	**811**	**3012**	**102**	**8**	**11**	**121**
	Radio Interview	INT-BE	2F	296	1189	769	317	1086	56	32	15	103
		INT-FR	2M	331	1402	996	346	1342	21	30	9	60
		total	**2M/2F**	**627**	**2591**	**1765**	**663**	**2428**	**77**	**62**	**24**	**163**
	Map Task	ITI-01	M	50	172	117	36	153	15	2	2	19
		ITI-02	2M	47	142	82	42	124	17	1	0	18
		ITI-03	F	100	419	270	104	374	30	11	4	45
		ITI-04	2F	204	790	538	197	735	50	1	4	55
		ITI-05	M	28	128	92	27	119	8	0	1	9
		ITI-06	1M/1F	128	431	298	106	404	17	8	2	27
		ITI-07	M	33	140	98	30	128	10	2	0	12
- formal		**total**	**6M/3F**	**590**	**2222**	**1495**	**542**	**2037**	**147**	**25**	**13**	**185**
	Life Story	NAR-BE	F	206	939	634	238	872	55	12	0	67
		NAR-CH	F	218	949	632	228	860	75	8	6	89
		NAR-FR	F	198	775	531	192	723	46	2	4	52
		total	**3F**	**622**	**2663**	**1797**	**658**	**2455**	**176**	**22**	**10**	**208**
	7 genres	**25**	**16M/12F**	**4183**	**17778**	**12403**	**4570**	**16973**	**558**	**117**	**130**	**805**