Une comparaison de la déclinaison de F0 entre le français et l'allemand journalistiques

Carolin Schmid[1] Cédric Gendrot[2] Martine Adda-Decker[3]
(1) Universität Trier, FBII-Phonetik, 65296 Trier, Deutschland
(2) Laboratoire de Phonétique et Phonologie, CNRS,UMR7018
ILPGA, 19, rue des bernardins, 75005 Paris
(3) LIMSI-CNRS bat. 508, BP 133, 91403 Orsay cedex
schm2801@uni-trier.de, cgendrot@univ-paris3.fr, madda@limsi.fr

RÉSUMÉ

L'objectif de cette étude est d'explorer la déclinaison de la F0 au cours de séquences comprises entre pauses en français et en allemand à l'aide de grands corpus journalistiques transcrits et segmentés automatiquement (au total environ 80.000 séquences de plus de 1000 locuteurs). Deux méthodes différentes ont été appliquées : (i) une analyse de régression simple pour calculer la déclinaison globale de la F0 et (ii) un algorithme de type convex hull afin de localiser les pics et les vallées de F0 et ainsi obtenir un contour des lignes inférieures et supérieures.
Les résultats montrent des aspects communs aux deux langues : La tendence globale de la F0 à baisser d'environ 2,5 st par seconde ainsi que des prédicteurs communs pour l'amplitude de la pente, tels que la durée de la séquence et la valeur du resetting, de l'intercept et du pic le plus haut. Néanmoins nous constatons une partie de la pente propre à chaque langue dans les mouvements des lignes supérieures et inférieures.

ABSTRACT

F0-declination : a comparison between French and German journalistic speech

The aim of the present study is to investigate F0-declination over the course of utterances in French and German journalistic speech by using large transcribed and automatically segmented corpora (a total of about 80,000 utterances of more than 1,000 speakers). Two different methods were applied : (i) regression-analysis in order to calculate the overall downtrend of F0 and (ii) convex-hull to detect local peaks and valleys in order to calculate the top- and bottom lines. The results show similar characteristics for both languages of the slope : there is an overall declining tendency for the F0 of about 2.5 st per second as well as the same predictors for the amplitude of the slope like utterance duration and the F0-value of the resetting, the intercept and the highest peak. Nevertheless we found language- specific parts of the slope in the mouvements of the top- and bottom lines.

MOTS-CLÉS : intonation, ligne de déclinaison, F0, régression, modelisation, inter-langue, resetting.

KEYWORDS: intonation, declination line, F0, regression, modelling, crosslinguistic, resetting.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 329–336,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La déclinaison de la F0 est définie comme la tendance globale de la fréquence fondamentale à baisser au cours d'une séquence, entre une ligne supérieure reliant ses pics locaux et une ligne inférieure reliant ses vallées locales qui baissent également. Un *resetting* de la F0 a lieu au début de chaque nouvelle séquence (cf. (T'Hart *et al.*, 1990), à voir dans la figure 1). Nous employons

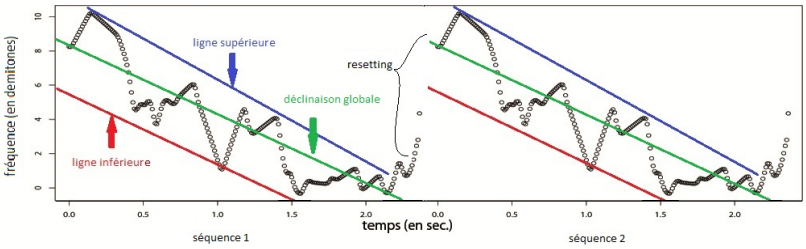

FIGURE 1 – caractéristiques de la ligne de déclinaison : baisse globale au cours d'une séquence, mouvements descendants et montants au niveau local entre deux lignes globalement descendantes(ligne supérieure et inférieure), resetting de la F0 entre deux séquences

le terme *séquence* par la suite en essayant de nous approcher de la notion de la phrase qui relie l'unité de sens et unité intonative et à laquelle la déclinaison donne une notion de cohérence (Cruttenden, 1987; Jun et Fougeron, 2000; Vaissière, 1983).
La tendance globale de la F0 à décliner est liée à la pression sousglottique (Lieberman, 1967), à la traction de la trachée (Maeda, 1976) et aux mouvements des muscles laryngés (Ohala et Ewan, 1973). Pourtant certaines incertitudes subsistent : l'aspect de la déclinaison est-il dépendant de la langue ou est-il controlé par le locuteur ?
Les difficultés à définir plus précisement la nature de la déclinaison sont liées au fait qu'il est délicat d'observer la déclinaison pure. La courbe globale de F0 est constituée de mouvements de différents niveaux prosodiques (Fujisaki, 1988), ce qui fait que la déclinaison est souvent masquée par d'autres facteurs comme des composants de l'accentuation ou d'autres séquences (montée de continuation, resetting, montée ou descente finale) ou des facteurs microprosodiques. Dans la présente étude nous observons également le degré de la pente dans la partie médiane de la séquence seulement, afin d'exclure les influences des changements locaux de F0 aux extrémités des séquences (500 ms de début et de fin).
Il n'existe à notre connaissance que peu d'études inter-langues sur la déclinaison et la plupart examinant de la parole acquise lors d'enregistrements dans des conditions de laboratoire, cf. (Cooper et Sorensen, 1981; Strik et Boves, 1995). La présente étude a pour but d'examiner l'aspect de la déclinaison et les conditions par lesquelles celle-ci est influencée, en mesurant à partir de grands corpus de parole journalistique les correlations entre le degré de la déclinaison et des facteurs qui l'influencent, tels que la durée des segments, la longueur des silences précédents et suivants, la valeur de l'intercept, de la plus haute valeur de la F0 de la séquence et du resetting. Une comparaison inter-langues permettra de montrer dans quelle mesure la ligne de déclinaison

est dépendante du système phono-prosodique de la langue. En effet (Jun et Fougeron, 2000) ont montré que le groupe accentuel, unité prosodique de base du français, est principalement réalisé par le schéma prosodique *LHLH%* alors qu'en allemand 84% des unités prosodiques de base sont réalisées par les schémas *H*L*, *L*H*, *HH*L* ou *L*HL* (Mixdorff, 2002).

2 Corpus et Méthode

Deux corpus audio constitués d'enregistrements d'émissions journalistiques ont été exploités pour réaliser la présente étude.
Le corpus français correspond à environ 30 heures de parole, extraites principalement d'émissions de *France Inter* et fut initié dans le cadre de la campagne *ESTER*(Galliano *et al.*, 2005). Le corpus allemand consiste en environ 20 heures de parole d'émissions d'*Arte*, collectées dans le cadre d'un *projet FP5* (décrit dans (Gendrot et Adda-Decker, 2005)). Les corpus audio ont d'abord été transcrits orthographiquement par des humains, qui ont également indiqué des ruptures prosodiques. Par la suite un alignement automatique des données audio et de leurs transcriptions à été effectué par le *speech-transcription system* du *LIMSI* afin de marquer les frontières des phonèmes et des mots ainsi que les silences(Gauvain *et al.*, 2002). Dans un premier temps les séquences individuelles des corpus ont été selectionnées et extraites avec leurs valeurs de F0. Les unités de parole définies comme des séquences correspondent à des extraits compris entre deux pauses (étant marquées soit par les transcripteurs, soit par l'alignement automatique) et dans lesquelles il n'y avait pas de silence plus long que 50 ms (sinon exclus).
Afin de retrouver dans les extraits la notion de *phrase* mentionnée plus haut nous sommes contraints de nous baser sur les silences, indicateur fiable de la présence des syntagmes de haut niveau (Cruttenden, 1987) (d'autres méthodes sont en cours pour tester l'impact de la demarche d'extraction des séquences).
Le logiciel PRAAT(Boersma et Weenink, 2012) a été utilisé pour extraire les valeurs de F0 (par defaut toutes les 10 ms et sur les positions centrales des segments voisés seulement). Les séquences étaient sauvegardées avec leurs informations sur le nom du locuteur, la langue, la durée (en ms), les silences (lieu et durée en ms), le resetting et les valeurs de la F0 afin de calculer non seulement la déclinaison mais aussi l'influence de certains facteurs sur son degré.
La valeur du resetting constitue la différence (en st) entre la première valeur de la F0 d'une séquence et la dernière valeur de la F0 de sa séquence précédente. Dans cette première approche, nous avons décidé de suivre le protocole de (Yuan et Liberman, 2010).
Les courbes de la F0 ont ainsi été optimisées selon le processus suivant : (i) en les interpolant afin d'obtenir un contour continu (interrompu précédemment par les segments non voisés), (ii)en les lissant par un filtre passe-bas, (iii) et en convertissant les valeurs en Herz en valeurs en demi-tones (st) par la formule suivante : $st = 12 * log_2(\frac{F_0}{F_0 - 5^{ième} quantile})$

Nous avons choisi comme fréquence de base pour chaque séquence le $5^{ième}$ quantile de la fréquence moyenne de toutes les séquences d'un même locuteur. Pour mesurer la déclinaison nous avons d'abord calculé la ligne de régression globale par *ordinary least square modelling* pour la séquence entière (à voir dans la figure 2, à gauche) ainsi que pour sa partie médiane (après avoir retiré les premières et dernières 500 ms des séquences).

Ensuite les positions des pics et des vallées des contours de F0 ont été mesurées par l'algorithme *convex hull* (Mermelstein, 1975) afin d'obtenir la ligne supérieure et la ligne inférieure du contour unique de F0 d'une séquence (voir figure 2, à droite). Afin de retrouver une éventuelle tendance propre à chaque langue, les pics et les vallées d'une séquence ont été moyennés respectivement à 6 points relatifs : les valeurs de F0 figurant entre 0 et 10% de la durée de la séquence, celles figurant entre 10 et 30%, entre 30 et 50%, entre 50 et 70%, entre 70 et 90% et entre 90 et 100%.

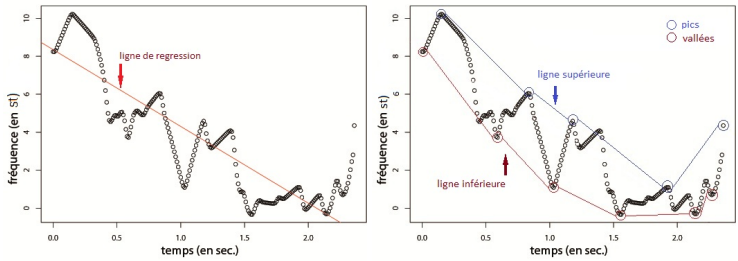

FIGURE 2 – à gauche : la ligne de régression calculée sur le contour de F0 d'une séquence. à droite : les ligne supérieures et inférieures reliant les pics et les vallées du contour de F0 d'une séquence, détectées par *convex hull*

3 Résultats

Le corpus français consistait initialement en 46.630 séquences d'une durée moyenne de 1,75 secondes et en allemand de 33.394 séquences avec une durée moyennne de 1,4 secondes. Dans le but de pouvoir comparer nos résultats aussi à ceux obtenus par (Yuan et Liberman, 2010) pour la ligne de déclinaison en Anglais et en Mandarin, nous avons décidé de baser nos analyses également sur les séquences d'une durée d'entre 1 et 4 secondes et avec une pente négative. Ce choix se justifie en outre par nos propres observations : les séquences d'une durée inférieure à 1 seconde montrent des valeurs de régression extrèmes et les séquences d'une durée supérieure peuvent éventuellement résulter d'une erreur de segmentation en séquences, tenant en compte la durée moyenne des séquences et le pourcentage relativement bas des séquences d'une durée supérieure à 4 secondes (2,8% en allemand et 6.5% en français).
Les séquences en allemand montrent un pourcentage plus élevé de pentes négatives que les séquences françaises (74,5% contre 60,8%). Nous supposons que le nombre important de séquences avec une pente globalement montante est lié à une intonation marquée à cause des montées de continuations (notamment pour le français) et des questions.
Toujours dans le but d'assurer une comparaison avec les travaux de (Yuan et Liberman, 2010), nous avons comparé exclusivement les séquences avec une ligne de régression négative. Cette restriction permet au final l'analyse de 16.987 séquences de plus de 700 locuteurs français et 13.413 séquences de plus de 400 locuteurs allemands au total.

3.1 Contour global

Nous avons pu constater des similitudes entre les 2 langues en ce qui concerne le degré des pentes negatives ainsi que des facteurs qui l'influencent. En allemand le degré moyen de la pente s'élève à -2,5 st/s et en français à -2,4 st/s pour le contour global de F0 sur la séquence complète. Le degré de la pente sur sa partie médiane seulement est de -2,3 st/s en allemand et de -2,4 st/s en français.
La correlation entre la durée de la séquence et le degré de la pente est de $r^2 = 0,4$ pour les deux langues (à voir dans la figure 3) : plus la séquence est courte, plus sa pente négative est raide. Entre la valeur de l'intercept et le degré de la pente, le coefficient de corrélation s'élève à

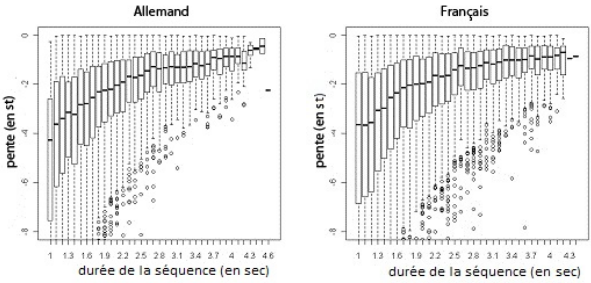

FIGURE 3 – comparaison de la correlation entre la valeur de la pente et la durée de la séquence en allemand et en français

$r^2 = 0,6$ pour les deux langues : plus la valeur de l'intercept est haute, plus la pente négative est raide. La plus haute valeur de la F0 de la séquence à également un effet sur le degré de la pente. Plus cette valeur est haute, plus la pente négative est raide (respectivement $r^2 = 0,2$).
Pour la durée des silences précédents et suivants les séquences nous n'avons pas pu observer de correlation avec le degré de la pente (r^2 toujours inférieur à +/- 0,1).
Une correlation entre la valeur du resetting au début de la séquence et le degré de la pente ne se montre que pour des valeurs positives du resetting à partir de 0 st/s ($r^2 = 0,2$ en allemand et $r^2 = 0,3$ en français) : plus le resetting est important, plus la pente négative est raide (voir figure 4). Les valeurs négatives du resetting suggèrent la présence de séquences entre lesquelles la F0 continue à baisser et pour lesquelles aucune correlation avec le degré de la pente peut être constatée dans les deux langues ($r^2 < 0,02$).
Si l'on considère le resetting de la F0 comme marqueur de frontière définissant la séquence (cf. section 1), ce résultat montre que notre approche de la notion de la phrase pourrait encore être améliorée dans des futures études en se basant non seulement sur les silences pour définir les séquences de la parole, mais également sur la présence d'un resetting positif.

3.2 Ligne supérieure et ligne inférieure

Nous avons pu observer des caractéristiques propres à chaque langue, et ce particulièrement pour les mouvements des lignes supérieures et inférieures en allemand et en français (figure 5).

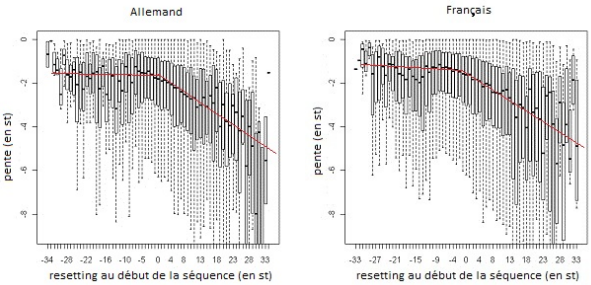

FIGURE 4 – comparaison de la correlation entre la valeur de la régression et la valeur du resetting au début de la séquence en allemand et en français

Dans les deux langues il existe des différences dans le degré des pentes de la ligne inférieure et supérieure. En allemand c'est la ligne supérieure (régression moyenne : -2,5 st/s) qui est plus raide que la ligne inférieure (régression moyenne : -2,1 st/s) : comme des tests-t montrent avec une différence moyenne de 0,4 st/s ($p < 0{,}0001$). En français c'est au contraire la ligne inférieure (régression moyenne : -2,4 st/s) qui est plus raide que la ligne supérieure (régression moyenne : -2,1 st/s) : avec une différence moyenne de 0,3 st/s ($p < 0{,}0001$). Les valeurs de F0 montrent

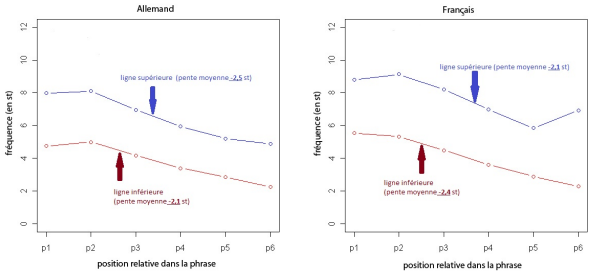

FIGURE 5 – comparaison des lignes inférieures et supérieures de F0 à l'aide des moyennes des pics et des vallées à des positions relatives dans les séquences en allemand et en français

un plus grand registre en français, avec une distance moyenne de 15 st/s entre la plus basse et la plus haute valeur ($p < 0{,}0001$). En allemand le registre comprend en moyenne 13 st/s ($p < 0{,}0001$). Ceci est lié au plus grand registre des valeurs de F0 sur la ligne supérieure en français qui s'étend sur 12 st/s avec une valeur maximale à 15 st/s ($p < 0{,}0001$), en allemand elle s'étend sur 10 st/s avec un maximum à 13 st ($p < 0{,}0001$).
A partir de la deuxième position relative et jusqu'à la partie finale de la séquence il y a dans les deux langues et sur les deux lignes une baisse de la F0 : en moyenne de -0,4 st/s en français et

de -0,7 st/s en allemand (p <0,0001). Dans la partie finale de la séquence par contre peut être

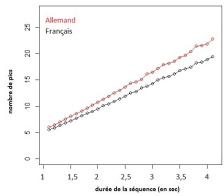

FIGURE 6 – fluctuations de la F0 : le nombre des pics par rapport à la durée de la séquence

observé un mouvement différent en comparant les deux langues. Si les lignes inférieures ont tendence à baisser dans chacune des langues, la ligne supérieure en français montre une montée finale avec une valeur moyenne de 1,6 st/s (p <0,0001) tandis qu'en allemand cette ligne montre une descente finale avec une valeur moyenne de -0,5 st/s. Cette différence (p <0,0001) fait référence aux systèmes prosodiques des deux langues, le français étant une langue à frontières (et utilisant des montées finales pour marquer ses frontières), l'allemand étant une langue à accent lexical et les montées de F0 étant situées à l'intérieur des sequences. Cette différence dans l'accentuation est également visible dans la figure 5, qui illustre la correlation entre la durée de la séquence et le nombre des pics du contour de F0. Dans les séquences de même durée se trouvent toujours plus de pics en allemand qu'en français.

4 Conclusion et Discussion

Cette étude a pu montrer une tendance à la déclinaison de F0 pour de la parole journalistique en allemand et en français. Le fait que cette tendance soit comparable autant pour toute la séquence que sur sa partie médiane seulement, montre que la déclinaison est indépendante des mouvements initiaux et finaux de la F0. Pour les pentes de ces deux langues nous avons pu constater un aspect semblable en ce qui concerne les lignes de régression du contour global (une pente moyenne d'environ -2,5 st/s). Ce résultat est comparable à celui trouvé par (Yuan et Liberman, 2010) pour l'Anglais. La correlation du degré de la pente avec d'autres facteurs tels que la durée de la séquence, l'intercept, le resetting et le pic le plus haut semble également similaires entre les 2 langues.
Pourtant les mouvements des lignes inférieures et supérieures du contour F0 semblent plus spécifiques à la langue. La ligne supérieure apparaît influencée par des mouvements locaux de F0 et correspond aux système phono-prosodique de la langue.
La pente de déclinaison pourrait ainsi être jugée en partie contrôlée par le locuteur. Celui-ci semble surtout avoir un contrôle sur la ligne supérieure qui relie les mouvements locaux de la F0, mais qui peut néanmoins avoir des effets sur l'aspect général de la déclinaison globale. La relative similarité inter-langues de la ligne inférieure nous mène pourtant à supposer que la déclinaison globale de la F0 est conditionnée partiellement physiologiquement.

Références

BOERSMA, P. et WEENINK, D. (2012). Praat : doing phonetics by computer[computer program]. version 5.3.04. `http://www.praat.org/`. [consulté le 12/01/2012].

COOPER, W. et SORENSEN, J. (1981). *Fundamental frequency in sentence production*. Springer-Verlag, New York.

CRUTTENDEN, A. (1987). *Intonation*. Cambridge University Press, Cambridge.

FUJISAKI, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. *In* FUJIMURA, O., éditeur : *Vocal Physiology : Voice Production, Mechanisms and Functions*, pages 347–355. Raven, New York.

GALLIANO, S., GEOFFROIS, E., MOSTEFA, D., CHOUKRI, K. et BONASTRE, J.-F. (2005). The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. *In Eurospeech-Interspeech*, pages 1149–1152.

GAUVAIN, J., LAMEL, L. et ADDA, G. (2002). The Limsi Broadcast News Transcription System. *Speech Communication*, 37(1-2):89–108.

GENDROT, C. et ADDA-DECKER, M. (2005). Impact of duration on f1/f2 formant values of oral vowels : an automatic analysis of large broadcast news corpora in french and german. *In INTERSPEECH*, pages 2453–2456.

JUN, S.-A. et FOUGERON, C. (2000). A phonological model of french intonation. *In* BOTINIS, A., éditeur : *Intonation : Analysis, Modeling and Technology*, pages 209–242. Kluwer Academic Publishers, Dordrecht.

LIEBERMAN, P. (1967). *Intonation, perception, and language*. MIT Press, Cambridge.

MAEDA, S. (1976). *A characterization of American English intonation*. Thèse de doctorat, MIT, Cambridge.

MERMELSTEIN, P. (1975). Automatic segmentation of speech intosyllabic units. "*J. Acoust. Soc. Am.*, 58(4):880–883.

MIXDORFF, H. (2002). Speech technology, tobi and making sense of prosody. *In Invited talk at Speech Prosody 2002, Aix, France*, pages 31–38.

OHALA, J. et EWAN, W. (1973). Speed of pitch change. "*J. Acoust. Soc. Am.*, 53(1):354.

STRIK, H. et BOVES, L. (1995). Downtrend in F0 and Psb. *J. Phonetics*, 23:203–220.

T'HART, COHEN et COLLIER (1990). *A perceptual study of intonation : An experimental-phonetic approach to speech melody*. Cambridge University Press, Cambridge.

VAISSIÈRE, J. (1983). Language-independent prosodic features. *In* CUTLER, A. et LADD, D., éditeurs : *Prosody :models and measurements*, pages 53–65. Springer, Berlin.

YUAN, J. et LIBERMAN, M. (2010). F0 declination in English and Mandarin broadcast news speech. *In Proceedings of Interspeech 2010*, pages 134–137.