CE QUE L'ANNOTATION AVEC TREETAGGER PERMET D'APPRENDRE SUR LES ADVERBES ET ADVERBIAUX DANS LES SCRIPTS D'AUDIODESCRIPTION ALLEMANDS

Eva Schaeffer-Lacroix
MCF HDR sections 7 et 12
Inspé de Paris

PLAN

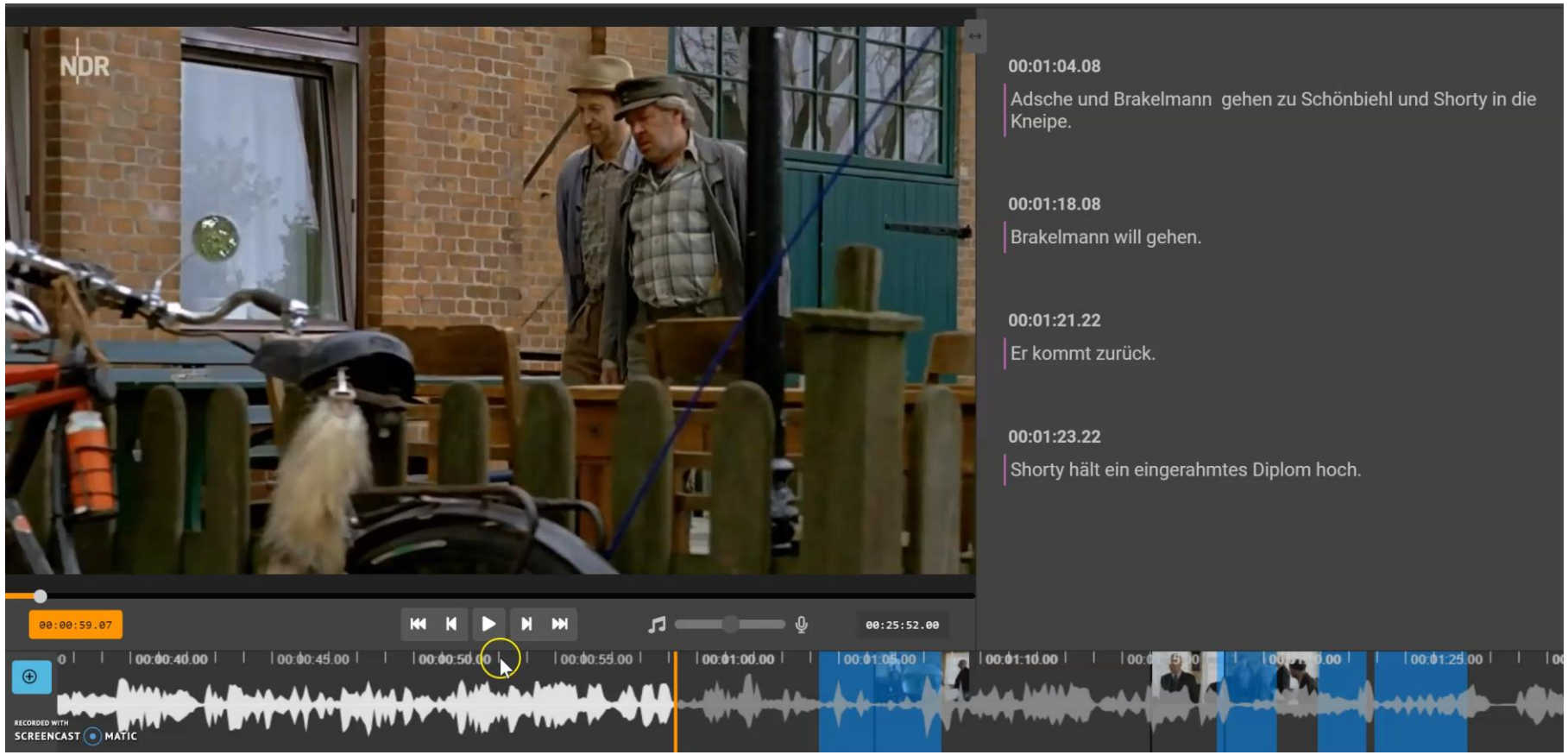
- Audiodescription
- Adverbes et adverbiaux allemands
- Corpus *Buettenwarder*
- Annotation XML
- Annotation par le TreeTagger
- Analyse des données
- Discussion

AUDIODESCRIPTION

Définition d'audiodescription" (AD) dans les directives du Norddeutscher Rundfunk (2019) :

- piste audio supplémentaire destinée à des personnes malvoyantes ou aveugles ;
- rend le contenu visuel d'un film accessible ;
- très dense en termes d'informations fournies.

EXEMPLE D'AUDIODESCRIPTION



Adsche et Brakelmann rejoignent Schönbiehl et Shorty au bar.

Brakelmann s'apprête à partir.

Il revient.

Shorty brandit un diplôme encadré.

Kömkuchen - épisode 45 de la série *Neues aus Büttenwarder* (Eberlein 1997-2017) & Frazier

ÉTAT DE L'ART

- En linguistique de corpus, peu de recherches sur les adverbes et adverbiaux car manque de données en qualité et quantité suffisante (Volk/Graën 2017)
- En audiodescription, recherches assez fréquentes sur les parties du discours
  - ➔ Adverbes allemands et finnois (Hirvonen 2013)
  - ➔ Adverbes néerlandais (Reviers et al. 2015 ; Reviers 2018)

> "[...] les adverbes apparaissent moins souvent dans le corpus AD que dans l'échantillon de langue générale [...]" (Reviers et al. 2015)

L'APPORT DES ADVERBES ET ADVERBIAUX POUR LE DOMAINE DE L'AUDIODESCRIPTION (1)

- La présence d'un adverbe provoquerait une impression de flottement au-dessus de la scène (Larrory 2021)

  ➔ Potentiel de créer des "instantanés cinématographiques"

  ➔ Potentiel de condenser l'information

L'APPORT DES ADVERBES ET ADVERBIAUX POUR LE DOMAINE DE L'AUDIODESCRIPTION (2)

- Adverbes et adverbiaux comme moyen pour analyser le degré de réalisation du postulat d'objectivité dans les directives du NDR (2019) :

  "L'audiodescription ne doit pas expliquer, ni évaluer, ni **interpréter.**"

- Postulat relativisé & les films pour enfants :

  "[Dans les films pour enfants,] il est important de décrire les expressions faciales et de classer les émotions (par exemple 'Attristée, Anna baisse les yeux')."

Source : *Vorgaben für die Audiodeskription* (NDR 2019)

(N.B. : Les traductions des citations dans cette présentation ont été faites avec DeepL & post-édition.)

HYPOTHÈSE

- Les adverbes modaux permettent d'exprimer "le rapport que le sujet énonciateur entretient avec le contenu propositionnel" (Le Querler 2004 : 646).

==> Moins il y a d'adverbes modaux dans les scripts AD, plus ces derniers sont conformes au postulat d'objectivité.

MÉTHODE

- Identification d'adverbes et adverbiaux dans les différentes sections des scripts AD du corpus *Buettenwarder* à l'aide du TreeTagger & post-édition
- Analyse quantitativé & TXM
  - Fréquences absolues
  - Fréquences pondérées
- Analyse qualitative sémantique
  - Espace
  - Temps
  - Modalité
    - Degré
    - Évaluation
    - Appréciation

ADVERBES

- Mots inconjugables et indéclinables
- Classe de mots fermée (≈ 100 unités ; liste)
- Exemples : *vielleicht* [peut-être], *leider* [malheureusement], *da* [là]

ADVERBIAUX

- (Groupes de) mots appartenant à une autre catégorie lexicale & fonction d'adverbe
  - ➔ *langsam* [lentement] : forme adjectivale
  - ➔ *gebückt* [courbé] : forme verbale
  - ➔ *lächelnd* [en souriant] : forme verbale

CARACTÉRISTIQUES FORMELLES

Zinsmeister & Heid (2003)

Utilisé comme adverbe	Utilisé de manière prédicative	Utilisé de manière attributive	Exemple	
ADV			*lediglich*	[seulement]
ADV	ADV		*vergebens*	[en vain]
ADV Signification 1		ADJA Signification 2	*nämlich*	[1 à savoir ; 2 le|la même]
ADV Signification 1	ADJD Signification 2	ADJA Signification 2	*eben*	[1 exactement ; 2 plat|plane]
ADJD	ADJD	ADJA	*wahrscheinlich*	[probablement]
	ADJD	ADJA	*schuldig*	[coupable]
		ADJA	*obere*	[supérieure, placée en haut]
forme non fléchie	forme non fléchie	forme fléchie		

*zu **ebener** Erde* [à même le sol]
***Eben** das wollte ich sagen*. [C'est exactement ce que je voulais dire.]

FONCTION DES ADVERBES

grammis (Leibniz-Institut für Deutsche Sprache, 2018) :

"La fonction prototypique des adverbes est de modifier sémantiquement un événement, un objet ou un fait. Le type de modification ou de spécification dépend des propriétés sémantiques de l'adverbe : spécification spatiale (...), temporelle (...), concessive (...) ou modale (...), etc."

CLASSIFICATION SÉMANTIQUE

Catégorie	Exemple	
local, directionnel	drüber	[au-dessus]
temporel fréquentatif duratif	jetzt oft lange	[maintenant] [souvent] [pendant longtemps]
modificatifs - degré - évaluation, jugement - appréciation	 sehr nur ziemlich	 [très] [seulement] [assez]
reste	alleine	[seul | e]

grammis
Inhaltlich begründete Adverb-Subklassen
[Sous-classes sémantiquement fondées)

PROJET PILOTE *BUETTENWARDER*

Série télévisée *Neues aus Büttenwarder* (Eberlein 1997-2017)

Projet en amont du projet *TADS* (Traduction de scripts d'audiodescription)

28 février – 10 juillet 2020

Partenaire : Kirsten Berland (Licence de LLCER Coréen, filière Traitement Numérique Multilingue INALCO, Paris), stage de 20 jours au laboratoire CeLiSo (Centre de linguistique en Sorbonne)

CORPUS BUETTENWARDER

- Épisodes 20, 45, 52-73 : annotations XML-Tei suffisamment revues et validées
- Implémentés dans TXM (71850 tokens = mots et signes de ponctuation)
- Propriétés : *word* [forme], *depos* [type de mot], *delemma* [lemme], *n* [nombre indiquant l'ordre d'une forme dans le corpus]
- Structures : *caption*, *prompt*, *sp*, *stage*, *time*.

Outil de textométrie TXM (Heiden et al. 2010)

SCRIPT D'AUDIODESCRIPTION

10:03:52 "Er hat gewisse Ansichten, ja." ("Aber" übersprechen) Drinnen.	10:03:52 "En effet, il en a de ces opinions." (Chevaucher "Mais") À l'intérieur.
10:15:19 "Wir haben noch eine Menge zu tun. Aber für heute ..." Schönbiehl geht zur Tür ... "habe ich genug." (KNIPSEN) s ... und macht das Licht aus.	10:15:19 "Nous avons encore beaucoup à faire. Mais pour aujourd'hui, ..." Schönbiehl va à la porte ... "j'en ai assez." (ÉTEIGNAGE) r ... et il éteint la lumière.

Source : épisode 60 de la série télévisée *Neues aus Büttenwarder* (Eberlein 1997-2017)

SECTIONS AD ET ÉTIQUETTES XML-TEI

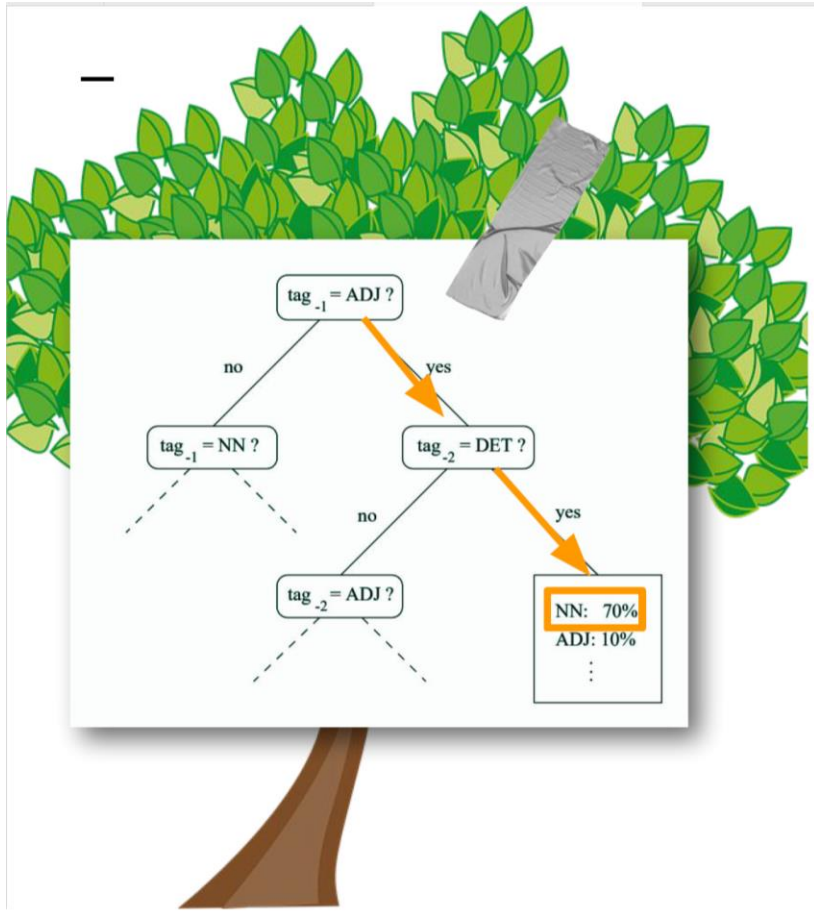
Section AD	Étiquette XML-TEI	Exemple
indication de temps	<time>	10:08:06
fin de dialogue de film (précède la partie à enregistrer)	<prompt>	"Kannst du mal kurz halten?" ["Tu peux tenir ça une seconde ?"]
didascalies (instructions pour l'audiodescripteur ou l'audiodescriptrice)	<stage>	("So, ganz vorsichtig" übersprechen) [Chevaucher "Bon, tout doucement")
parties à enregistrer (speaker)	<sp>	Adsche stützt den Kopf der Frau. [Adsche stabilise la tête de la femme.]
débit	<stage type="delivery">	s, ss, n (schnell, sehr schnell, normal) [rapide, très rapide, normal]
texte qui s'affiche à l'écran	<caption>	Auf Jürgens Kittel steht 'Fräulein Erika'. [La blouse de Jürgen porte l'étiquette 'Mademoiselle Erika'.]

FICHIER XML

```
<time>10:07:24</time>
<prompt>"Kannst du mal kurz halten?"</prompt>
<stage>(<prompt>"So, ganz vorsichtig"</prompt> übersprechen)</stage>
<sp>Adsche stützt den Kopf der Frau.</sp>

<time>10:08:06</time>
<prompt>"Ach, Adsche."</prompt>
<stage type="delivery">ss</stage> <sp>Auf Jürgens Kittel steht <caption>'Fräulein
Erika'</caption>.</sp>
```

TREETAGGER



Decision Trees

Gesucht: P(NN|DET, ADJ) = 0,7

- ➔ Sobald zweimal ja gesagt wird, wird der Wahrscheinlichkeitsvektor ausgegeben
- ➔ Beispiel STTS: Maximal 106 Tests für 54 Tags
- ➔ Entscheidungsbäume werden durch einen Algorithmus berechnet, der die Tests so wählt, dass möglichst wenig Schritte gemacht werden müssen

Source :
Deborah Watty

DÉFINITION "ADVERBE" DANS LE STTS (STUTTGART-TÜBINGEN TAGSET)

Guidelines für das Tagging deutscher Textcorpora mit STTS (Schiller et al. 1999, p. 56) :

"Seuls les vrais modificateurs de verbes, d'adjectifs, d'adverbes et de phrases entières qui ne sont pas dérivés d'adjectifs et ne peuvent être infléchis sont compris comme des adverbes.

Les formes verbales qui se présentent également comme des adjectifs attributifs et qui sont utilisées comme adverbes, mais qui n'ont sémantiquement rien à voir avec l'adjectif et qui ne peuvent généralement pas non plus être utilisées comme prédicatifs, sont comptées comme des adverbes (par exemple *nämlich* [à savoir])."

ANNOTATIONS TREETAGGER

	$"	"	
Oh	ITJ	oh	
,	$,	,	
eine	ART	eine	
Rede	NN	Rede	
.	$.	.	
Nee	PTKANT		Nee
,	$,	,	
dann	ADV	dann	
lieber	ADJD	lieb	
singen	VVINF	singen	
.	$.	.	
"	$"	"	
Genervt		VVPP	nerven
setzt	VVFIN	setzen	
Schönbiehl		NE	Schönbiehl
sich	PRF	sich	
wieder		ADV	wieder
.	$.	.	

"Oh, un discours ? Ah non, je préfère encore chanter."
Énervé, Schönbiehl se rassoit.

Erreurs d'étiquetage :

- "lieber" [de préférence] est un adverbe, et son lemme s'appelle "gern" (il est le comparatif I de cette forme).
- "Gernervt" [Énervé] est un adverbial (participe passé utilisé comme adverbe).

ANNOTATION À L'AIDE DU TREETAGGER

Repérage d'éléments spécifiques à l'audiodescription : s, ss, n.

Requête ss

	depos	Fréquence
	ADJA	260
	VVFIN	236
	VVIMP	105
	NN	23
	ADJD	15
	NE	1

ss (*sehr schnell* – très vite)

étiquettes :

- adjectif attribut du sujet
- verbe conjugué
- verbe à l'impératif
- nom commun
- adjectif ayant la fonction d'adverbe
- nom propre

Keiner will der erste sein.	ss_ADJA	Adsche stellt Kreuze auf.
"Echt?"	ss_ADJD	... und gibt Leonie die Papiere.
("Verstehe" übersprechen)	ss_VVFIN	Leonie sitzt im Auto.
"Jetzt brauchen wir nur noch zwei."	ss_VVIMP	Auf der Dorfkrug-Terrasse liest Griem Krischans Quittung.

QUELLE ÉTIQUETTE POUR S, SS, N ?

a) XY ("non-mot")
b) ADVAD ("adverbe dans le cadre d'un script d'audiodescription)
c) <stage type="delivery">ss</stage">

Re: Frage zu TreeTagger-Etiketten im Rahmen einer Analyse von Audiodeskriptionsskripten

7 Juillet 2020 18:45

Expéditeur : schmid

À: elacroix thomas schmidt

Liebe Frau Lacroix

> Wir haben ein kleines deutschsprachiges Korpus in TXM implementiert und prüfen die Etiketten, die der TreeTagger den Daten beim Hochladen zugeordnet hat. Wir fragen uns, wie die Elemente annotiert werden sollten, die den Sprecher|inne|n angeben, wie etwas gesprochen werden soll, z.B. "s" (schnell), "ss" (sehr schnell), "n" (normal). Sie tragen unterschiedliche Etiketten, z.B. "ADJA", "ADJD" oder "VVIMP" (siehe beigefügte Exceldatei). Wir fragen uns, ob es sinnvoll ist, sie alle als "ADJD" oder als "VVIMP" zu deklarieren. Oder glauben Sie, es wäre sinnvoll, extra ein Etikett zu erfinden, das speziell auf die Audiodeskription zugeschnitten wäre, z.B. "ADVAD" (Adverb im Rahmen der Audiodeskription)? Oder wäre "ADJDAV" angemessener (aber sicher zu lang)?

Ich würde diese Marker wohl am ehesten als Nichtwörter (Tag XY) annotieren. In der Tiger-Baumbank werden bspw. Autorenkürzel als XY annotiert. Sie könnten dann Einträge für diese Wörter in einem kleinen Ergänzungslexikon auflisten und dem TreeTagger mit der Taggeroption -lex übergeben. Dann sollte er diese Marker korrekt für Sie annotieren.

Sie können auch das Tagset erweitern und ein neues Tag vergeben. Dann müssen Sie aber den TreeTagger neu trainieren, um damit auch annotieren zu können.

ÉTIQUETTES TREETAGGER POUR ADVERBES ET ADVERBIAUX DANS DES SCRIPTS AD

Étiquette	Définition	Exemple	
ADV	Adverbe	schon, bald	déjà, bientôt
ADJD	Adverbial ou adjectif prédicatif	[er fährt] schnell	[il avance] rapidement
PAV	Adverbe pronominal	dafür	pour cela
ADVAD	Adverbe spécifique aux scripts AD	s, ss, n	r, tr, n [rapide, très rapide, normal]

ANALYSE QUANTIFICATIVE

Fréquence relative & fréquence pondérée

	ADV	Pond.	ADJD	Pond.	ADVAD	Pond.	PAV	Pond.
prompt	1752	1000	524	11,3	1	-107,4	60	2,1
speaker	436	-185,3	725	0,8	9	-202,6	103	4
stage	136	-3	181	9,7	779	1000	1	-4,5
caption	9	-1,2	3	-1,9	0	-2,2	0	-0,5
Total	2333		1433		789		164	

CLASSIFICATION SÉMANTIQUE DES ADV DANS *STAGE*, *PROMPT* ET *SPEAKER*

Section	local directionnel	temporel fréquentatif duratif	modificatif (degré, évaluation, appréciation)	Reste	Total
stage	10	25	(30)	(71)	76
prompt	164	423	261	847	1695
speaker	152	224	23	37	436
Total	326	672	314	955	2207

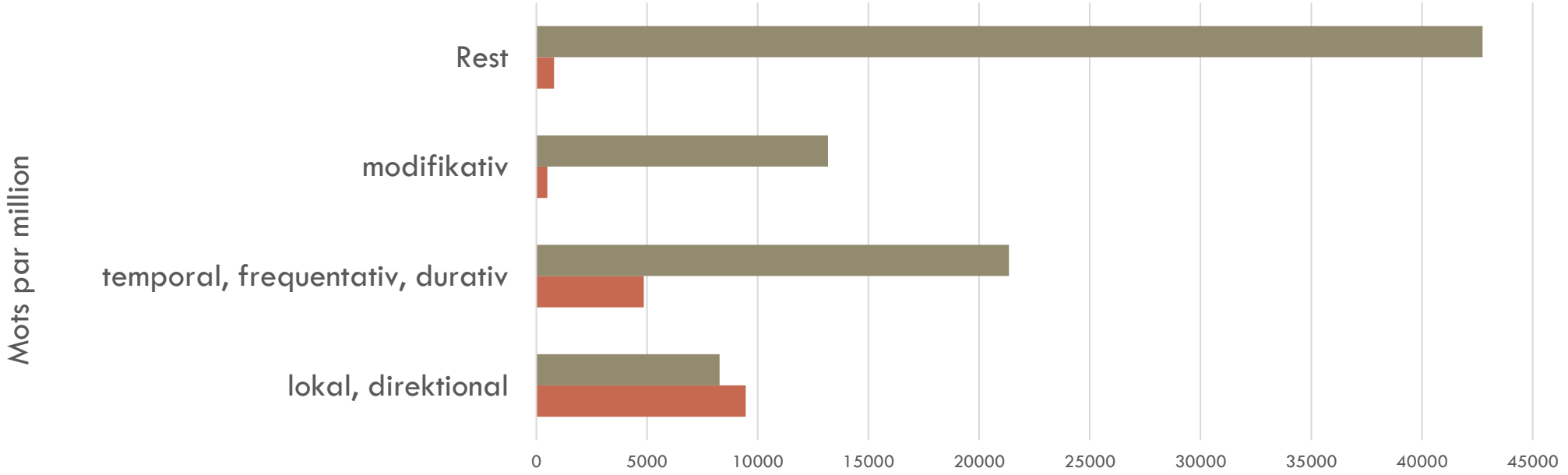
N.B. : Il s'agit de fréquences absolues. Les fréquences entre parenthèses sont des citations de *prompts* au sein de la section *stage*.

CLASSIFICATION SÉMANTIQUE DES ADJD DANS *SPEAKER*

classe de mots	local, directionnel	temporel, fréquentatif, duratif	modificatif		Total
			Degré, évaluation, appréciation	Mimique, émotion	
ADJD (participe I)	29	1	0	91	118
ADJD (autres formes)	118	99	83	305	608
Total	147	100	83	396	726

N.B. : Il s'agit de fréquences absolues.

FRÉQUENCE RELATIVE DES FORMES ADV DANS LES SECTIONS "PROMPT" ET "SPEAKER"



	lokal, direktional	temporal, frequentativ, durativ	modifikativ	Rest
prompt	8276,5	21347,4	13171,8	42745,3
speaker	9457,7	4859	498,9	802,6

RÉSULTATS (1)

- **Section *stage* :**

- peu de variance lexicale ; couvre seulement deux concepts sémantiques : degré et fréquence

- **Section *speaker* :**

- couvre un large éventail de concepts (lieu, temps, degré, fréquence, concomitance, etc.)

- **Section *prompt* :**

- couvre également un large éventail de concepts

- sujette à des erreurs d'étiquetage TreeTagger & présence de fragments de texte et d'éléments hétérogènes (particules conversationnelles, les particules modales, les particules de degré, interjections).

➔Solutions possibles :

- L'étiquetage des données de cette section à l'aide du jeu de balises STTS 2.0. = une alternative au jeu de balises STTS 1.0.

- Ne pas tenir compte de cette section, qui renseigne peu sur la nature des scripts AD.

RÉSULTATS (2)

Ce qui a été obtenu :

- Précision du degré d'objectivité des différentes sections du corpus Buettenwarder

➔ Il y a moins d'énoncés de jugement et de coloration subjective dans *speaker* que dans *prompt,* donc conformité par rapport au postulat d'objectivité.

- Dans la section *speaker,* on trouve toutefois un certain nombre d'occurrences d'audiodescription de mimiques et d'émotions, de façon privilégiée à l'aide de participes I et II utilisés comme adverbes.

- Les étiquettes TreeTagger pour les mots qui sont formellement des participes II et susceptibles d'être utilisés comme adverbes doivent encore être vérifiées (ADV ou pas ?).

➔ La détection automatique des adverbes et adverbiaux illustre le fait qu'en règle générale, il est difficile de distinguer les adverbiaux des autres parties du discours.

RÉSULTATS

La reconnaissance automatique des adverbes et adverbiaux permet de...

- montrer à quel point il est difficile de distinguer les adverbiaux des autres parties du discours ;
- documenter le degré de discours objectif dans les différentes sections des scripts AD.

RÉFÉRENCES BIBLIOGRAPHIQUES

Hirvonen, Maija. 2013. Perspektivierungsstrategien und -mittel kontrastiv - Die Verbalisierung der Figurenperspektive in der deutschen und finnischen Audiodeskription. trans-kom - Journal of Translation and Technical Communication Research 6(1). 8–38.

Larrory, Anne. 2021. „Seitenweise Sonderpreise“. Zu einigen ‚weise-Adverbien‘ im Deutschen". Colloque Deutsche Adverbien und Adverbiale aus deskriptiver, theoretischer und vergleichender Sicht. Maison des Sciences de l'Homme d'Aquitaine, Université Bordeaux Montaigne, 29-30 avril 2021.

Leibniz-Institut für Deutsche Sprache. 2018. Semantisch begründete Adverb-Subklassen. “*Wissenschaftliche Terminologie*”. *Grammatisches Informationssystem grammis*. https://grammis.ids-mannheim.de/progr@mm/5278 (6 August, 2021).

Le Querler, Nicole. 2004. Les modalités en français. Revue belge de Philologie et d’Histoire. Persée - Portail des revues scientifiques en SHS 82(3). 643–656. https://doi.org/10.3406/rbph.2004.4850.

Norddeutscher Rundfunk. 2019. Vorgaben für Audiodeskriptionen. NDR. https://www.ndr.de/fernsehen/barrierefreie_angebote/audiodeskription/Vorgaben-fuer-Audiodeskriptionen,audiodeskription140.html (7 March, 2020).

Reviers, Nina, Aline Remael, Walter Daelemans, Anna Jankowska & Agnieszka Szarkowska. 2015. The Language of Audio Description in Dutch: Results of a Corpus Study. In New Points of View on Audiovisual Translation and Media Accessibility, 167–190. Peter Lang. http://www.clips.ua.ac.be/~walter/papers/2015/rrd15.pdf (30 September, 2019).

Reviers, Nina. 2018. Studying the language of Dutch audio description: An example of a corpus-based analysis. Translation and Translanguaging in Multilingual Contexts 4(1). 178–202. https://doi.org/10.1075/ttmc.00009.rev.

Volk, Martin & Johannes Graën. 2017. Multi-word Adverbs – How well are they handled in Parsing and Machine Translation? s.n. https://doi.org/10.5167/UZH-141846. https://www.zora.uzh.ch/id/eprint/141846 (11 July, 2021).

Zinsmeister, Heike & Ulrich Heid. 2003. Identifying predicatively used adverbs by means of a statistical grammar model. In Dawn Archer, Paul Rayson, Andrew Wilson & Tony McEnery (eds.), *UCREL Technical Paper*, vol. 16 (Special issue), 932–939. Lancaster. http://ucrel.lancs.ac.uk/publications/CL2003/papers/zinsmeister.pdf.

RESSOURCES ET OUTILS

Eberlein, Norbert. 1997-2017. *Neues aus Büttenwarder*. https://www.ndr.de/mediathek/mediatheksuche105_broadcast-129.html.

Heiden, Serge, Jean-Philippe Magué & Bénédicte Pincemin. 2010. TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, vol. 2(3), 1021–1032. Edizioni Universitarie di Lettere Economia Diritto. https://halshs.archives-ouvertes.fr/halshs-00549779/document (30 November, 2018).

Pajonczek, Lukas & Christian David. 2019. *Frazier*. VIDEO TO VOICE GmbH. https://accessibility.studio/ (26 March, 2020).

Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf (15 July, 2021).

ACTIVITÉS DE L'ÉQUIPE TADS

- Rencontre de l'équipe TADS et de Thomas Schmidt (Université de Bâle, Suisse) du 25-27 novembre 2021 : ateliers TXM et EXMARaLDA ; développement de notre méthodologie.
- Colloque international : *Discussing the Limits of Objective Audio Description*, 5-7 octobre 2022, Institut für Übersetzungswissenschaft und Fachkommunikation, Université de Hildesheim (Allemagne). Subvention *Procope plus 2022* obtenue.
- En collaboration avec Nathalie Mälzer et Maria Wünsche, direction de la publication d'un numéro spécial *JAT* (*Journal of Audiovisual Translation*) à paraître fin 2022. Titre de travail du numéro : *Discussing the limits of Objectivity in Barrier-free Communication*