2. COMMUNICATION

La communication est l'action de transmettre un message contenant de l'information à un récepteur. La communication se distingue par une hétérogénéité des récepteurs. Qui peuvent être des personnes, des groupes, des masses, des animaux, des machines, des robots. On parle de communication interpersonnelle, de communication de groupe ou de communication de masse en fonction du nombre des récepteurs. L'information est de l'ordre du « contenu », la communication est de l'ordre de la « relation ».

Le message transmis est conditionné par son contexte. Les facteurs environnementaux comme la localisation, l'événement, la diversité culturelle, l'importance, le temps, la confidentialité, qui forment le contexte, influencent le contenu du message. Le message transmis est conditionné en outre par son code qui doit être commun entre émetteur et récepteur. Une donnée aussi banale qu'une date dans un message, par exemple le 12/3/2020, est interprétée comme le 12 mars 2020 par un citoyen européen et comme le 3 décembre 2020 par un citoyen américain, à défaut de spécifier le format de date utilisé pour l'encodage et le décodage. Les signes utilisés pour coder et décoder un message peuvent être des chiffres, des lettres, des images, des vidéos, de la parole, des programmes informatiques.

La communication repose sur différents outils. La communication écrite se fait par des livres, journaux, affiches et blogs. Le téléphone permet la communication verbale à distance, tout comme la radio. Le smartphone, l'ordinateur et les messageries supportent les communications verbales et textuelles. La télévision et les réseaux sociaux peuvent transmettre des contenus animés. On parle de diffusion si le message s'adresse à des récepteurs multiples, comme dans les cas de la presse, de la radio, de la télévision, des blogs et des réseaux sociaux. Si les récepteurs sont des consommateurs, la communication devient publicité.

La communication est un besoin humain fondamental. Vivre, c'est communiquer. Les êtres humains souhaitent communiquer pour différentes raisons : apprendre, collaborer, comprendre, convaincre, échanger, exprimer ses sentiments, partager, séduire.

L'Union européenne et le Conseil de l'Europe avaient proclamé l'année 2001 comme « Année européenne des langues ». L'apprentissage des langues élargit les horizons était le message de cette initiative.

2.1. Le langage, la langue et la parole

Le langage désigne la capacité qui permet à chacun d'entre nous de communiquer et d'interagir avec d'autres personnes. Le langage humain possède une créativité extrêmement développée puisqu'à partir d'un nombre limité de sons et de mots, chacun d'entre nous peut exprimer une infinité de messages.

La langue désigne un outil permettant de communiquer. Selon un récit de la Genèse, les humains ne parlaient qu'une seule langue avant la construction d'une tour à Babylone qui, par sa hauteur, touchait le ciel et permettait un accès direct au paradis. Dieu trouvait les humains trop orgueilleux et les punit en brouillant leur langue, bien qu'ils ne se comprissent plus. Ils furent alors contraints d'abandonner la tour de Babel et se dispersèrent sur la terre, formant ainsi des peuples étrangers les uns des autres. C'est en référence à ce récit que l'on utilise parfois le terme tour de Babel pour parler d'un lieu où règnent le brouhaha et la confusion.

La langue n'est donc plus commune à tous les êtres humains, mais seulement à un groupe de personnes. Une langue est souvent associée à une nation ou à une réalité géopolitique. On parle de dialecte si une langue est parlée par un groupe restreint de personnes et d'un patois si ce groupe est très restreint. Une langue peut continuer à exister même si plus personne ne la parle. Il s'agit de langues mortes comme le latin ou le grec ancien. Les langues parlées sont dites vivantes, elles évoluent avec le temps par l'admission de nouveau mots et par le changement de la grammaire.

On appelle langue naturelle une langue qui s'est formée au cours du temps par la pratique de ses locuteurs. C'est le cas de la majorité des langues parlées dans le monde. Au contraire, on appelle langue construite, parfois improprement langue artificielle, une langue qui résulte d'une création normative consciente d'un ou de plusieurs individus. Un exemple est l'espéranto créé en 1887 par Ludwik Lejzer Zamenhof, un ophtalmologue polonais. Cette langue est utilisée par environ 2 millions de personnes dans 120 pays. On estime que le nombre de langues qui existent aujourd'hui sur terre varie entre 3.000 et 7.000, en fonction des critères de comptage utilisés.

Dans le passé, la langue luxembourgeoise a été souvent considérée comme un dialecte germanique, mais par la loi du 24 février 1984 le luxembourgeois a été proclamé comme langue nationale. En 2023, la langue luxembourgeoise a été ancrée dans la nouvelle constitution.

Contrairement au langage, la langue nécessite un apprentissage et s'acquiert au fur et à mesure de sa vie. Toute langue constitue un système complexe réunissant un ensemble de mots (= le lexique) et un ensemble de règles de fonctionnement (= la grammaire, les règles d'agencement des sons, les règles de conjugaison, ...).

La parole désigne l'utilisation concrète de la langue par les individus. Elle désigne donc la manière d'utiliser l'outil. La parole prend en compte la prononciation, l'accent, le rythme, l'intonation ou encore le type de mots ou d'expressions utilisés. Elle est donc plus concrète et plus individuelle que la langue.

Dans la communication avec les machines, le langage est également devenu un outil de plus en plus important. Les machines s'expriment moyennant la synthèse vocale (TTS) et elles interprètent nos messages moyennant la reconnaissance automatique de la parole (STT). Dans ce contexte, le luxembourgeois est considéré comme une langue à faibles ressources, car les bases de données audio enregistrées avec transcriptions, disponibles pour réaliser des systèmes TTS et STT, sont très limitées.

2.1.1. Orthographie et grammaire

Pour apprendre une langue, il faut disposer d'abord de règles d'écriture qui sont définies moyennant une orthographie et une grammaire. En général, l'orthographie est complétée par un dictionnaire. La grammaire règle la construction d'une phrase.

Un premier dictionnaire appelé « Lexicon der Luxemburger Umgangsprache (LLU) » a été publié en 1847 par Jean-François Gangler. Le contenu original est disponible sur le portail « infolux.uni.lu » de l'université du Luxembourg.

Commission de dictionnaire 1897

Une première commission de dictionnaire luxembourgeois a été instituée par la loi du 19 février 1897. Elle était composée au début de sept membres, ensuite des dix membres suivants : Nicolas Gredt, Jean-Pierre Henrion, Charles Mullendorff, Henri Schliep, Caspar Mathias Spoo, Joseph Weber, Nicolas van Werveken, Paul Clemen, André Duchscher et Willy Goergen. Le résultat des travaux de cette commission s'est concrétisé dans la publication « Wörterbuch der luxemburgischen Mundart » en 1906. Ce contenu est également disponible dans la rubrique « Wörterbücher » sur le portail « infolux.uni.lu ».

Commission de dictionnaire 1935

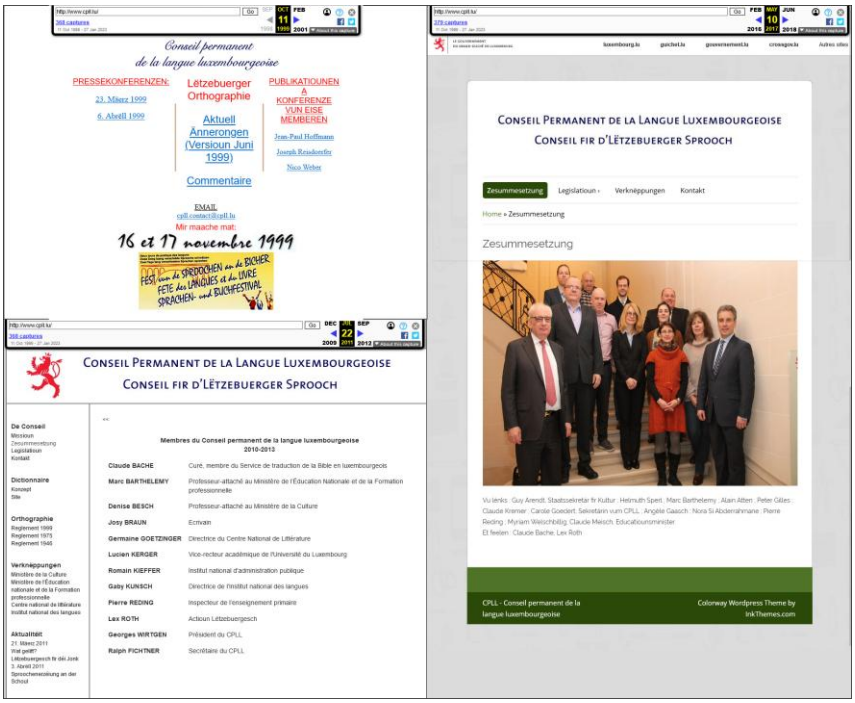
Une deuxième commission de dictionnaire a été nommée en 1935. Les travaux ont été interrompus pendant la guerre. Un système officiel d'orthographie luxembourgeoise a été introduite après la deuxième guerre mondiale par arrêté ministériel du 5 juin 1946, mais n'a pas été accepté par le public. En 1948, la commission instituée en 1935 a été chargée de reprendre ses travaux. Le dictionnaire résultant a été publié en 22 fascicules entre 1950 et 1975 sous le nom de « Luxemburger Wörterbuch », avec un complément en 1977. Les membres de cette deuxième commission étaient Joseph Tockert, Rober Bruch, Joseph Hess, Ernest Ludovicy, Joseph Meyer, Hélène Palgen et Isy Comes.

Commission de dictionnaire 1992

Une troisième commission de dictionnaire a été créée par arrêté gouvernemental du 12 juin 1992. Ses attributions étaient l'élaboration, l'actualisation permanente et l'édition d'ouvrages lexicologiques de la langue nationale ainsi que les réajustements ponctuels de l'orthographe officielle. Après trois ans d'interminables palabres et de manoeuvres d'obstruction de certains membres, la commission a arrêté sa mission. « Péan funèbre pour une commission défunte » est le titre d'un article afférent rédigé le 29 décembre 1995 par Joseph Reisdoerfer dans le Lëtzebuerger Land. Comme les membres de cette troisième commission de dictionnaire sont cités comme pionniers dans le cadre d'autres projets dans le présent livre, je préfère rester discret et ne pas relever leurs noms dans ce sous-chapitre. Le lecteur curieux peut faire ses propres recherches dans les archives sur le web.

L'ancien « Luxemburger Wörterbuch » a été réimprimé la même année (1995), mais il a été retiré de la vente en 1997 à cause des critiques concernant l'inclusion de nombreux proverbes antisémites et misogynes.

Conseil permanent de la langue luxembourgeoise



Pages web du CPLL sur la Wayback Machine

La troisième commission en arrêt de ses travaux a été abolie en 1998 pour donner suite à la mise en place d'un conseil permanent de la langue luxembourgeoise (CPLL), avec siège au centre national de littérature à Mersch, par règlement ministériel du 5 janvier 1998. Ce conseil, secondé par plusieurs groupes de travail, a été créé une deuxième fois avec le même contenu par le règlement grand-ducal du 29 juillet 1999, toujours sous la tutelle de la ministre de la Culture, Erna Hennicot-Schoepges. Le CPLL est chargé d'étudier, de décrire et de diffuser la langue luxembourgeoise. Le premier président du CPLL était Georges Wirtgen, directeur de l'institut supérieur d'études et de recherches pédagogiques (ISERP) situé à Walferdange. Son successeur était Marc Barthelemy en 2017 qui lui-même fut remplacé par Myriam Welschbillig comme président du CPLL par arrêté du gouvernement en conseil en 2019. Le CPLL comprend 11 membres dont les noms sont relevés ci-après par ordre chronologique depuis 1999 : Josy Braun, Guy Dockendorf, René Faber, Germaine Goetzinger, Jeannot Hansen, Jean-Paul Hoffmann, Viviane Mertens, Pierre Reding, Joseph Reisdoerfer, Michel Schmit, Nico Weber, Claude Bach, Denise Besch, Lucien Kerger, Romain Kieffer, Gaby Kunsch, Lex Roth, Guy Berg, Peter Gilles, Josiane Kartheiser, Helmuth Sperl, Caude Kremer, Angèle Gaasch, Nora Si Abderrahmae, Max Kuborn.

Commissaire à la langue luxembourgeoise

Par la loi du 20 juillet 2018 un commissaire à la langue luxembourgeoise a été institué. Il est appelé à contribuer à la mise en œuvre de la politique de la langue luxembourgeoise et à proposer au gouvernement un projet de plan d'action et après adoption du plan d'action par le gouvernement, à superviser et coordonner sa mise en œuvre. Marc Barthelemy, qui travaillait depuis 18 ans au sein du ministère de l'Éducation nationale, comme premier conseiller et professeur attaché, fut nommé premier commissaire de la langue luxembourgeoise en octobre 2018. Le lecteur averti a déjà fait sa connaissance quelques lignes ci-dessus comme deuxième président du CPLL. Après la création du CPLL, d'abord par règlement ministériel, ensuite par règlement grand-ducal, sa création a été confirmée une troisième fois dans la loi du 20 juillet 2018, je suppose pour souligner son importance. Fin décembre 2022, Marc Barthelemy est parti en retraite. Avant de partir, il a fait le bilan de son mandat et a présenté un nouveau plan d'action de 50 mesures pour promouvoir la langue nationale. Ce plan a été approuvé le 14 décembre 2022 par le gouvernement en conseil. Son successeur à partir du 1er janvier 2023 est Pierre Reding, premier conseiller de gouvernement et chef de la direction générale de l'intégration au ministère de l'Éducation nationale, de l'enfance et de la jeunesse. Il était membre du CPLL depuis le début.

Zenter fir d'Lëtzebuerger Sprooch

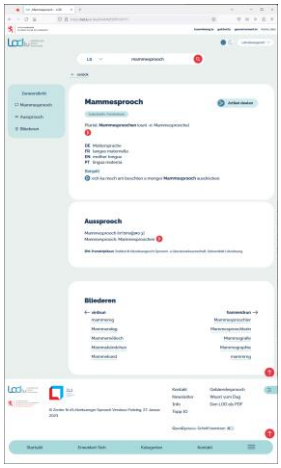
La loi du 20 juillet 2018 a institué un autre nouvel organisme, le centre pour la langue luxembourgeoise (ZLS : « Zenter fir d'Lëtzebuerger Sprooch »), attaché au ministère de l'Enseignement. Depuis sa création, le ZLS est dirigé par Luc Marteling, ancien journaliste et rédacteur en chef de rtl.lu.

L'orthographie actualisée luxembourgeoise a été approuvée en novembre 2019 par le gouvernement en conseil. La version la plus récente a été mise à jour en septembre 2022. Elle est disponible gratuitement comme fichier PDF à télécharger ou comme livre auprès du ZLS.

2.1.2. Lëtzebuerger Online Dictionnaire (LOD)

Avec plusieurs millions de recherches par an et plusieurs mille visiteurs par jour, le dictionnaire luxembourgeois en ligne (LOD : Lëtzebuerger Online Dictionnaire) est certainement apprécié par une grande partie des résidents au Luxembourg. Wikipédia définit le dictionnaire comme un ouvrage de référence contenant un ensemble de mots d'une langue.

LOD sous tutelle culturelle



Page web du LOD

Les origines du LOD remontent à 1992. À l'époque, un groupe « Lëtzebuerger Dictionnaire », institué en 1992, travaillait sous l'égide du ministère de la Culture et sous la responsabilité administrative de Ralph Fichtner. Il était le secrétaire de l'Institut Grand-Ducal, Section Linguistique et le premier secrétaire du Conseil Permanent de la langue luxembourgeoise (CPLL). Le dictionnaire était basé sur le corpus linguistique LuxText qui contenait plus de 1,8 million de mots courants provenant de sources écrites et parlées. Ce corps a été corrigé avec l'aide d'étudiants universitaires qui ont été engagés par le ministère de la Culture pendant les vacances scolaires.

La première trace du LOD sur la Wayback Machine des archives Internet date du 8 février 2007. Ensuite, il y a un vide à partir du 9 février 2008 jusqu'au 6 février 2013 quand on retrouve le LOD sur le web avec l'addition de l'anglais comme langue d'interface. Dans l'édition du Lëtzebuerger Land du 27 juin 2014 François Schanen s'inquiétait sur la situation du LOD dans un article intitulé « Qui trop embrasse… ». En septembre 2015, on découvre une nouvelle page d'accueil qui ne change plus pendant quelques années.

LOD sous tutelle éducative

Le développement et support du LOD ont été confiés au ZLS lors de sa création en 2018, mais ce n'est que sur la copie du site web lod.lu prise le 29 janvier 2019 par la machine Wayback qu'on découvre sur la page d'accueil le remplacement du ministère de la Culture par le ministère de l'Éducation nationale, comme maître d'œuvre. Une année plus tard le site web du LOD a fait peau neuve et se présente sous la responsabilité du ZLS. Dans la suite la disposition et les fonctionnalités du site web lod.lu n'ont guère évoluées jusqu'en juin 2022. Une version flambant neuve (fuschnei) de la plateforme internet LOD, au look moderne et aux fonctionnalités élargies, a été mise en ligne et présentée au public le 20 juin 2022, en présence du ministre de l'Éducation nationale, de l'Enfance et de la Jeunesse, Claude Meisch.

Claude Meisch, ministre de l'Éducation nationale, entouré de l'équipe du LOD en juin 2022

La photo ci-dessus présente l'équipe complète du ZLS, qui se compose de seize personnes, lors de la présentation du nouveau LOD. Parmi le personnel actuel du ZLS, figure l'unique lexicographe du Grand-Duché, Alexandre Ecker, un des pionniers de la création du dictionnaire auprès du ministère de la Culture.

Aujourd'hui, le LOD ne constitue pas un simple dictionnaire en ligne, mais tout un écosystème. Le site web n'affiche pas seulement un champ de recherche sur la page d'accueil, mais également le mot du jour, des informations utiles et des nouvelles sur la langue luxembourgeoise. Le visiteur peut choisir parmi cinq langues pour l'interface : luxembourgeois, allemand, français, anglais et portugais. Le mot recherché est retourné avec l'orthographie correcte, des informations linguistiques, des synonymes, des traductions dans d'autres langues et une transcription en phonèmes moyennant l'alphabet IPA. Les résultats sont complétés par un exemple d'une phrase contant le mot avec la prononciation sonore enregistrée par Max Kuborn.

Une recherche avancée permet d'utiliser des filtres. Si le mot recherché n'est pas trouvé, l'utilisateur peut demander son inclusion dans le dictionnaire avec un simple formulaire en ligne. Une commission du LOD décide régulièrement sur le suivi à donner à ces demandes.

Les fichiers actualisés régulièrement avec la liste des mots du LOD peuvent être téléchargés sur la plateforme de données luxembourgeoises « data.public.lu » sous une licence CC0. Dans ce cadre, il convient également de relever que le LOD dispose d'une interface de programmation (API) pour inclure les fonctionnalités du dictionnaire en ligne dans une application tierce.

2.1.3. Outils informatiques de correction d'orthographie

Pour promouvoir l'écriture d'une langue comme le luxembourgeois, il ne suffit pas de définir une orthographie et une grammaire officielles et de disposer d'un dictionnaire, mais il faut également offrir des moyens pour faciliter la correction de fautes. Des programmes informatiques pour détecter des erreurs dans un texte et de proposer des corrections sont des outils appropriés pour aider les auteurs luxembourgeois à parfaire leurs documents.

EPISTOLE-PC


ESSAI LOGICIEL

EPISTOLE PC

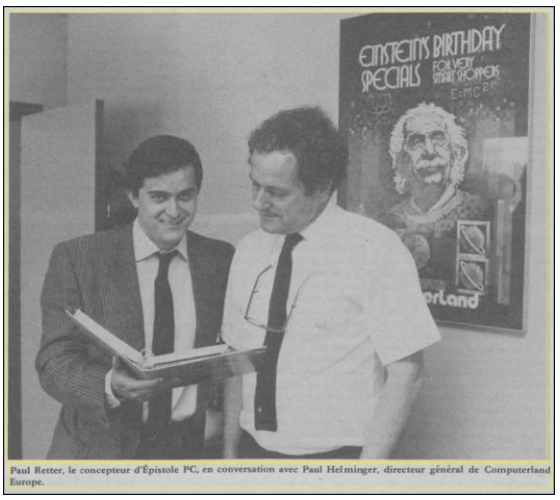
A l'instar d'autres logiciels, Epistole (développé sur Apple puis sur PC) sort à présent en version Junior pour les nouveaux compatibles bon marché.

EPISTOLE PC

PASSABLE BON EXCELLENT

Epistole-PC

Le premier outil de correction automatique de l'orthographie luxembourgeoise a été intégré dans le logiciel de traitement de texte EPISTOLE, adapté au PC par Paul Retter vers le milieu de la décennie 1980. [5] Pour commercialiser son logiciel Paul Retter avait créé en 1988 la société Silis s.à r.l. qui existe encore aujourd'hui. Les traces de ce projet sur le web sont rares. On trouve des tests dans quelques magazines informatiques de l'époque et des annonces de recrutement de dactylos pour des services de l'Etat dans les quotidiens au début des années 1990 où il est mentionné que les connaissances en traitement de texte EPISTOLE PC sont considérées comme avantage. Le logiciel EPISTOLE a été progressivement remplacé par Microsoft Word auprès des administrations, ministères et communes au Luxembourg dans les années suivantes. Hélas, ce programme ne disposait alors pas encore de correcteur d'orthographie luxembourgeoise. En 2004 Thierry Fromes, country Manager de Microsoft Luxembourg, annonçait la disponibilité d'un paquet de configuration (LIP : « Language Interface Pack ») pour utiliser le luxembourgeois dans l'interface de Windows XP. Le LIP pour la langue nationale a été développé en collaboration avec le CPLL et le CRPGL. Ces paquets ont été perfectionnés dans les versions Windows qui suivaient et des correcteurs orthographiques pour le luxembourgeois ont été intégrés dans les logiciels Microsoft Office dans la décennie 2010.



Paul Retter, le concepteur d'Épistole PC, en conversation avec Paul Helminger, directeur général de Computerland Europe.

Paul Retter présente EPISTOLE PC à Paul Helminger

CORTINA

En 1999, le CPLL avait initié un projet appelé CORTINA, l'abréviation pour « Correction ORThographique INformatique Appliquée à la langue luxembourgeoise ». La réalisation du projet a été confiée à la cellule de recherches, d'études et de développements en informatique du centre de recherche public Gabriel Lippmann (CRPGL).

Pierre Mousel

Le chef de projet était Pierre Mousel. Il était secondé par l'informaticien Yannick Durand et par Johannes Kiel qui effectuait un stage au CRPGL. Johannes Kiel rédigeait son mémoire en linguistique informatique à l'université de Trèves et il apportait son expertise concernant « l'Eifeler Regel » relatif à la langue luxembourgeoise dans le projet. Le responsable du CPLL était son président Georges Wirtgen, qui était assisté par le linguiste Jérôme Lulling. Ce

dernier réalisait à l'époque une thèse de doctorat avec le titre « La créativité lexicale dans la langue luxembourgeoise » à l'université Paul Valéry à Montpellier, sous la supervision du professeur des universités François Schanen, originaire du Luxembourg.

Jérôme Lulling

Le projet CORTINA a été présenté le 1er octobre 2001 lors d'une conférence de presse par Erna Hennicot-Schoepges, ministre de la Culture, de l'enseignement supérieur et de la recherche. Une version améliorée (CORTINA 2) a été développée en 2002 et le logiciel de correction a été ajouté d'une façon artisanale dans le code propriétaire de Microsoft Word. Grâce à l'édition par Pierre Mousel d'une description détaillée sur le projet CORTINA qui est encore accessible aujourd'hui sur le web, nous disposons d'une riche documentation sur cet outil de correction de l'orthographie luxembourgeoise. On pouvait entrer des mots isolés ou des textes complets dans un champ sur la page web pour les faire analyser et corriger en cas d'erreurs.

CORTINA était basé sur une architecture client-serveur et fonctionnait comme applet JAVA dans un navigateur Internet standard, à condition de supporter la version 1.1 de JAVA, ce qui posait parfois des problèmes aux usagers. Comme le dictionnaire élaboré par le groupe de travail du CPLL n'était pas encore prêt au début de la décennie 2020, les chevilles ouvrières du projet CORTINA avaient créé leur propre dictionnaire comme cœur du projet.

SpellChecker

Le logiciel de traitement informatique du luxembourgeois le plus connu est certainement l'application « spellchecker.lu » qui a été lancée en 2006 par trois étudiants à l'université technique de Kaiserslautern, Tom Goedert, Luc Heischbourg et Michel Weimerskirch. Depuis 2008, c'est Michel Weimerskirch seul qui continuait à assurer le développement et la maintenance du SpellChecker luxembourgeois. Il est actuellement un partenaire-dirigeant de l'agence créative digitale Lightbulb qu'il a co-fondée en 2013.

Spellchecker App

Les grandes étapes de l'évolution de l'application « spellchecker.lu » figurent sur la page « Historique » du site web du projet. Un module de correction pour le logiciel de traitement de texte LibreOffice et une première version pour mobiles introduits en 2009, un nouveau layout du site web présenté en 2012 et des apps pour iOS et Android lancées en 2017 sont les jalons du projet. Pendant toutes ces années Michel Weimerskirch a essayé d'obtenir une copie du dictionnaire en élaboration auprès du ministère de la Culture pour l'inclure dans l'application SpellChecker, sans succès. Il a été contraint de créer sa propre liste de mots luxembourgeois. Une grande refonte de cette liste a été effectuée en 2016 avec l'assistance de Sandra Souza Morais, étudiante en informatique. Après la reprise du LOD par le LZS, des bons contacts ont été établis entre les deux acteurs. Aujourd'hui, le SpellChecker se base sur l'orthographie approuvée en 2019 et utilise une liste de mots actualisée par le LZS.

Et comme le monde est petit, on ne s'étonne pas que l'agence LightBulb, créée par le développeur du SpellChecker Pierre Weimerskirch, a conçu l'architecture technique et la présentation (design) du nouveau site web lod.lu en 2022.

2.1.4. Outils de traduction automatique

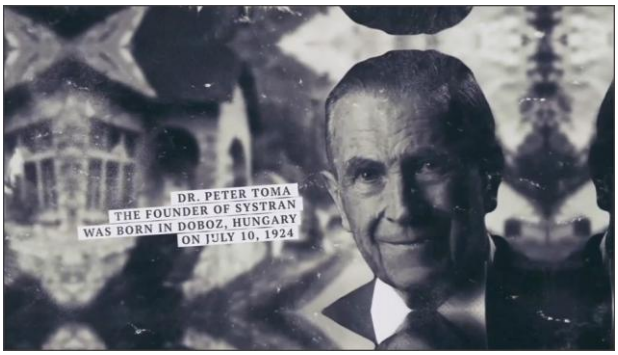
Si deux personnes ne parlent pas la même langue, il faut convertir la langue source dans la langue cible, ce qu'on appelle traduction. Le Luxembourg figure parmi les pionniers dans l'utilisation de la traduction automatique du fait que les nombreuses institutions européennes, ayant leur siège dans le pays, doivent gérer leurs documents dans les différentes langues officielles de la Communauté européenne.

SYSTRAN made in Luxembourg

Siège de la CECA

Tout a commencé au début des années 1950 avec la fondation de la Communauté européenne du charbon et de l'acier (CECA) par six nations européennes, dont le Luxembourg. La CECA a été dirigée par la Haute Autorité qui siégeait à Luxembourg et qui gérait quatre langues officielles : français, allemand, italien et néerlandais. La Haute Autorité de la CECA fusionnait en 1965 avec les commissions de la Communauté économique européenne (CEE) créée en 1957, puis en 1967 avec la Communauté européenne de l'énergie atomique (Euratom), sur base du traité de fusion des exécutifs des trois communautés. En janvier 1969 l'Office des publications officielles des Communautés européennes, devenu en 2009 l'Office des publications de l'Union européenne, a été créé et domicilié à Luxembourg. L'Office des publications publie quotidiennement le Journal officiel de l'Union européenne dans les 23 langues officielles de l'Union et propose plusieurs services en ligne destinés et aux professionnels et au grand public du monde entier. Dans son rapport d'activités des années 1975 à 1977, la Direction Générale XIII de la Commission européenne, en charge de la gestion de l'information et de l'innovation, dont le siège se trouvait également à Luxembourg, annonçait le démarrage d'un projet pilote de traduction automatique avec le logiciel SYSTRAN.

Le logiciel SYSTRAN a été développé par Peter Toma, né en juillet 1924 à Doboz, un petit village hongrois. Marqué par les misères de la deuxième guerre mondiale, il était persuadé que les barrières linguistiques constituaient un frein pour garantir la paix. Après avoir immigré aux Etats-Unis en 1952, il a d'abord travaillé à partir de 1956 comme assistant à l'Institut de technologie de Californie (Caltech), ensuite à Université de Georgetown où il participait aux travaux du groupe GAT (General Analysis Technique, rebaptisé dans la suite Georgetown Automatic Translation).



Peter Toma dans une vidéo de Systran

Peter Toma a été engagé en 1961 par la société Computer Concepts à Los Angeles où il valorisait son savoir-faire pour réaliser les projets AUTOTRAN et TECHNOTRAN qui tournaient sur des ordinateurs IBM 7090. En 1964 Peter Toma présentait les systèmes de traduction automatique à des scientifiques européens à l'Université de Bonn en Allemagne. Le problème était toutefois que l'ordinateur IBM 7090 ne disposait pas de mémoire et de puissance de calcul suffisant pour traduire plus que des courtes phrases. Lors du vol vers l'Allemagne, Peter Toma avait lu le manuel d'instruction du nouvel ordinateur 360, annoncé par IBM. D'emblée, il était persuadé que cet ordinateur permettait de réaliser vraiment ses rêves et il concevait dans sa tête l'idée comment transformer AUTOTRAN et TECHNOTRAN dans un nouveau produit qu'il nommait SYSTRAN.

Comme l'environnement de travail était plus propice en Europe qu'aux Etats-Unis pour progresser avec la traduction automatique, Peter Toma décidait de rester en Allemagne. Il commençait à enseigner la programmation à l'université de Bonn, puis à l'Université de la Sarre. En parallèle, il émulait le fonctionnement de l'ordinateur IBM 360 sur le modèle IBM 7090 pour créer un modèle SYSTRAN embryonnaire. En même temps, il s'était inscrit comme étudiant à Bonn où il achevait son doctorat en 1970 avec une thèse sur la traduction automatique.

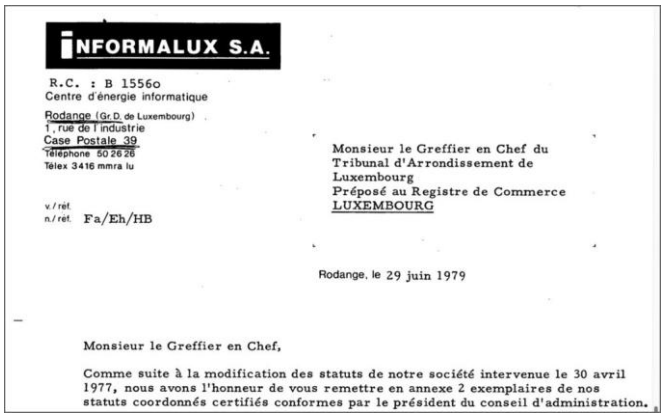
Lorsque Heinz Unger, le directeur du département mathématique à l'Université de Bonn, a obtenu les crédits pour l'installation et l'exploitation d'un ordinateur IBM 360 dans son institut, Peter Toma a pu montrer le fonctionnement correct de son système de traduction SYSTRAN. En 1965, la Fondation allemande pour la recherche (DFG: Deutsche Forschungsgesellschaft) avait invité les meilleurs informaticiens et linguistes allemands pour évaluer le projet SYSTRAN. Après une journée de test et de discussions, une bourse de recherche a été accordée à Peter Toma pour parfaire son système.

En 1967, le laboratoire de Rome, le centre de recherche et de développement de la Force aérienne américaine (RADC), lançait un appel d'offres pour la traduction de textes russes en anglais. Toma soumettait une offre pour son projet SYSTRAN qui a été retenue par le RADC, parmi des concurrents comme IBM et Thompson Ramo Wooldridge.

Ce succès a incité Peter Toma à créer sa propre société en 1968 pour continuer le développement de SYSTRAN, à La Jolla en Californie, sous le nom de Latsec (Language translation system and electronic communications). D'autres contrats ont été conclus avec des clients américains, par exemple avec la NASA. Avec des hauts et des bas, le projet SYSTRAN a été perfectionné, à Bonn et à La Jolla, jusqu'en 1975 quand Peter Toma a fondé le World Translation Center (WTC) pour gérer les contrats SYSTRAN en dehors des Etats-Unis.

En 1973, le Danemark, l'Irlande et le Royaume-Uni ont adhéré à la Communauté économique européenne. L'introduction de l'anglais comme langue majeure et du danois comme sixième langue officielle posait la Commission européenne et l'Office des Publications Européennes devant des problèmes monstrueux. Il fallait engager des centaines de traducteurs diplômés additionnels pour gérer le nombre grandissant de documents à publier et à distribuer. C'était le démarrage de l'introduction d'un système de traduction automatique pour les besoins des institutions européennes et le départ pour la mise en service de SYSTRAN au Luxembourg.

Il existe deux récits de la manière d'adoption de SYSTRAN. Le premier récit crédite aux institutions européennes le don d'une vue à long terme en ayant cherché proactivement un système sur le marché. Le deuxième récit dit que c'était l'initiative de Peter Toma qui cherchait un marché européen pour son produit SYSTRAN. La vérité se trouve probablement au milieu. Il est toutefois établi que Peter Toma présentait en juin 1975 un prototype SYSTRAN anglais-français (développé par sa filiale canadienne WTC-C) à la Commission européenne à Luxembourg et que celle-ci évaluait un deuxième produit appelé TITUS, conçu par l'Institut Textile de France. Le projet TITUS a été retiré pendant les négociations du fait qu'il ne pouvait pas traduire des textes complets. Comme SYSTRAN pouvait être utilisé sans modification sur un IBM 360, ordinateur que la Commission possédait à cette époque, un premier contrat pour l'utilisation du système SYSTRAN par les institutions européennes a été signé fin 1975 avec WTC. Il portait sur l'adaptation du projet pilote anglais-français aux besoins de la Commission, et le développement d'un début de système français-anglais. Le chef de projet de la Commission était Loll Rolling, responsable des développements linguistiques et technologiques dans le domaine de l'information à la DG XIII, la direction générale qui gérait également le projet Euronet-Diane à l'époque.



INFORMALUX S.A.

R.C. : B 15560
Centre d'énergie informatique

Rodange (Gr.D. de Luxembourg)
1, rue de l'industrie
Case Postale 39
Téléphone 50 26 26
Télex 3416 mmra lu

Monsieur le Greffier en Chef du
Tribunal d'Arrondissement de
Luxembourg
Préposé au Registre de Commerce
LUXEMBOURG

v/réf.
n/réf. Fa/Eh/HB

Rodange, le 29 juin 1979

Monsieur le Greffier en Chef,

Comme suite à la modification des statuts de notre société intervenue le 30 avril 1977, nous avons l'honneur de vous remettre en annexe 2 exemplaires de nos statuts coordonnés certifiés conformes par le président du conseil d'administration.

Extrait d'une lettre d'Informalux

Un des premiers contractants externes de la Commission européenne pour le projet SYSTRAN était la société Informalux S.A. domiciliée à Rodange. Elle a été constituée en 1977 avec le slogan « centre d'énergie informatique ». Les actionnaires majoritaires étaient InfoArbed et la Banque Générale. Léonard Siebenaler était le directeur général.

Six traducteurs diplômés des institutions européennes ont été affectés au projet pour assister les informaticiens au développement. Après quelques mois, il ne restait qu'un seul traducteur qui faisait confiance au projet, Ian M. Pigott. Les autres étaient retournés à leurs anciens postes de travail et s'étaient solidarisés avec la majorité des traducteurs et du personnel administratif des institutions européennes qui contestait l'utilité du projet de traduction automatique. Dans la suite Ian M. Pigott a été désigné comme chef de projet SYSTRAN et il contribuait à faire évoluer le système sensiblement. Au début des années 1980, le système permettait de traduire des textes anglais en français et en italien et des textes français en anglais, avec un taux d'erreurs suffisamment bas pour faciliter la tâche des traducteurs humains à produire des traductions avec la qualité voulue. De plus en plus d'employés des institutions européennes commençaient à apprécier cet outil.

Une licence SYSTRAN a été vendue au Japon en 1980, combinée avec la création d'une société locale Systran Corporation. En juin 1982, une société luxembourgeoise Systran International GmbH a été constituée pour commercialiser le système de traduction automatique SYSTRAN. Le registre de commerce et des sociétés luxembourgeois nous renseigne que les 500 parts de la société étaient détenues par le professeur universitaire allemand Helmut Fischer (495 parts) et l'informaticien Cay-Holger Stoll (5 parts), résidant à Gonderange. Ce dernier était membre de l'équipe projet SYSTRAN comme linguiste et figurait comme gérant de la nouvelle société qui a cessé ses activités en 1987.

En 1983, InfoArbed et Informalux ont créé la joint-venture « ECAT – European Center for AutomaticTranslation s.à r.l. ». Chaque actionnaire détenait 2000 parts. En juillet 1984, le capital de cette société a été doublé et les 4.000 nouvelles parts ont été souscrites par Systran International GmbH. Après la liquidation de cette société quelques années plus tard, la dénomination sociale d'ECAT a été changée en TELECTRONICS s.à r.l. en 1991.

En février 1986, une conférence internationale SYSTRAN a eu lieu à Luxembourg. Dans l'introduction, Peter Toma racontait pourquoi il avait demandé un faible montant pour la mise à disposition du logiciel SYSTRAN à la Commission européenne. Il soulignait que son principal objectif était le maintien de la paix et qu'une meilleure communication entre les pays européens, grâce à la traduction automatique, constituait un moyen pour y parvenir. Il annonçait également qu'il avait vendu récemment l'ensemble de ses sociétés et des droits et licences SYSTRAN (à l'exception de la filiale japonaise) à un fabricant de robinets industriels en France, la famille Gachot, qui souhaitait diversifier ses activités. Avec les revenus de cette vente, Peter Toma comptait organiser deux projets de taille dans l'intérêt du maintien de la paix.

En 1986, le siège social de Systran a été transféré à Paris et la gestion de la société mère Systran S.A. a été assurée par Jean Gachot. Son fils Denis Gachot était déménagé en Californie pour diriger les filiales Latsec et WTC à La Jolla. En 1989, Denis Gachot a présenté « The SYSTRAN Renaissance » avec l'annonce de plusieurs nouveautés, entre autres des versions SYSTRAN pour PC, la location du logiciel SYSTRAN, l'accès à distance par Telenet, et même un accès Minitel pour le grand public en France.

Malgré ces innovations, la réussite financière n'était pas au rendez-vous pour la famille Gachot. En 1993, les activités SYSTRAN ont été cédées à quelques anciens actionnaires minoritaires, parmi eux Dimitrios Sabatakakis qui dans la suite a dirigé le groupe SYSTRAN pendant vingt ans.

En 1996 l'équipe de développement et de maintenance SYSTRAN au Luxembourg, composée de linguistes, traducteurs et informaticiens, était passée de deux personnes à une quarantaine. La Commission européenne planifiait de confier l'exploitation du système au Centre de traduction des organes de l'Union européenne (CdT) qui a été établi à Luxembourg en 1994. Le CdT n'était pas attaché à la Direction Générale Traduction de la Commission européenne, mais il disposait de sa propre personnalité juridique.

En mars 1996, une société TELINGUA s.à r.l. a été constituée par Telindus, probablement dans le contexte du remaniement des contrats par les institutions européennes. En septembre 1997, cette société a été convertie en société anonyme Systran Luxembourg par les actionnaires Systran S.A. France (250 parts), Telindus (175 parts), Norbert von Kunitzki (50 parts) et Pierre Musman (25 parts). Le contrat SYSTRAN avec la Commission européenne a été cédé à SYSTRAN Luxembourg.

Au Luxembourg les sociétés Informalux S.A. et Telectronics s.à r.l. ont fusionné en 2001, la nouvelle entité a été dénommée TELECTRONICS S.A. Elle est restée active pendant presque vingt ans et a été absorbée par Proximus Luxembourg S.A. en 2020.

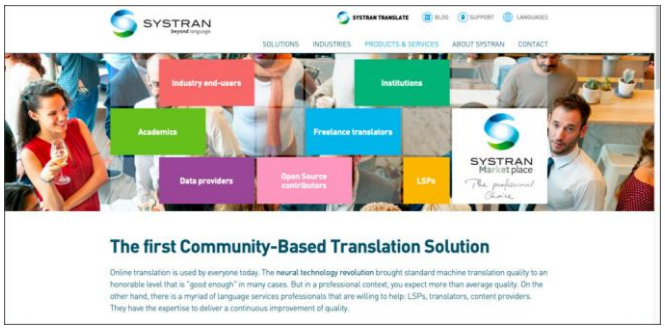
Les institutions européennes avaient lancé en octobre 2003 un nouvel appel d'offres pour l'exploitation du système de traduction automatique sans la prise en compte des intérêts du groupe SYSTRAN. L'appel d'offres a été remporté par la société luxembourgeoise Gosselies S.A. (constituée en 1994 et liquidée en 2011) qui n'avait aucune expertise en matière de linguistique informatique.

En 2004 Pierre Musman est sorti du Conseil d'Administration de Systran Luxembourg S.A. et Norbert von Kunitzki est décédé en 2005 par suite d'une chute dans les montagnes. À partir de ce moment SYSTRAN Luxembourg a été administré uniquement par Denis Gachot et par Dimitrios Sabatakakis (PDG de Systran S.A. France). Guillaume Naigeon de Systran S.A. France assurait la fonction de commissaire aux comptes.

Dimitris Sabatakakis et Park Ki-Hyun

En 2013, la société coréenne SLCI a lancé avec succès une offre publique d'achat (OPA) pour acquérir SYSTRAN. Avec l'obtention de 85% des actions de Systran par SLCI, l'acquisition a été conclue à Seoul en mai 2014. Les administrateurs français ont quitté la société SYSTRAN Luxembourg S.A. et ont été remplacés par les administrateurs coréens Ji Changjin, Kim Dong Pil et Park Ki-Hyun.

Pendant une année Chang-Jin Ji et Guillaume Naigeon assuraient l'intérim de PDG. En juillet 2015, Jean Senelart, un ingénieur et informaticien linguistique de renom, a pris les rênes du groupe SYSTRAN. Il avait rejoint SYSTRAN en 1999, d'abord comme chef de projet, ensuite comme directeur des équipes R&D, avec lesquelles il avait lancé quatre générations de produits SYSTRAN. En 2008, il est devenu Chief Scientist et en 2014 Global CTO du groupe. Mais le jeu des chaises musicales n'était pas encore fini. En août 2020, la majorité des actions du groupe SYSTRAN a été achetée par un consortium d'investisseurs institutionnels coréens : STIC Investments, SoftBank Korea, Korea Investment Partners et Korea Investment Securities. La présidence et direction du groupe ont été confiées en mai 2022 à Vincent Godard. De formation ingénieur, il avait occupé le poste de Directeur Commercial de 2009 à 2013 auprès de SYSTRAN, ce qui facilitait sa prise de fonction. Jean Senelart continue de siéger au Conseil d'administration de SYSTRAN en tant que conseiller scientifique.



Page web d'accueil Systran

En 2022 SYSTRAN Luxembourg S.A. est administrée complètement par SYSTRAN S.A. France, le commissaire aux comptes est la société luxembourgeoise F.C.G. S.A. La filiale luxembourgeoise et la maison mère semblent se porter bien. SYSTRAN est toujours considéré comme pionnier et leader du marché de la traduction automatique, avec des technologies basées sur l'intelligence artificielle très innovantes. Hélas malgré un séjour de presque 50 ans au Luxembourg SYSTRAN n'a jamais appris à traduire le luxembourgeois.

Il reste à signaler que des litiges entre le groupe SYSTRAN et la Commission européenne concernant la violation de droits d'auteur et entre anciens et nouveaux actionnaires de SYSTRAN concernant la non-observation d'obligations ont occupé la Cour de Justice Européenne et les tribunaux nationaux jusqu'à la fin de la dernière décennie, bien que l'exploitation du système auprès des institutions européennes a été arrêtée en 2010, l'année de décès de Peter Toma.

EUROTRA

Lors de la mise en place de SYSTRAN en 1976, il n'y avait pas seulement de la résistance auprès du personnel des services de traduction des institutions européennes, mais également beaucoup de critiques de la part de linguistes dans les universités en Europe qui reprochaient à la Commission européenne d'acheter un produit américain au lieu de promouvoir le savoir-faire européen. Le directeur général de la DG XIII, Raymond Appleyard, et le directeur responsable pour la gestion de l'information à la DG XIII, Georges J. Anderla, étaient sensibles à ces arguments et proposaient le lancement d'un projet européen de traduction automatique, tout en continuant à supporter la solution pragmatique SYSTRAN. Le projet, appelé EUROTRA, n'a été approuvé qu'en 1982 par le Conseil européen et par le Parlement européen et il n'a démarré qu'en 1985. Malgré un investissement de 70 millions d'euros pendant 10 ans, le système n'est jamais devenu opérationnel.

CRETA

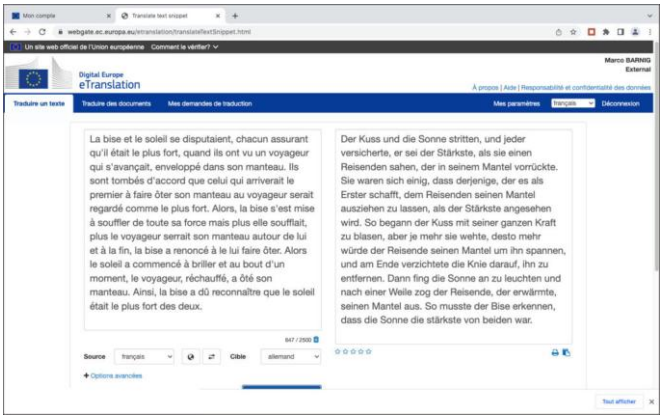
L'Institut Européen pour la Gestion de l'Information (IEGI), créé en 1982 au Luxembourg, participait dès le début au projet EUROTRA. En 1989, l'IEGI a été intégré dans le centre de recherche public du centre universitaire (CRP-CU). Le centre de recherches et d'études en traduction automatique (CRETA) de l'IEGI a été converti en cellule de recherche spécialisée au sein du CRP-CU dans le domaine de la technologie de l'information. Les activités des membres de CRETA se situaient au niveau de la mise en place d'une base de données CRIS et de l'évaluation et test de logiciels utilisés dans le développement informatique du projet par d'autres équipes européennes.

MT@EC

Le nombre de paires de langues supportées par SYSTRAN ne cessait de croître, mais le travail lexicographique sous-jacent rendait difficile l'extension de la couverture aux nouvelles langues résultant des élargissements successifs de l'Union européenne. En décembre 2010, le système SYSTRAN a été arrêté et remplacé par MT@EC, un système de traduction automatique basé sur les statistiques et exploitant les énormes corpus bilingues et multilingues des institutions européennes tels qu'Euramis. S'appuyant sur le système à source ouverte Moses, développé à l'Université d'Édimbourg, MT@EC supportait en 2016 plus que 500 paires de langues, avec des degrés divers de fiabilité et de qualité. Mais la durée de vie du projet MT@EC était largement inférieure à celle de SYSTRAN, le système a été rapidement remplacé par eTranslation.

eTranslation

Comme la technologie de l'intelligence artificielle se développait à un rythme stupéfiant, la méthode de traduction automatique fut surpassée par une nouvelle méthode dès 2017. La traduction automatique neuronale était arrivée, rendue possible par les progrès en matière de puissance de calcul faisant sortir l'intelligence artificielle des laboratoires. Cette nouvelle technologie exigeait encore davantage de puissance de calcul, et plus spécifiquement, des processeurs graphiques. Testé depuis 2019, un nouveau service en ligne appelé eTranslation a été ouvert en mars 2020, non seulement pour les institutions européennes, mais également pour les entreprises européennes et pour des professionnels, et ceci à titre gratuit. Markus Foti est le responsable de ce projet.



Page web de traduction avec eTranslation

J'ai testé en 2022 le service eTranslation avec la traduction française-allemande de la fable d'Ésope « La bise et le soleil se disputaient… ». Cette fable est extraite d'un ensemble de fables en prose qu'on attribue à l'écrivain grec Ésope. Le texte est utilisé régulièrement par des linguistes dans le cadre de projets de recherche sur le traitement du langage naturel (NLP: Natural Language Processing). Elle a été traduite dans des centaines de langues mondiales et régionales. eTranslation propose 3 termes différents pour traduire le mot bise en fonction du contexte : der « Kuss », die « Knie » et der « Bise ». Le modèle de traduction utilisé pour eTranslation ne semble donc pas encore être à la hauteur de ses compères Google, Facebook ou IBM

Si vous souhaitez traduire des textes étrangers en luxembourgeois, c'est en vain. La langue luxembourgeoise n'est pas encore supportée. Il faut s'adresser à Google Translate.

Google Translate

Le service de traduction en ligne Google Translate a été lancé en avril 2006. Il était d'abord basé sur la technologie SYSTRAN. En octobre 2007, Google a remplacé SYSTRAN par sa propre technologie de traduction automatique statistique, supportant 25 langages. En janvier 2010, Google a introduit une version mobile pour Android, une année plus tard une application pour iOS.

Une application de réalité virtuelle pour traduire des mots ou textes, photographiés avec un smartphone ou une tablette, appelée Word Lense, a été intégrée dans Google Translate à la suite de l'acquisition par Google en mai 2014 de la start-up américaine Quest Visual qui a développé ce système.

En février 2016, Google avait ajouté 13 langues additionnelles à son service Google Translate, dont le luxembourgeois, pour supporter alors un total de 103 langues. Au début, ce n'était pas parfait, par exemple « Lëtzebuerg ass mei Land » a été traduit en « Irland ist mein Land ».

En novembre 2016, Google a introduit la version neuronale Google Neural Machine Translation (GNMT) de son service de traduction automatique pour 40 langues. GNMT fait partie du projet Google Brain, lancé en 2011 par Jeff Dean et Greg Corrado des laboratoires de recherche Google X et par Andrew Yan-Tak Ng de l'université de Stanford. En août 2017, le nouveau système a été étendu aux autres langues et le support du luxembourgeois a été bien amélioré.

Le service Google Translate est gratuit pour des traductions en ligne de textes courts, mais c'est un service payant qui fait partie des produits Google Cloud pour des traductions de taille.

Yandex Translate

Les internautes en Russie et dans différents autres pays de l'Est préfèrent utiliser Yandex Translate au lieu de Google Translate. Le service est intégré dans le moteur de recherche de même nom créé en 1997 par Arkadi Voloj. Yandex est une société russe dont la maison mère est basée à Amsterdam. Yandex Translate utilise un système statistique de traduction automatique par auto-apprentissage. Ce qui surprend, c'est que le service supporte la langue luxembourgeoise. Un test de la traduction anglais-luxembourgeois avec la fable d'Esope montre que la qualité de la traduction n'est pas très élevée.

OpenNMT

OpenNMT est le projet de traduction automatique à source ouverte le plus ancien. Yoon Kim, membre du groupe NLP de l'université de Harvard, a développé en juin 2016 un logiciel qui est à l'origine du projet. Yoon Kim est aujourd'hui maître assistant au MIT. En collaboration avec SYSTRAN, un premier modèle OpenNMT a été publié en décembre 2016. L'auteur principal était Guillaume Klein de SYSTRAN. Une première version OpenNMT pour la bibliothèque logicielle Python d'apprentissage machine à source ouverte PyTorch a été publiée en mars 2017 en collaboration avec Facebook Research. Une version pour l'outil d'apprentissage automatique à source ouverte développé par Google, au nom de TensorFlow, suivait en juin 2017. Le projet OpenNMT est actuellement maintenu par SYSTRAN et Ubiqus. L'ancien PDG de SYSTRAN, Jean Senellart, est un des administrateurs du forum OpenNMT.

Des modèles de traduction pré-entraînés pour OpenNMT sont rares. Sur le site web OpenNMT on trouve seulement un modèle anglais-allemand et un modèle allemand-anglais.

MarianNMT

MarianNMT est un outil neuronal de traduction automatique entièrement programmé en C++. Pour cette raison, il est deux fois plus rapide qu'OpenNMT si on effectue des tests dans les mêmes conditions. Le projet a été développé à l'Université d'Édimbourg et à l'université Adam Mickiewicz à Poznań en Pologne, en collaboration avec Microsoft, à partir de 2017.

Le chef de projet est Marcin Junczys-Dowmunt, actuellement scientifique NLP principal auprès de Microsoft. Il est né à Bydgoszcz en Pologne, la ville natale de Marian Rejewski, un mathématicien et cryptologue polonais renommé. En hommage à ce scientifique, Marcin Junczys-Dowmunt a utilisé son prénom comme nom de travail du projet et le nom n'a jamais été changé. La technologie Marian est utilisée par Microsoft pour son service de traduction commercial Microsoft-Translator sur la plateforme Azure. À l'exemple de Google, une version légère est intégrée dans le moteur de recherche Bing de Microsoft pour traduire gratuitement des textes courts. Mais la langue luxembourgeoise n'est pas supportée par Microsoft.

Une multitude de modèles de traduction pour différentes paires de langues a été publiée par l'Université d'Helsinki sous l'initiative de Jörg Tiedemann, professeur des technologies de langage au département des humanités numériques. Ces modèles neuronaux sont connus sous les noms de Helsinki-NLP/opus-mt-xx-yy (xx = langue source, yy = langue cible). Un total de 1.466 modèles a été entraîné avec les données du corpus public ouvert OPUS, respectivement avec les données de la collection Tatoeba, qui comprend également une base de données luxembourgeoise.

J'ai converti le couple de modèles *en-lu* et *lu-en* en PyTorch et j'ai programmé une application pour un espace de démonstration. Hélas, le corpus Tatoeba luxembourgeois n'a ni la taille ni la qualité pour obtenir des résultats de traduction valables. Dans l'attente de la découverte d'une base de données plus appropriée pour pouvoir entraîner un meilleur modèle de traduction anglais-luxembourgeois, et éventuellement des modèles luxembourgeois avec d'autres langues comme source ou cible, je laisse la présente version en place sur la plateforme HuggingFace comme preuve de concept.

Facebook NLLB (No Language Left Behind)

Précédé par son modèle « pytorch/translate », Facebook Research a présenté en 2019 son outil de modélisation de séquences fairseq en source ouverte qui permet, entre autres, de faire de la traduction automatique. À l'instar d'OpenNMT et de MarianNMT, des modèles de traduction automatique pré-entraînés pour quelques couples de langues ont été mis à la disposition des chercheurs.

En octobre 2021 Facebook Inc., la maison mère des services Facebook, Instagram, WhatsApp et Messenger, a changé son nom en Meta Platforms Inc. et Facebook AI est devenu Meta AI. Le 6 juillet 2022, Meta AI a annoncé la mise au point du premier modèle d'intelligence artificielle unique qui permet la traduction en 200 langues différentes avec une qualité de pointe. Ce modèle, appelé NLLB-200, fait partie de l'initiative « No Language Left Behind » de Meta. Le modèle NLLB-200 est disponible en plusieurs versions, avec des tailles allant de 4,4 GB (600 millions de paramètres) jusqu'à 404 GB (54,5 milliards de paramètres.

Comme le modèle NLLB nécessite également un réglage fin pour les langues à ressources limitées comme le luxembourgeois moyennant un entraînement supplémentaire, ce qui n'a pas été réalisé jusqu'à présent par Meta pour l'outil de traduction affiché en bas des contributions publiées sur le réseau social Facebook, ce qui explique la traduction parfois amusante de textes luxembourgeois sur cette plateforme.

Amazon Translate et Sockeye

Le service de traduction automatique commercial d'Amazon repose également sur un projet à source ouverte, appelé Sockeye. Les origines du projet remontent à 2017. Si on passe le texte de « La bise et le soleil se disputaient… » au traducteur Amazon Translate, on s'étonne que la traduction allemande « Der Nordwind und die Sonne stritten sich… » est correcte. Mais à la deuxième vue, on découvre que dans les autres phrases la bise devient der « Kuss » comme dans la majorité des autres systèmes de traduction.

Google T5 et mt5

Les premiers modèles neuronaux entraînés pour la traduction automatique étaient basés sur une architecture de réseau de neurones récurrents. Tout a changé en juin 2017 lorsque Google a présenté une nouvelle architecture de réseau neuronal, appelée Transformer, dans sa publication « Attention is All You Need ». D'un jour à l'autre les transformer, qui sont conçus pour gérer des données séquentielles telles que le langage naturel, sont devenus le modèle de choix pour la traduction automatique. Cela a conduit au développement de systèmes pré-entraînés tels que BERT (Bidirectional Encoder Representations from Transformers) par Google et GPT-1 (Generative Pre-Training Transformer) par OpenAI en 2018.

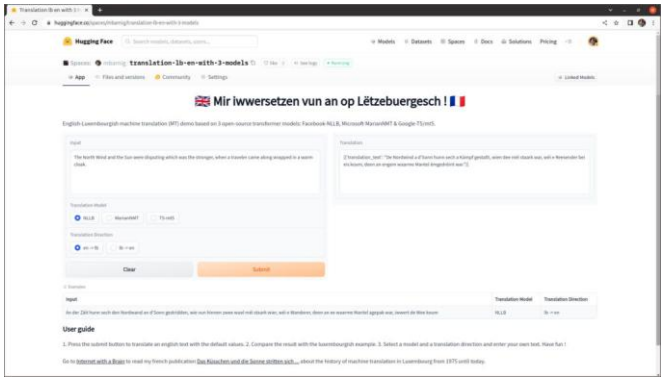
Fin 2019, Google a introduit son transformer T5, pré-entraîné avec une vaste base de données. Le modèle T5 supporte 4 langues (anglais, allemand, français et roumain) et existe en cinq versions. La plus petite comprend 60 millions de paramètres, la plus large dépasse 11 milliards.

Même si l'application principale du modèle T5 n'est pas la traduction, celle-ci fonctionne entre les quatre langues supportées.

Un modèle multilingue mt5 supportant 101 langues, dont le luxembourgeois, a été présenté par Google une année après le modèle T5, en octobre 2020. Comme le modèle mt5 nécessite un réglage fin (finetuning) avant de pouvoir servir pour des applications de traitement du langage, comme la traduction automatique, quelques développeurs ont réalisé des outils informatiques pour faciliter l'apprentissage automatique par les machines.

Comme le réglage fin des modèles T5 et mt5 nécessite des puissances de calcul et des tailles de mémoire très élevées, typiquement une TPU (Tensor Processing Unit) comme accélérateur sur la plateforme de coopération Google Colab, je me suis limité à l'entraînement du modèle mt5-small, qui a une taille de 1,2 GB, avec le corpus Tatoeba anglais-luxembourgeois.

Mir iwwersetzen vun an op Lëtzebuergesch



Application de démonstartion sur la plateforme HuggingFace

J'ai programmé une application comme espace de démonstration sur la plateforme publique de collaboration HuggingFace pour comparer les résultats de la traduction anglais-luxembourgeois et vice-versa avec les trois modèles à source ouverte NLLB, MarianNMT et T5/mt5.

En entrant les textes testés avec la démo HuggingFace dans l'application Google Translate, on se rend compte de la performance supérieure de ce modèle dont le code est fermé et propriétaire.

Si on passe en revue l'histoire de la traduction automatique, on constate une évolution exponentielle. Il y a eu plus d'innovations et de changements au courant des trois dernières années que pendant la période de 50 ans (de 1970 à 2019) précédente.

EUROSCRIPT et les autres agences de traduction

En-dehors des projets publics de traduction automatique et des logiciels à source ouverte, il y a de nombreuses agences de traduction au Luxembourg et à l'étranger qui proposent leurs services de gestion de l'information. Il suffit de faire une recherche sur Google avec le mot-clé « traduction Luxembourg » pour trouver des sociétés comme Alphatrad, DeepL, Eurotraduc, LexiLux, Transat, Translatores, ViaVerbi etc. Il existe même une association luxembourgeoise des traducteurs et interprètes (ALTI), fondée en 2011 dans le but de défendre les intérêts des traducteurs et interprètes professionnels du Grand-Duché du Luxembourg et de faire mieux connaître leur métier.

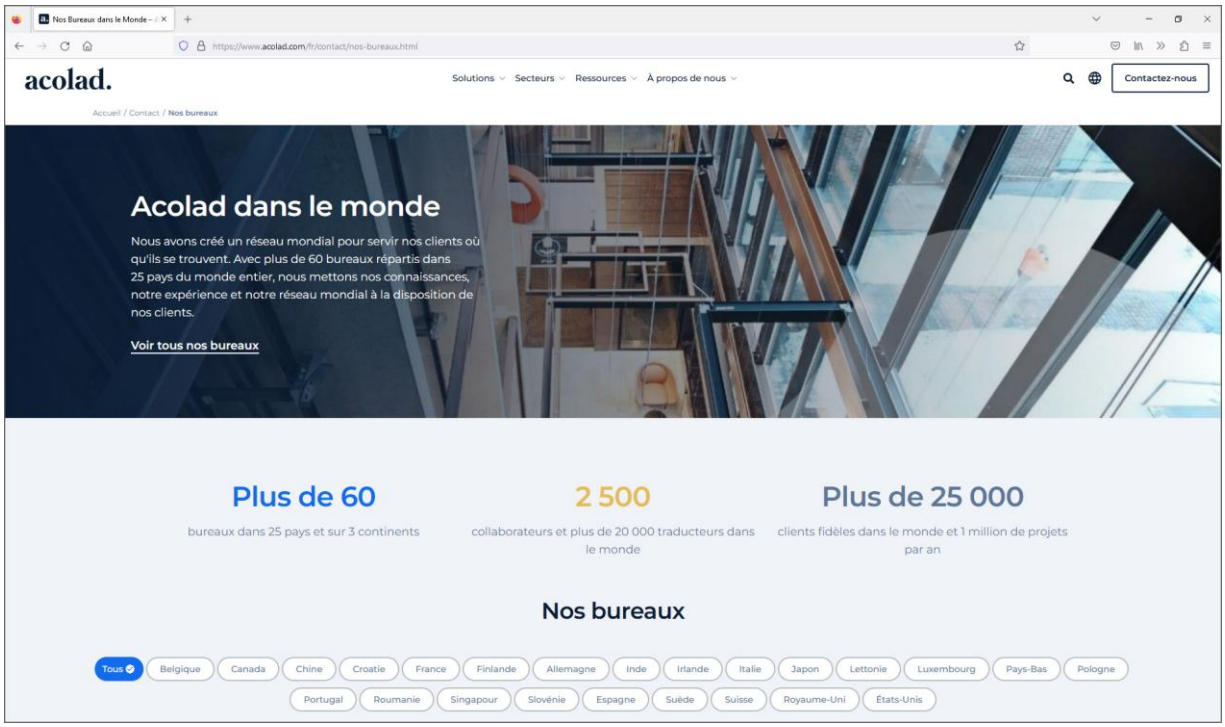
Bâtiment Euroscript à Bertrange

Le doyen parmi les agences de traduction est la société Euroscript Luxembourg s.à r.l. Elle a été constituée le 24 juin 1987 avec siège à Bertrange par les actionnaires « Saarbrücker Zeitung Verlag und Druckerei » et « Heimat-Presseverlag GmbH » de Völklingen. L'objet social de la société était la conception, l'édition et la traduction de documents dans toutes les langues de la communauté. Les institutions européennes étaient le client le plus important. Euroscript a eu une existence mouvementée avec des changements fréquents d'administrateurs et de gérants, des acquisitions de sous-traitants, des fusions avec des partenaires et des licenciements de personnel en cas de non-prolongation d'un contrat européen.

Le premier gérant d'Euroscript domicilié au Luxembourg, était Brigitte Hennemann, de formation ingénieur. Elle a dirigé l'entreprise à partir de 1989 jusqu'en 2008. En 2005, le groupe Euroscript employait 600 personnes en Europe, réalisait un chiffre d'affaires de 45 millions d'euros et comptait des implantations en Allemagne, Belgique, Hongrie, Lettonie, Pologne et Suisse. À l'époque, Euroscript se plaçait en huitième position parmi les 20 plus grands bureaux de traduction au monde. En 2007 Euroscript a fusionné avec la société française Eurodoc pour former la nouvelle entité Euroscript International S.A. Brigitte Hennemann prenait le poste de directeur commercial de la nouvelle entité qui était dirigée par Mark Evenepoel. En 2014 Euroscript International a effectué l'acquisition d'Amplexor, une entreprise spécialisée dans la gestion de contenus basée à Louvain en Belgique.

Une année plus tard, les actionnaires d'Euroscript décidaient d'opérer sous une seule et unique dénomination : Amplexor. Le nom de la société changeait en Amplexor Luxembourg s.à r.l. En 2020

Amplexor s'est rapprochée du groupe français ACOLAD et deux and plus tard Amplexor est devenu ACOLAD, une équipe composée de 2.500 experts et d'un réseau de 20.000 linguistes. Le siège social d'ACOLAD est en France, avec 60 agences dans 25 pays, dont le Luxembourg avec un bureau à l'ancienne adresse d'Euroscript à Bertrange.



Page web d'accueil de Acolad

Fin 2021 la société Amplexor, successeur d'Euroscript, a été rayée dans le registre de commerce luxembourgeois. Dans quelques années, le nom Euroscript sera oublié.

Wordbee

Wordbee est un éditeur de logiciel spécialisé dans l'ingénierie linguistique, la traduction et la terminologie. C'est également une plateforme collaborative innovante d'aide à la traduction et une société de conseil qui accompagne ses clients dans la mise en œuvre de leurs projets de gestion de contenu multilingues et de localisation. La start-up Wordbee SA a été fondée en mars 2008 par José Vega et Stephan Böhmig. C'était à l'époque la 25e start-up accueillie au Technoport à Esch-sur-Alzette et la 13e société hébergée, les 12 autres avaient déjà quitté le Technoport avec succès.

Les responsables du portail de la sécurité de l'information luxembourgeois CASES, qui fait partie du ministère de l'Économie, témoignaient qu'ils utilisaient Wordbee Translator tous les jours pour gérer leur site web multilingue.

José Vega est arrivé au Luxembourg en 1997 pour conduire des projets linguistiques en contact avec les institutions européennes. Il y a cofondé et dirigé des firmes similaires à Wordbee, par exemple myXml et Lucid'IT.

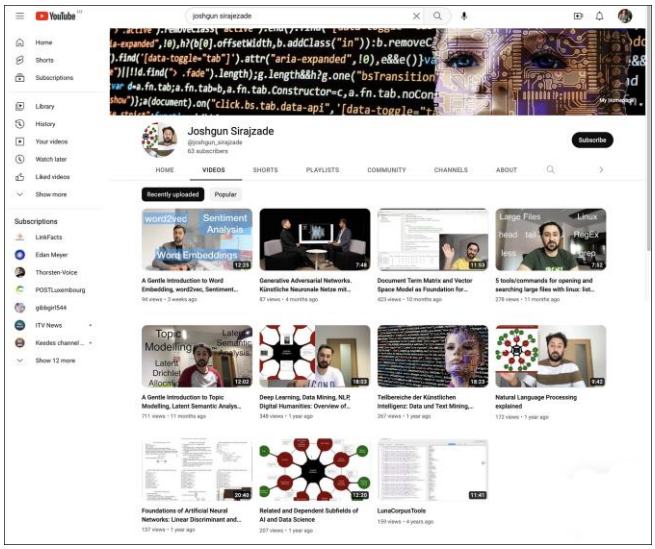
En mars 2011, Wordbee recevait à la CeBIT le prix « European Seal of e-excellence » dans la catégore technologie du langage. En juin 2013, c'était le « LT-Innovate award » qui fût décerné à Bruxelles à la société. En 2015 Wordbee figurait parmi les finalistes de la prestigieuse récompense « Red Herring's Top 100 Europe Award ».

Aujourd'hui, le siège de Wordbee se trouve toujours au Luxembourg, au boulevard du Jazz à Belvaux, et l'entreprise continue à parfaire et à commercialiser ses produits Translator, Beebox et Flex API.

2.1.5. Logiciels de traitement du langage naturel

Pour communiquer avec un ordinateur uniquement en langage naturel, on utilise des logiciels particuliers permettant la compréhension automatique de la langue utilisée. Les technologies à la base de ces outils sont appelées NLP (Natural Language Processing). Il s'agit d'une branche de l'intelligence artificielle (AI). Les algorithmes les plus communs du NLP sont le tokenizing, le part-of-speech-tagging, le stemming, l'analyse du sentiment, la segmentation du topic ou la segmentation de l'unité nommée. Une bibliothèque logicielle de traitement du langage naturel très puissante est le kit NLTK (Natural Language Toolkit : datasciencetest.com/nltk). Dans le cadre de ce livre, je me limite à présenter les bibliothèques qui supportent la langue luxembourgeoise.

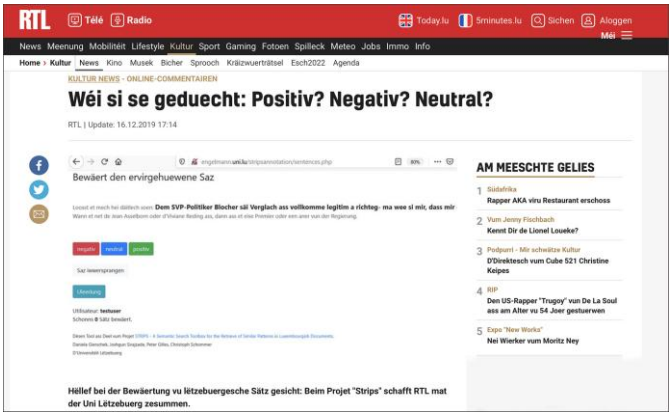
LuNa et STRIPS



Collection de vidéos de Joshgun Sirajzade sur Youtube

Joshgun Sirajzade est chercheur postdoctoral au département des sciences informatiques à l'université du Luxembourg. Originaire de l'Auzerbäidjan, il a poursuivi ses études aux universités de Würzburg et de Trèves. En 2014, il a obtenu un prix promotionnel pour sa thèse au sujet de l'œuvre de Michel Rodange, soutenue à Trèves sous la direction de Claudine Moulin. Au Luxembourg, il s'est focalisé sur la réalisation de projets en relation avec l'intelligence artificielle. Il a publié plusieurs vidéos sur Youtube pour expliquer cette technologie au grand public.

À Trèves Joshgun Sirajzade a développé l'outil informatique LuNa pour annoter et traiter des textes luxembourgeois. Le développement a été achevé à l'université du Luxembourg et couronné par une publication en 2019 avec Christophe Schommer comme coauteur. LuNa était à la base du projet STRIPS (A Semantic Search Toolbox for the retrieve of Similar Patterns in Luxembourgish Documents), développé par Peter Gilles et Christophe Purschke en collaboration avec les experts de l'intelligence artificielle Christophe Schommer et Joshgun Sirajzade. Les travaux se sont déroulés dans la période de 2019 à 2021, en partenariat avec RTL qui a mis à disposition des chercheurs une archive informatique de taille contenant l'intégralité des nouvelles (news > 1998) et des commentaires (> 2008) publiés sur le site web rtl.lu. Pour obtenir des évaluations humaines d'une partie des documents aux fins d'entraîner un modèle d'intelligence artificielle en mode supervisé, un appel à des bénévoles a été lancé sur rtl.lu et infolux.lu pour annoter sur une page web dédiée des phrases présélectionnées dans cette archive. Le premier recours à l'outil STRIPS a eu lieu dans le cadre d'une thèse de doctorat au sujet de la reconnaissance de sentiments dans des textes luxembourgeois.



Page web au sujet de Strips sur RTL News

Spellux

Christophe Purschke, professeur associé pour la linguistique informatique à l'université du Luxembourg, a développé en 2020 une suite de programmes en Python 3, appelée Spellux, pour corriger des textes luxembourgeois dans une application tierce. Je me suis servi de ce logiciel en 2022 pour corriger des traductions automatiques de descriptions anglaises, contenues dans les métadonnées d'images téléchargées sur Internet, aux fins d'entraîner un modèle AI de reconnaissance de photos.

Le code de Spellux est disponible en source ouverte sur le dépôt Github de l'auteur. Dans son projet Christophe Purschke utilise des fonctionnalités de la version 2.2.2 de SpaCy ainsi que le dictionnaire SpellChecker de Pierre Weimerskirch, qui est également publié sur Github.

SpaCy

SpaCy est une bibliothèque NLP développée par Matthew Honnibal et Ines Montani. Il s'agit d'un logiciel multi-plateforme libre, de qualité industrielle, publié sous licence MIT. La première version a été introduite en février 2015, une version 3.5.0 qui supporte 72 langues date du 20 janvier 2023. Depuis la version 2.2.0 SpaCy supporte la langue luxembourgeoise. En novembre 2019, Peter Gilles a présenté les fonctionnalités de SpaCy pour la langue luxembourgeoise sur le site web InfoLux de l'université du Luxembourg.
En 2016, les deux développeurs de SpaCy ont fondé une start-up allemande au nom Explosion AI qui compte aujourd'hui 36 spécialistes. À côté de SpaCy ils proposent les outils « prodigy » et « THiNC ».

CyanogenMod

CyanogenMed n'est pas exactement un logiciel de traitement du langage naturel, mais il est intimement lié aux langues. CyanogenMod était un système d'exploitation de remplacement sur des smartphones et tablettes Android. Il offrait des fonctionnalités et des options indisponibles sur les versions d'Android distribuées par les vendeurs sur leurs appareils.

Développé à partir de 2010 par Stefanie Kondik, CyanogenMod était jusqu'à son arrêt le 24 décembre 2016 le système d'exploitation alternatif le plus répandu pour les smartphones Android.

Le problème majeur était que l'installation était réservée à des technophobes et que l'utilisateur perdait ses droits de garantie sur l'appareil. Au début Google avait demandé la suppression de cette application dans sa boutique en ligne Google Play. Ensuite, les dirigeants de Google avaient lancé une offre d'achat de la start-up Cyanogen, sans succès.\

En 2015 Michel Weimerskirch, avec l'assistance de Sandra Souza Morais, avait traduit les menus de CyanogenMod en luxembourgeois. Le duo de développement a été supporté par Orange Luxembourg S.A., qui proposait l'aide de ses experts pour remplacer le système d'exploitation d'origine par le système d'exploitation luxembourgeois.

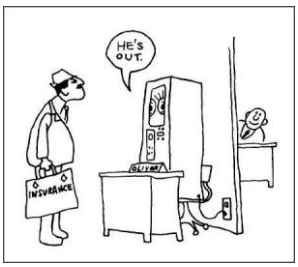
À l'époque Orange a même vendu dans ses boutiques des smartphones Galaxy S4 Black Edition du constructeur Samsung avec le système d'exploitation remplacé.

2.1.6. Outils de traitement de la parole luxembourgeoise

Après avoir présenté les projets en relation avec le langage et la langue au Luxembourg, il convient d'explorer le domaine de la parole luxembourgeoise. Pour s'entretenir avec une machine en langage naturel, elle doit pouvoir parler. La technologie pour le faire est la synthèse vocale, connue sous le sigle TTS (Text-to-Speech). En 2020, j'ai publié un livre sur l'histoire de la synthèse vocale mécanique, électrique, électronique et informatique avec le titre « Synthèse de la parole ». Pour comprendre ce qu'un interlocuteur dit, la machine doit reconnaître la parole. Les technologies afférentes s'appellent STT (Speech-to-Text) respectivement ASR (Automatic Speech Recognition). Pour parler et pour comprendre la langue nationale les ordinateurs doivent maîtriser la phonétique luxembourgeoise.

Mon premier livre

En 1960, J.C.R. Licklider a publié son fameux article au sujet de la symbiose homme-machine qui préfigurait l'informatique interactive. Huit ans plus tard, en 1968, il publiait, ensemble avec Robert W. Taylor, la contribution visionnaire The « Computer as a Communication Device ». Le dessin humoristique à droite, réalisé à l'époque pour cette publication par Rowland B. Wilson, montre l'ordinateur OLIVER qui annonce à un visiteur que son patron n'est pas au bureau.



Ordinateur Oilver en 1968

Si on fait des recherches sur Google ou Bing en relation avec des sujets du domaine de la linguistique informatique, un nom est souvent retourné dans les résultats : Martine Adda-Decker. Elle dirige le laboratoire de phonétique et phonologie à la Sorbonne-Nouvelle à Paris et elle est directrice de recherche au CNRS. Sur le site web « ResearchGate » on trouve une liste impressionnante de 277 publications et 2.897 citations avec son nom, dont plusieurs se réfèrent à des projets portant sur la langue luxembourgeoise. Elle a dirigé et codirigé de nombreuses thèses de doctorat effectuées à différentes universités, parmi eux l'université du Luxembourg.

Dans les prochains sous-chapitres, je vais d'abord introduire les outils de transcription phonétique et ensuite décrire les projets TTS et STT / ASR qui supportent le luxembourgeois.

Transcription phonétique luxembourgeoise

Consonants		
IPA	Examples	English approximation
Native		
b	***B**een* [beːn][1]	**b**all
ɕ	*Ii**ch*t [liːɕt], *Bie**g*** [biəɕ][1][2]	**sh**e, but more of a *y*-like sound
d	***d**ëll* [ˈdɛl][1]	**d**one
f	***F**ësch* [fəʃ][1]	**f**uss
g	***G**ëtt* [gɛt][1]	**g**uest
dz	*spa**ds**éieren* [ʃpɑˈdzəəʁən][1][7]	hea**ds**
dʒ	*Jeans* [dʒiːns][1]	**j**eans
pf	***Pf**licht* [pflict]	cu**pf**ul
w	*zwee* [tswɛː], *Comptoir* [ˈkõːtwaːʁ][8]	**w**e
ʒ	*hé**j**en* [ˈhɛɪzən][1][2]	mea**s**ure, but more of a *y*-like sound

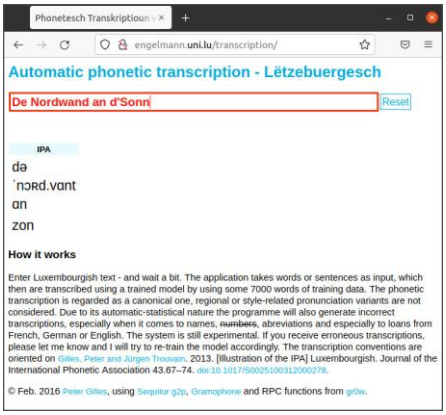
Vowels		
IPA	Examples	English approximation
Monophthongs		
ɑ	*K**a**pp* [kɑp]	**a**rt
ɑː	*K**aa**p* [kɑːp], ***waa**rm* [vɑːm][5]	*Australian* **a**bad
æ	*K**ä**pp* [kæp]	b**a**ck
ə	*F**ë**sch* [fəʃ],[3] *Drog**e**n* [ˈdʁoːgən],[1][10] *Bäcker* [ˈbɔkɐ][11]	*rough like* h**u**rt
ɛ	*dr**é**cken* [ˈdʁɛkən][5]	l**e**t
ɛːɐ̯	*St**ä**r* [ʃtɛːɐ̯][5]	*traditional RP* sq**ua**re
iə	*Ii**e**sen* [ˈliəzən], *Bi**e**rgem* ◀) [ˈbiəʒəm][13]	*roughly like* y**ea**rn
iːɐ̯	***wie**r* [viːɐ̯][5][13]	s**ee** other
oːɐ̯	***Joe**r* [joːɐ̯][5]	*Scottish* n**o** other
uə	*Bu**e**dem* [ˈbuədəm], *Lëtzebu**e**rg* [ˈlətsəbuəɕ][13]	*roughly like* w**or**d
uːɐ̯	***kuer**z* [kuːɐ̯ts][5][13]	t**oo** upbeat

Extrait de l'IPA des phonèmes luxembourgeois

La transcription phonétique consiste à découper la parole en segments sonores insécables et à employer une séquence de symboles unique pour chaque son. Un alphabet phonétique international a été développé par l'association phonétique internationale fondée à Paris en 1986. On utilise le même sigle IPA et pour l'alphabet (International Phonetic Alphabet) et pour l'association (International Phonetic Association). Une première charte de l'alphabet IPA a été publiée en 1988 et révisée une dizaine de fois dans la suite. La dernière révision date de 2020 et comprend 118 symboles graphiques. La combinaison de ces symboles permet de définir la prononciation de toutes les langues du monde.

La documentation de référence pour la phonétique luxembourgeoise constitue la contribution « Luxembourgish » de Peter Gilles et Jürgen Trouvain publiée en 2013 dans le journal de l'association phonétique internationale. La rubrique « Phonetik » sur le portail InfoLux de l'université du Luxembourg présente un ensemble de textes et d'illustrations concernant la structure sonore de la langue luxembourgeoise. Un tableau de supervision globale avec 68 symboles pour spécifier les phonèmes luxembourgeois est présenté sur Wikipédia.

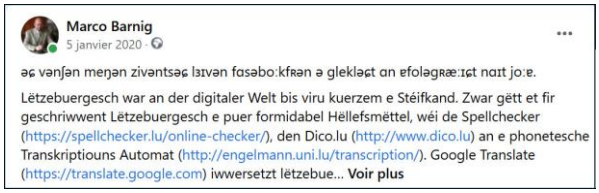
Outils de transcription phonétique automatique



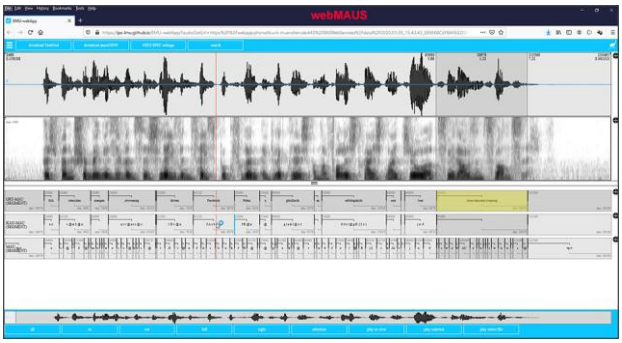
Outil de transcription de Peter Gilles

En février 2016, Peter Gilles a mis en ligne un outil de transcription phonétique automatique sur le serveur de l'université du Luxembourg qui permettait de tester deux méthodes pour effectuer la conversion de textes luxembourgeois en symboles IPA. Ce type de conversion est encore désigné comme « conversion de graphèmes en phonèmes ». Au début, les algorithmes « Sequitor G2P » et « Gramophone » ont été appliqués par Peter Gilles pour entraîner les modèles AI afférents avec ses dictionnaires phonétiques luxembourgeois, disponibles dans son dépôt Github. Aujourd'hui seul l'outil G2P est encore opérationnel sur le site web.

Pour la nouvelle année 2020, j'avais souhaité mes meilleurs vœux à mes amis sur Facebook avec la phrase luxembourgeoise convertie en symboles IPA suivante :« əɕ vənʃən meŋən zivəntsəɕ lɜɪvən fɑsəboːkfʀən ə glekləɕt ɑn ɐfoləgʀæːɪɕt nɑɪt joːɐ ». Je laisse le déchiffrage du contenu exact au lecteur.



Transcription de mes meilleurs voeux pour 2020



Outil de transcription webMAUS

Un outil beaucoup plus sophistiqué qui supporte la langue luxembourgeoise est le service en ligne webMAUS mis à disposition par l'institut phonétique et de traitement de la parole de l'université Louis-et-Maximilien de Munich, fondée en 1472. Cet outil ne supporte non seulement pas la conversion de graphèmes en phonèmes, mais également l'alignement des séquences IPA avec des fichiers audio. Les données et fichiers pour intégrer la langue luxembourgeoise dans webMAUS ont été fournis par Peter Gilles.

Gruut

Dans le prochain sous-chapitre, nous allons découvrir la start-up Coqui AI et son projet Coqui STT qui utilise l'outil de transcription phonétique intégré dans le système de synthèse vocale renommé eSpeak. En printemps 2021, il y a eu des débats dans le forum de discussion de Coqui TTS sur la conformité de la licence eSpeak avec l'utilisation d'une partie du code afférent par une société commerciale. La fondation Mozilla, qui est à l'origine du projet Coqui TTS, n'avait pas ce problème parce que son statut était compatible avec la licence eSpeak.

Le code en question a été retiré immédiatement dans le dépôt Coqui TTS sur Github et le projet se trouvait d'un jour à l'autre sans outil de conversion phonétique. La communauté partait à la recherche d'un logiciel de remplacement. J'avais proposé comme candidat le logiciel Gruut qui fait partie du projet d'assistant vocal « Rhasspy » développé par Michael Hansen, alias « Synesthesian », pendant ses temps de loisir. À l'époque, Michael Hansen travaillait comme assistant scientifique AI dans les laboratoires de recherche des forces aériennes aux Etas-Unis.

« Rhasspy » est un ensemble d'outils pour la création d'un système domotique sous Linux avec assistance vocale. Pour appuyer la candidature de Gruut et pour avancer avec mon propre projet Coqui TTS, j'avais ajouté le code et les fichiers requis pour le support de la langue luxembourgeoise dans cet outil de conversion phonétique. Gruut n'est pas un programme isolé, mais comprend également le logiciel Gruut-IPA et se base sur des paquets de logiciels externes comme « Babel, num2words. pycrfsuite et pydateparser ».

L'intégration de mon code gruut-lb dans le logiciel maître a été effectuée par Synesthesiam le 6 décembre 2021. Depuis cette date le luxembourgeois figure parmi les autres langues supportées par Gruut : arabe, tchèque, allemand, anglais, espagnol, persan, français, italien, néerlandais, russe, suédois et swahili.

2.1.7. Outils de synthèse vocale luxembourgeoise

J'ai toujours été fasciné par la synthèse vocale. En 1976, j'ai supervisé un travail de diplôme à l'institut d'électronique de l'EPFZ, réalisé par Kurt Mühlemann. L'objectif était de créer un circuit de synthèse vocale avec des filtres réglés. À la fin, le synthétiseur était capable de prononcer la phrase « Ich bin ein Computer » avec une tonalité robotique.

Aujourd'hui, les sociétés GAFAM (acronyme pour Google, Apple, Facebook, Amazon et Microsoft) proposent des services cloud de synthèse de la parole avec des voix artificielles ayant une qualité qui ne permet plus de faire la différence avec une voix humaine. Hélas, la langue luxembourgeoise n'est pas encore supportée par ces grands groupes.

Les modèles d'intelligence artificielle TTS afférents sont entraînés en mode apprentissage profond (deep learning) avec des larges corps de textes orthographiques, sans passer par une conversion phonétique. Il est évident qu'une machine qui sait apprendre la conversion de graphèmes en phonèmes, suivi d'un apprentissage de conversion de phonèmes en sons, peut aisément sauter une étape et apprendre la conversion de graphèmes en sons sans le recours à des phonèmes.

Dans quelques années, les informations auront oublié l'alphabet IPA et seuls quelques linguistes vont continuer à s'en servir dans le cadre de leurs travaux de recherche sur les langues. Mais le chemin pour arriver à ce stade était long.

Dans le monde académique les pionniers de la synthèse vocale informatique sont les universités d'Édimbourg en Écosse (> 1984), l'université Carnegon Mellon (CMU) à Pittsburgh aux États-Unis (> 1986), l'université de Mons en Belgique (> 1995), l'université de technologie Nagoya au Japon (> 1995) et l'université de la Sarre en Allemagne (> 2000).

Parmi les GAFAM, il faut souligner le rôle pionnier de Microsoft avec son interface SAPI (Speech Application Program Interface) introduit en 1995. Parmi les pionniers individuels de la synthèse vocale informatisée deux noms sont à mettre en exergue :

- Dennis H. Klatt pour le développement du programme « KlatTalk » présenté en 1982 qui est à l'origine de l'équipement populaire DECTalk, commercialisé par la société Digital Equipment Corporation (DEC) à partir de 1984
- Alan W. Black pour le projet Festival (> 1996), la fondation de l'entreprise CEPSTRAL (> 2000) et le lancement du concours TTS annuel « Blizzard Challenge » (> 2005)

En ce qui concerne les programmes de synthèse vocale qui n'ont pas été développés ni par des universités ni par des grandes entreprises, mais par un développeur indépendant, un nom est à retenir : Jonathan Duddington. Dès 1995, il a démarré le développement du logiciel « Speak » destiné aux ordinateurs personnels ACORN avec le système d'exploitation RISC.

Je vais commencer à présenter le logiciel Speak même si d'un point de vue chronologique ce n'était pas le premier projet de synthèse de la parole qui supportait la langue luxembourgeoise.

Évolution de Speak vers eSpeakNG-lb

Page web eSpeak en 1997 sur la Wayback Machine

Sur « l'Internet Wayback Machine », on trouve l'ancienne page web de Jonathan Duddington (version du 14 juin 1997) où il annonce la disponibilité d'une nouvelle version pour les ordinateurs RiscOS 3.0 au prix de 19,50 £. Une version 2 améliorée a été annoncée en 1999. À côté du module TTS, il y avait un éditeur de parole, un éditeur de phonèmes, un module « Talk as you Type » et un outil de vérification des règles phonétiques.

À l'époque Jonathan Duddington a présenté son logiciel aux membres de différents clubs d'utilisateurs d'ordinateurs ACORN dans la région du Coventry au Royaume-Uni, par exemple au Derbyshire Acorn RISC Club (DARC), au ARM Club, au Manchester Acorn User Group et à la RISC OS 99 Show.

Speak utilisait la méthode de synthèse par formants. L'application comprenait plusieurs voix masculines et féminines, mais seulement pour la langue anglaise. Vers 2006 Jonathan Duddington portait Speak sur Linux et offrait le code en source libre sur la plateforme Sourceforge. La première version disponible était speak-1.5, publiée le 17.2.2006. À côté de la langue anglaise, cette version supportait l'esperanto et l'allemand. Plusieurs nouvelles versions ont été ajoutées au fil de l'année. En août 2006, l'application a été renommée eSpeak et un site web dédié afférent a été mis en ligne. eSpeak supportait des langues supplémentaires, l'africain, le grec, l'espagnol, l'italien et le polonais.

À partir de la version 1.16, les fichiers avaient également été renommés espeak. Des nouvelles versions ont été publiées à un rythme effréné jusqu'à la version espeak-1.48.4 en mars 2014. Cette version supportait 45 langues ainsi que les voix Mbrola.

eSpeak jouissait d'une très grande communauté et Jonathan Duddington était très engagé à fournir son aide aux utilisateurs et développeurs inscrits sur les listes de messages du projet. Pendant les quelques années, il avait édité plusieurs centaines de réponses et d'annonces.

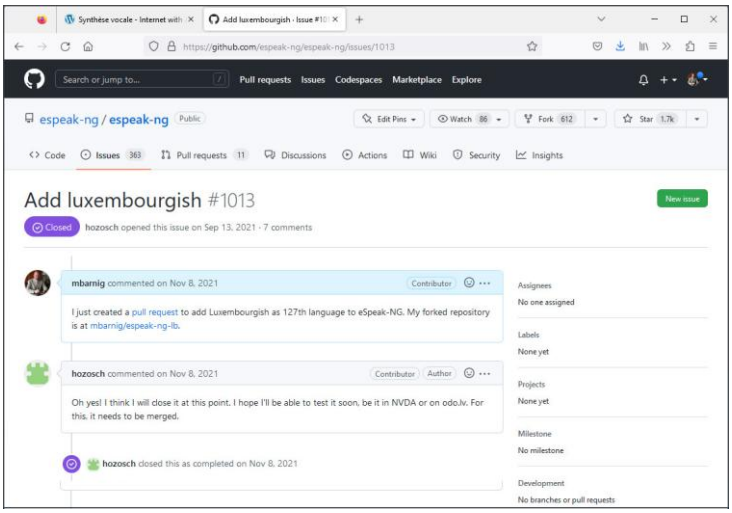
En septembre 2014, j'avais démarré un projet d'intégration de la langue luxembourgeoise dans le logiciel eSpeak. Pour ajouter une nouvelle langue, il fallait ajouter plusieurs fichiers spécifiques et modifier quelques lignes de code dans le logiciel de base. Comparé à la complexité des logiciels de synthèse vocale contemporains, basés sur des réseaux neuronaux, le fonctionnement d'eSpeak est assez simple. Les phonèmes des voyelles sont convertis en sons moyennant des générateurs de tonalités en spécifiant la fréquence et la durée. Pour les voyelles composées comme « ei », « au » ou « oi » on indique les voyelles au début et à la fin et la durée totale. Pour produire des consonants, on utilise deux méthodes. La première consiste à utiliser des générateurs de bruit avec l'application de filtres numériques. Pour certaines voyelles, on utilise simplement des sons pré-enregistrés qui sont sauvegardés dans des fichiers wav. Il suffit alors de lire ces fichiers au bon moment et de les restituer avec une durée indiquée. La description détaillée du fonctionnement d'eSpeak dépasse largement le cadre du présent livre, mais pour se faire une idée du codage, je présente ci-après quelques extraits du fichier « ph_letzebuergesch » qui spécifie la génération des sons dans eSpeak :

Intro	Vowels	Consonants

// Lëtzebuergesch
// virtual class of vowels :
#@, #a, #e, #i, #o, #u
// IPA Vokaler (20) :
ɑ, aː, ɛː, e, æ, eː, ə, ɐ, i, iː, o, oː, u, uː, y, yː, ɑ̃ː, ɛ̃ː, õː, œː
// IPA Vokalkoppelen (9) :
æːɪ, ɑʊ, æːʊ, ɑɪ, ɜɪ, oɪ, iə, əʊ, uə
// Konsonanten : Total : 27
// Nasal Phonemen :
m, n, ŋ
// Plosiv Phonemen :
p, b, t, d, k, g
// Affricate Phomenen :
ʦ, dʒ
// Frikativ Phonemen :
f, v, w, s, z, ʃ, ʒ, X, ɕ, ʁ, ʑ, h
// Approximant Phonemen :
l, j
// Trill Phonem :
ʀ
	phoneme a // K[a]pp ;
	kurz geschwate Vokal a
	vwl starttype #a endtype #a
	ipa ɑ
	length 120
	FMT(vowel/a)
	endphoneme
	
	phoneme a: // K[a]p ;
	laang geschwate Vokal a
	vwl starttype #a endtype #a
	ipa aː
	length 190
	FMT(vowel/aa_6)
	endphoneme
	
	phoneme aE // St[ä]ren
	vwl starttype #a endtype #e
	ipa ɛː
	length 190
	FMT(vdiph/ae_2)
	endphoneme
		phoneme m
		vcd blb nas
		ipa m
		FMT(m/mj)
		endphoneme
		
		phoneme h
		vls glt apr
		ipa h
		IF nextPh(#@) THEN
		WAV(h/h@)
		ELIF nextPh(#a) THEN
		WAV(h/ha)
		ELIF nextPh(#i) THEN
		WAV(h/hi)
		ELIF nextPh(#o) THEN
		WAV(h/ho)
		ENDIF
		endphoneme

Un deuxième fichier au nom de « lb_rules » règle la conversion des graphèmes en phonèmes. Le fait qu'eSpeak ne maîtrise pas l'alphabet IPA (sauf dans les commentaires) et qu'il faut donc définir son propre alphabet pour désigner les phonèmes ajoute une difficulté supplémentaire à la tâche de programmation. Il faut se rappeler que les origines d'eSpeak remontent à 1995 où la majorité des informaticiens ne connaissait que l'ASCII (American Standard Code for Information Interchange). Un troisième fichier s'appelle « lb_dict ». C'est un dictionnaire qui contient les exceptions par rapport aux règles de conversion définies, notamment des noms, des nombres, des abréviations et des mots compliqués. Un quatrième fichier « lb_emoji » ajoute la cerise sur le gâteau et contient la prononciation pour des pictogrammes, pour des symboles et, comme le nom le dit, pour des émojis. Pour illustrer le tout, j'ai assemblé le tableau suivant avec quelques extraits de ces trois fichiers :

lb_rules	lb_list	lb_emoji
.group a a a // a : IPA ɑ _) an (_ an // an : IPA ɑn) af (_ aːf // Af ; IPA aː C) a (C_ aː // Kap : IPA aː a (CC a // Kapp, blann : IPA ɑ a (CA aː // Fabel IPA : aː age arR@E aller al@Er awer aːv@Er .group aa aa aː // aacht, naass : IPA aː aach aː	Marco marko: Barnig barniS Simone zimon _3 draI _4 vOIeR Aachtasiwwenzeg aːxtaziv@EnTS@EX antiquitéiten eAntikitOIt@En naturmusée natu:rmy:ze: ouerestääbchen oU@Er@EStaEpX@En pedalléieren p@EdalOI@Er@En virfinanzéieren fi:rRfinanTSOI@Er@En	& an € eUro ⌨ tasta:tur 👲 kapuzin 🧒 tessa 🎪 TSirkus 🦍 gorila: 🦒 giraf 🐘 elefant 🦏 rino:zerus .....



Page web Github pour ajouter le luxembourgeois dans eSpeakNG

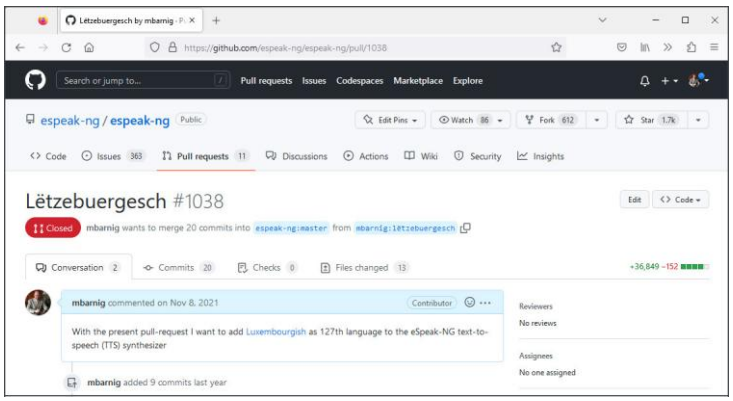
Au courant de 2015, quand mes fichiers étaient finalisés et testés, Jonathan Duddington était aux abonnés absents. Son dernier message date du 16 avril 2015. Depuis ce jour, il ne répondait plus aux questions. À la suite de l'allongement du silence de Jonathan Duddington, la communauté eSpeak commençait à s'inquiéter. Malgré plusieurs tentatives, personne n'a réussi à savoir ce qui s'était passé avec le développeur et on ne trouvait plus trace d'aucune biographie sur lui sur le net.

Après de longs débats au sein de la communauté eSpeak vers la fin de l'année 2015, la candidature de Reece Dunn comme nouveau coordinateur du projet eSpeak, rebaptisé eSpeakNG (nouvelle génération), a été approuvée. Reece Dunn était un contributeur de longue date au projet eSpeak et il l'avait porté sur Android. Son répositoire sur la plateforme Github, créé en mars 2013 pour le développement d'eSpeak pour Android, a été utilisé dans une première phase pour héberger le projet eSpeakNG. Un autre développeur, Alon Zakai alias kripken, a converti eSpeak en Javascript moyennant l'outil Emscripten.

À l'époque, j'avais perdu l'intérêt dans le projet d'intégration du luxembourgeois dans eSpeakNG qui supportait alors déjà 106 langues. La raison n'était pas seulement la présentation fin 2015 de MaryLux, le premier projet TTS luxembourgeois opérationnel, mais surtout à cause du fait qu'il fallait adapter mes fichiers aux modifications apportées entretemps à la nouvelle génération d'eSpeak.

Tout a basculé en septembre 2021 quand Eric Röder, un utilisateur allemand d'eSpeakNG, a demandé dans le forum de discussion du projet le support de la langue luxembourgeoise pour les besoins de son ami luxembourgeois aveugle. Même à ce jour eSpeakNG est le seul outil qui fonctionne sur toutes les plateformes et qui est compatible avec l'application libre et gratuite NVDA (Non Visual Desktop Access), utilisée par les aveugles et malvoyants. Le logiciel NVDA permet d'avoir accès à une lecture d'écran grâce à la synthèse vocale et à des fonctionnalités d'impression en braille.

Après un échange de courrier avec Eric Röder, je fus motivé à finaliser mes travaux entamés en 2014. J'ai pu bénéficier de plusieurs changements intervenus au courant des sept années précédentes. Entretemps, j'avais gagné en expérience dans l'utilisation des plateformes de collaboration et de partage de logiciels comme Github grâce à la réalisation d'autres projets. J'ai également pu profiter des dictionnaires LOD et SpellChecker avec transcriptions phonétiques publiés par Peter Gilles sur Github en 2019 dans son dépôt public « Luxembourgish language resources ». À l'aide de ces ressources, j'ai pu assembler un dictionnaire corrigé et étendu avec les symboles eSpeak pour les phonèmes, convertis à partir de l'alphabet IPA contenu dans les fichiers sources. Ainsi mon fichier « lb_dict » ne contenait pas seulement des exceptions aux règles « lb_rules », mais presque l'intégralité des mots luxembourgeois, ce qui a amélioré considérablement la transcription des graphèmes en phonèmes dans l'outil eSpeak.



Page web Github avec pull-request lb

Une troisième raison pour me remettre au défi d'ajouter le luxembourgeois dans l'écosystème eSpeakNG était mon engagement dans un nouveau projet de synthèse vocale à base d'intelligence artificielle, appelé Coqui-TTS. Ce système se basait sur eSpeak pour la transcription des textes en phonèmes.

Le 8 novembre 2021, j'ai lancé ma demande de « pull-request » pour ajouter mes fichiers luxembourgeois dans le répertoire central d'eSpeakNG. Quelques jours plus tard, Valdis Vitolins, alias valdisvi, une des chevilles ouvrières de la communauté eSpeak, procédait à quelques corrections et exécutait l'intégration de notre langue nationale. Le 11 novembre 2021, Lëtzebuergesch figurait comme 127e langue dans l'application eSpeakNG. J'avais annoncé cet événement sur les réseaux sociaux avec les exemples suivants :

An der 🕳 hunn sech den 🧭 🍽 an d'🤗 gestridden, wie vun hinnen zwee wuel méi 💪 wier, wéi e 🚶 , deen an ee waarme 👗 agepak war, iwwert de 🛏 koum.

Haut sinn ☝ mat mengen Enkelkanner 👫 , 👴 , 👧 , an 👩 an den 🎪 gaangen. Do hunn mer e 🐂, eng 🦒, en 🐘 an en 🦏 gesinn.

MaryLux

MaryLux est le nom de la voix synthétique du projet MaryTTS démarré à l'université de la Sarre en 2000. Le nom Mary est l'abréviation de « Modular Architecture for Research in sYnthesis ». Au début, c'était l'institut de phonétique de l'université de la Sarre et le laboratoire des technologies du langage du centre de recherche allemand pour l'intelligence artificielle (DFKI) qui étaient à l'initiative du projet. Marc Schröder et Jürgen Trouvain ont présenté le projet lors du quatrième atelier de travail de l'association internationale des communications vocales (ISCA) qui a eu lieu du 29 août au 1er septembre 2001 à Pertshire en Ecosse.

MaryTTS est programmé en Java et supporte la synthèse par sélection d'unités, la synthèse paramétrique statistique (HMM) et l'apprentissage moyennant des réseaux neuronaux. Au début, le système ne comprenait que les voix MBROLA. Dans la suite, l'anglais a été ajouté en utilisant une partie du code du projet à source libre FreeTTS, qui lui-même est un portage du projet FLITE de Alan W. Black.

D'autres langues ont été ajoutées et MaryTTS est le premier système de synthèse de la parole qui supporte le luxembourgeois. Le projet MaryLux a été présenté lors du 5e congrès de l'IGDD (Internationale Gesellschaft für Dialektologie des Deutschen) qui a eu lieu du 10 au 12 septembre 2015 à l'université du Luxembourg. Les auteurs étaient Ingmar Steiner, Jürgen Trouvain, Judith Manzoni et Peter Gilles. Le projet MaryLux a en outre été présenté le 10 novembre 2015 dans un colloque interne à l'université (InfoLux) et en avril 2016 sur le portail sciences.lu.

La voix synthétique de MaryLux est celle de Judith Manzoni. Elle a enregistré en 2014 des textes en luxembourgeois (63 minutes), allemand (22 minutes) et français (47 minutes), qui ont été ensuite traités par les différents outils informatiques du système MaryTTS pour créer le modèle de la voix, avec la méthode de sélection d'unités. Cette technique est devenue obsolète aujourd'hui, suite à la progression fulgurante des technologies d'apprentissage approfondi des machines (deep machine learning) sur base de réseaux neuronaux.

Coqui TTS

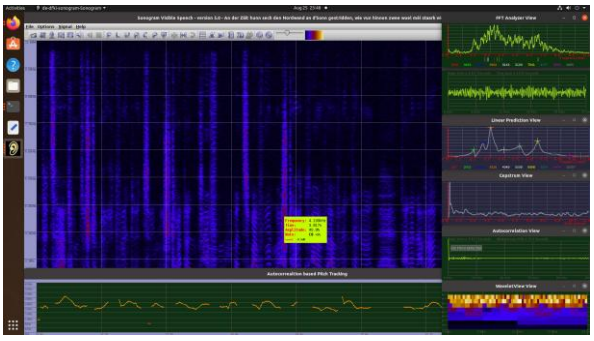
La fondation Mozilla, qui est surtout connue pour son navigateur Firefox, est également un pionnier de la synthèse (Mozilla TTS) et de la reconnaissance (Mozilla-Deepspeech) de la parole. Mozilla a lancé en juillet 2017 le projet Common Voice pour collecter de vastes échantillons de données vocales. Hélas, le luxembourgeois ne fait pas encore partie des langues supportées. En août 2021, la CEO de Mozilla informait son personnel que les activités de technologie de la voix seraient abandonnées au profit d'une focalisation sur Firefox. 250 personnes ont été licenciées. Dans la suite quelques anciens employés de Mozilla ont fondé en mars 2021 la start-up allemande Coqui AI sur les ruines des anciens projets vocaux de Mozilla. Le nom Coqui est dérivé d'une petite grenouille au même nom, endémique à Puerto Rico, qui est connu pour ses cris clairs et forts.

Les idées avancées par les fondateurs de Coqui AI m'avaient incité à reprendre mes anciens travaux de développement d'une voix synthétique luxembourgeoise. J'ai résumé cette histoire en juin 2021 dans mes contributions « Text-to-Speech sound samples from Coqui-TTS » et « The best of two breeds » sur mon site web. Ces travaux sont également à l'origine de l'intégration du luxembourgeois comme 127e langue dans l'application eSpeakNG décrite ci-avant.

Les premiers résultats tangibles étaient prêts en janvier 2022 et ont été publiés sur mon site web dans les articles « Mäi Computer schwätzt Lëtzebuergesch » et « Mäi Computer schwätzt Lëtzebuergesch mat 4 Stëmmen .

Au début, je me suis basé sur l'architecture Tacotron. Google a présenté ce nouveau système de synthèse vocale qui repose sur la superposition de deux réseaux neuronaux lors de la conférence Interspeech 2017 à Stockholm. Les résultats étaient proches d'une prononciation par des humains. La publication académique afférente porte les noms de 14 auteurs. Depuis cette date, ce modèle a été perfectionné et de nouveaux modèles neuronaux TTS ont été développés : Tacotron2-DCA, Tacotron2-DDC, GlowTTS, Fast-Pitch, Fast-Speech, AlignTTS, Speedy-Speech, VITS, … Depuis 2021, on a pu découvrir tous les quelques jours une nouvelle publication scientifique au sujet de TTS dans l'archive ouverte de prépublications électroniques ArXiv sur Internet.

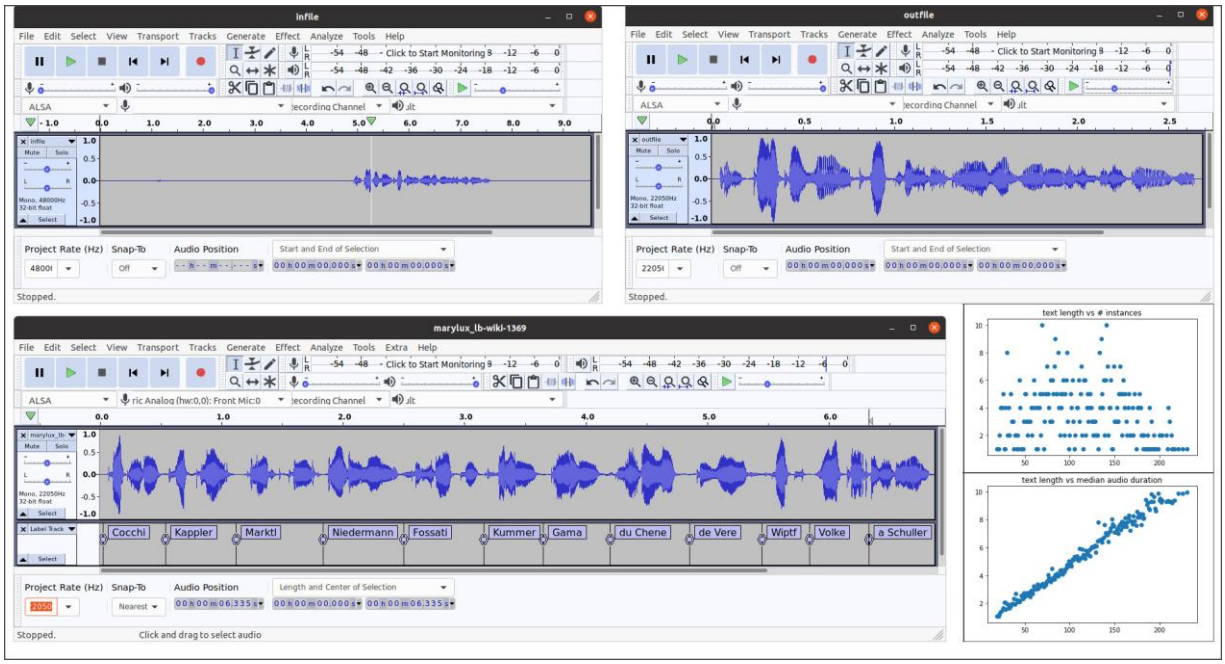
Comme les modèles TTS neuronaux sont gourmands en données et entraînés en général avec des enregistrements audio d'une durée de plusieurs dizaines d'heures, j'ai exploré dans une première phase les contraintes du modèle TTS Tacotron.



Sonogramme d'un enregistrement luxembourgeois

Pour démarrer l'entraînement d'un modèle Tacotron luxembourgeois, j'ai utilisé les données de la base de données MaryLux qui sont disponibles sur Github et régies par une licence permissive « Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License ».

J'ai ajusté les fichiers correspondants pour les adopter d'une façon optimale aux besoins de l'apprentissage automatique profond (deep machine learning). Il fallait changer la fréquence d'échantillonnage de 48 KHz à 22.050 Hz respectivement à 16.000 Hz, convertir le format audio FLAC en WAV, éliminer les silences, normaliser les amplitudes, couper les enregistrements en segments ayant des durées entre 2 et 10 secondes, adapter et corriger les transcriptions, convertir les textes en phonèmes et supprimer les enregistrements qui s'écartaient trop de la déviation moyenne, le tout avec un mélange de procédures manuelles et automatisées. La figure qui suit donne un aperçu visuel sur les outils et métriques employés pour effectuer ces manipulations.



Outils et métriques pour optimiser les enregistrements audio

Pour me conformer à la licence des données MaryLux, j'ai publié la nouvelle base de données publique sous le nom de « Marylux-648-TTS-Corpus » sur mon dépôt de développement Github. Il s'agit de 648 clips audio luxembourgeois, chacun ayant une durée inférieure à 10 secondes, et des transcriptions y associées. La durée totale est de 57 minutes et 31 secondes.

Pour entraîner un modèle TTS avec l'apprentissage automatique profond, on a besoin de calculateurs puissants. J'ai utilisé les infrastructures suivantes :

- Mon ordinateur personnel avec carte graphique NVIDIA RTX 2070, système d'exploitation Linux Ubuntu 20.4 et système de développement Python 3.8
- Mon compte Google-Colab pro dans les nuages, avec CUDA P100 et système de développement Python 3.7

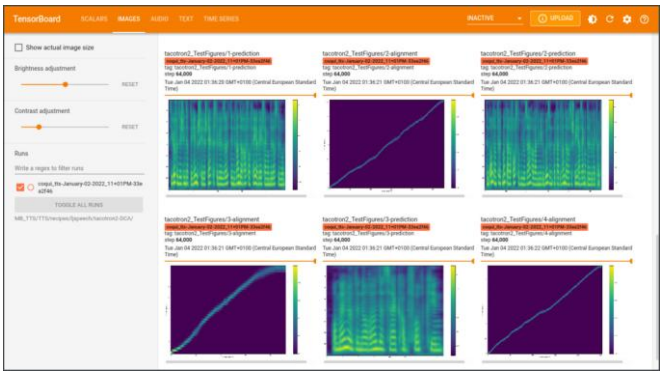
L'entraînement d'un modèle TTS avec une nouvelle base de données peut se faire à partir de zéro (from scratch) ou à partir d'un modèle existant. Dans le deuxième cas, on parle de transfert d'apprentissage (transfer learning) ou de fin réglage (fine tuning).

Les 648 échantillons de la base de données Marylux-648 ont d'abord été mélangés (shuffling), puis répartis en 640 exemples pour l'entraînement proprement dit et en 8 exemples pour l'évaluation, effectuée après chaque cycle d'apprentissage. Les six phrases de la fable « De Nordwand an d'Sonn », qui ne faisaient pas partie du jeu d'apprentissage, ont été utilisées pour les tests automatiques réalisés après chaque évaluation.

Un cycle d'apprentissage complet est appelé une époque (epoch). Les itérations sont effectuées par lots (batches). La durée d'une itération est fonction de la taille du lot. On a donc intérêt à choisir une taille élevée pour un lot. La différence s'exprime par des durées d'apprentissage de plusieurs heures, jours, semaines ou voir des mois. Hélas, la taille des lots est tributaire de la taille de mémoire disponible sur la carte graphique (CUDA).

Sur mon ordinateur personnel, je ne pouvais pas dépasser une taille de lot supérieure à 10 sans provoquer une interruption de l'entraînement à cause d'un débordement de la mémoire (memory overflow). Sur Google-Colab, je ne pouvais guère dépasser une valeur de 32. Sur mon ordinateur personnel, une époque prenait donc 64 itérations, sur Google-Colab, le nombre se réduisait à 20. Le temps d'exécution d'une époque était en moyenne de 95 secondes sur mon ordinateur personnel, ce qui faisait environ 26 heures pour l'entraînement complet d'un modèle TTS avec Marylux-648 (64.000 itérations).

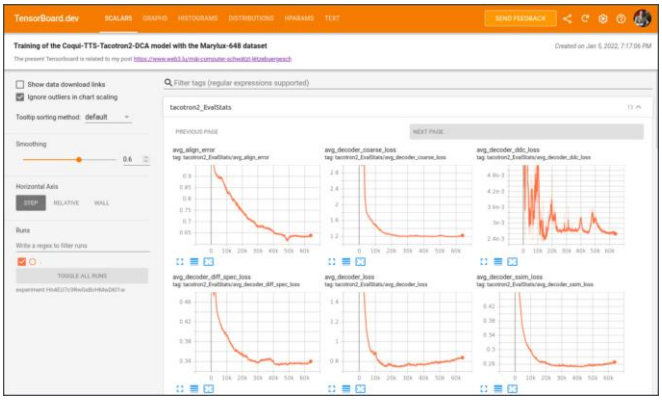
Avec un lot de 32, on s'attend à une réduction du temps d'entraînement d'un facteur 3.2, c.à.d. à environ 8 heures. Or, à cause du partage des ressources entre plusieurs utilisateurs sur Google-Colab, le gain est plus faible. J'avais observé un temps de calcul moyen de 72 secondes par époque, ce qui donnait une durée totale d'entraînement d'environ 20 heures pour 1000 époques (20.000 itérations).



Outil Tensorboard de Google pour suivre entraînement AI

L'entraînement d'un modèle TTS se base sur la comparaison de spectrogrammes, générés à travers un réseau neuronal, avec ceux calculés à partir des échantillons originaux. On peut suivre le progrès de l'entraînement à l'aide de métriques indiquées dans le terminal de gestion ou affichées sous forme graphique en temps réel dans l'outil informatique TensorBoard, développé et mis à disposition par Google. Un alignement diagonal de la courbe dans le graphique « attention » et la similitude des spectrogrammes sont un bon signe pour un apprentissage correct. Les graphiques avec les scalaires relatifs aux pertes, aux justesses et à d'autres paramètres d'apprentissage machine fournissent une vue plus fine sur les résultats.

Dans une première étape, j'avais utilisé le modèle Tacotron2-DCA (Dynamic Convolution Attention) à source ouverte mis à disposition par Coqui AI. Il s'agit d'une nième évolution du premier modèle TTS Tacotron avec une nouvelle architecture qui est configurée par un ensemble de plus de 150 hyperparamètres, en format JSON.



Outil Tensorboard de Google pour suivre apprentissage AI

Les premiers essais de synthèse vocale n'étaient pas fameux. Malgré mes efforts d'obtenir des meilleurs résultats en modifiant certains paramètres suivant une procédure « essai-erreur » (trial and error), avec des attentes de 26 heures entre chaque test, il n'y a pas eu d'évolution positive.

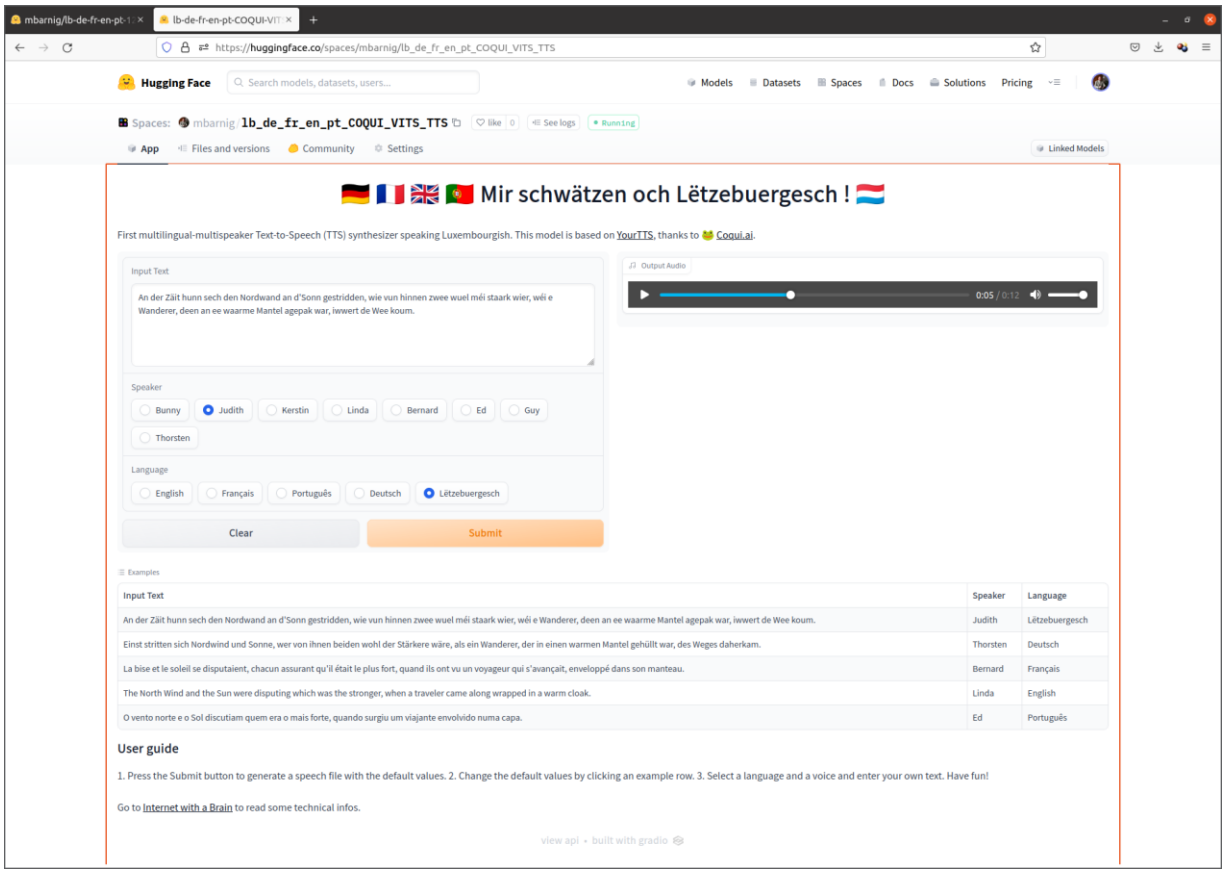
Une prolongation de l'apprentissage en continuant l'entraînement pendant plusieurs époques supplémentaires ne faisait pas de sens, car l'analyse du livre de bord montrait que l'entraînement du modèle TTS Tacotron2-DCA avec la base de données Marylux-648 entrait dans une phase de sur-apprentissage (overfitting), ce qui ne permettait plus de synthétiser correctement des mots ou phrases non vus lors de l'apprentissage. J'ai dû me rendre compte que la taille de la base de données Marylux-648 est insuffisante pour l'entraînement d'un modèle TTS Tacotron2-DCA. Il est vrai que Tacotron est connu pour être gourmand en données.

Dans une seconde étape, j'ai utilisé un autre modèle appelé VITS (Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech), le dernier né des familles TTS neuronaux que la communauté Coqui AI avait adapté et publié en source ouverte entretemps. Et là, le succès était au rendez-vous. La qualité de la voix synthétique « Judith » résultante était supérieure à la technologie par sélection d'unités de MaryLux, mais inférieure à la qualité d'une voix humaine. Pour progresser deux pistes restaient à explorer : augmenter la base de données d'enregistrements luxembourgeois et profiter d'un modèle VITS pré-entraîné avec des langues étrangères.

Avec l'accord de Peter Gilles, j'ai téléchargé les fichiers sonores avec transcriptions de dictées luxembourgeoises disponibles sur le site web de l'université du Luxembourg pour les cours de

formation en langue luxembourgeoise. Après un traitement de ces données avec les mêmes procédures décrites ci-avant pour la base de données MaryLux et avec l'accord des oratrices, j'ai réalisé un modèle de synthèse vocale VITS « multivoix » avec quatre voix féminines : Judith, Caroline, Nathalie et Sara. En utilisant le modèle VITS-VCTK de Coqui AI, pré-entraîné avec le corpus anglais VCTK pendant un million d'itérations et avec un large jeu de phonèmes IPA, y compris ceux que j'ai retenu pour mes modules de phonétisation luxembourgeoise Gruut et eSpeakNG, j'ai pu améliorer sensiblement la qualité de la synthèse vocale. La base de données de référence VCTK comprend environ 44.000 échantillons, prononcés par 46 orateurs et 63 oratrices.

La dernière étape consistait à ajouter des voix masculines et à développer une application de synthèse vocale multivoix et multilingue. J'ai pu profiter des prouesses technologiques réalisées par des membres de la communauté Coqui AI en relation avec le modèle TTS multilingue YourTTS. Avec l'autorisation aimable des dirigeants et responsables des contenus de RTL, Steve Schmit et Tom Weber, j'ai téléchargé des fichiers audio avec transcriptions à partir du site web rtl.lu pour étendre ma base de données d'échantillons d'entraînement TTS. « Same procedure as every year » était mon slogan pour convertir les données dans les formats requis par l'intelligence artificielle.



Application de démonstration d'un système de synthèse de la parole multivoix et multilingue sur HuggingFace

En juillet 2022, l'application était prête et j'ai créé un espace de démonstration sur la plateforme AI de collaboration et de partage HuggingFace sous le nom « Mir schwätzen och Lëtzebuergesch ». Le modèle TTS de démonstration supporte cinq langues : anglais, français, allemand, portugais et luxembourgeois. Pour chaque langue, la voix de synthèse peut être sélectionnée parmi huit voix de langues maternelles différentes qui gardent leurs accents typiques quand ils s'expriment dans une langue étrangère. Le projet a été bien apprécié par la communauté Coqui-TTS et par les dirigeants de Coqui AI.

Apprentissage profond et clonage de voix

Machine Pattern Playback fin 1940

L'apprentissage profond de machines est une technique passionnante. Le comportement des réseaux neuronaux à la base des architectures de systèmes de synthèse de la parole me rappelle parfois les réactions de mes cinq petits-enfants lorsqu'ils faisaient de nouvelles découvertes ou lors d'un nouvel apprentissage. J'ai également constaté que des anciennes techniques réapparaissent dans des nouveaux systèmes. Les spectrogrammes utilisés dans les modèles neuronaux TTS ont déjà été utilisés dans la machine Pattern Playback, développé par Franklin S. Cooper à la fin des années 1940.

Il me reste à signaler qu'un modèle de synthèse vocale comme Coqui TTS, bien entraîné avec une voix ou plusieurs voix, pour une langue ou plusieurs langues, permet le clonage (voice cloning) facile de la voix de n'importe quelle personne adulte. Un enregistrement vocal d'une durée de 15 à 30 secondes suffit pour créer une voix synthétique de cette personne avec une qualité difficile à distinguer de la voix originale. La production de voix synthétiques sur mesure est d'ailleurs un des créneaux commerciaux de la start-up Coqui AI. Avec de tels outils le « deep fake » est à la portée de tout le monde.

2.1.8. Outils de reconnaissance de la parole luxembourgeoise

En jargon technique, les outils de reconnaissance de la voix sont appelés STT (Speech-to-Text) ou ASR (Automatic Speech Recognition). Un vétéran parmi les outils ASR est le projet KALDI qui a démarré en 2009 à l'université Johns-Hopkins aux Etats-Unis. Aujourd'hui les géants du web comme Google, Microsoft, AWS, etc., proposent des services commerciaux de reconnaissance de la voix en ligne pour de nombreuses langues. Ils ne sont toutefois pas intéressés aux langues à faibles ressources comme le luxembourgeois. D'autres géants comme Meta (Facebook) et des start-up's comme OpenAI et Coqui AI ont développé des modèles à source ouverte qu'ils mettent à disposition de la communauté des chercheurs. L'état d'art actuel consiste à entraîner des larges réseaux neuronaux avec plusieurs langues et avec des millions de paramètres pour faire ensuite un réglage fin (fine-tuning) de ces modèles avec un entraînement supplémentaire pour une langue spécifique.

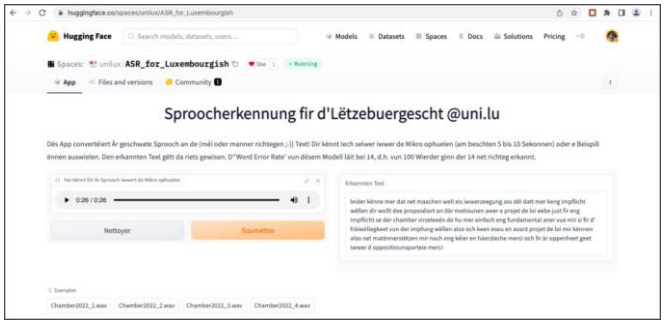
Modèle Wav2Vec-XLSR-53

Le modèle wav2vec, développé par le laboratoire de recherche de Facebook (meta.ai) au début 2019 est le plus ancien et le plus connu des modèles ASR. Rien que sur la plateforme AI HuggingFace on trouve 4.430 versions pré-entraînées de ce modèle. Le doyen de ces modèles au nom de Wav2Vec2-XLSR-53 a été pré-entraînée par Meta avec 56.000 heures d'enregistrements audio, sans transcriptions, en 53 langues, mais sans le luxembourgeois. Ce modèle a été utilisé par Le Minh Nguyen, un étudiant luxembourgeois en sciences des technologies vocales à l'université de Groningen, pour réaliser une application de reconnaissance vocale luxembourgeoise. En 2021, il a réussi son bachelor à l'université du Luxembourg avec la mention très bien. Dans le cadre du programme Erasmus il a passé un semestre à l'université technique de Vienne. En printemps 2022, Le Minh Nguyen a effectué un stage auprès du ZLS et en été 2022, il a obtenu son master à l'université de Groningen avec distinction cum laude.

Dans sa thèse de fin d'études « Improving luxembourgish speech recognition with cross lingual speech » Le Minh Nguyen présente les détails de son développement. Il a entraîné le modèle Wav2Vec2-XLSR-53 avec 842 heures d'enregistrements audio luxembourgeois sans transcriptions (unlabeled speech) et avec 14 heures d'enregistrements audio avec textes synchronisés (labeled

speech). Un modèle de langage 5-gram, entraîné avec des textes luxembourgeois totalisant 20 millions de mots, a été ajouté au décodeur vocal. Les enregistrements audio et les textes ont été fournis par la Chambre des Députés et par RTL. Actuellement, Le Minh Nguyen travaille à distance comme chercheur/programmeur pour la société Deepgram à San Francisco, spécialisée dans les plateformes vocales intelligentes.

Modèle Wav2Vec2-XLS-R

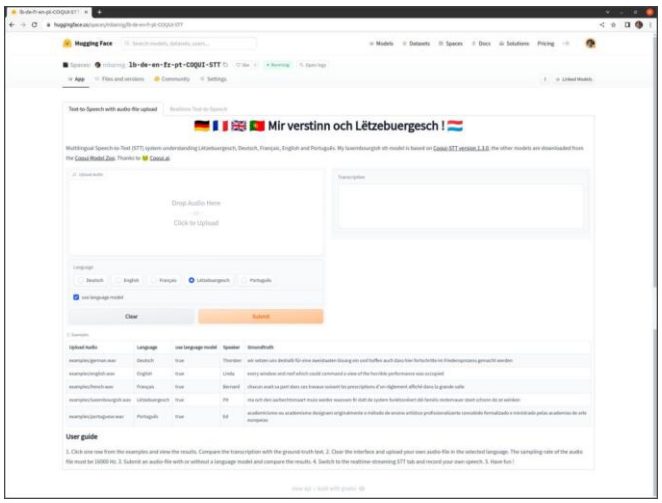


Premier modèle ASR luxembourgeois sur HuggingFace

Une version plus récente du modèle, XLS-R, pré-entraîné avec 436.000 heures d'enregistrements audio en 128 langues, incluant le luxembourgeois, a été présentée par Facebook (meta.ai) en décembre 2021. A partir du début 2022, Peter Gilles, professeur de langage et linguistique et directeur du département Humanités à la faculté des Sciences Humaines, des Sciences de l'Éducation et des Sciences Sociales (FHSE) de l'université du Luxembourg, a développé le premier prototype d'une application de reconnaissance vocale luxembourgeois basée sur ce modèle. En février 2022, il a présenté ce modèle dans un espace de démonstration sur la plateforme HuggingFace où il pouvait être testé par des initiés.

Cette application incluait également un modèle de langage n-gram, mais ne supportait pas encore l'usage des lettres majuscules pour les substantifs. Peter Gilles a présenté ses résultats en mai 2022 sur le site web Infolux de l'université du Luxembourg avec l'attribut « A very first model ». Ce modèle était un jalon dans le développement des technologies de la voix au Luxembourg.

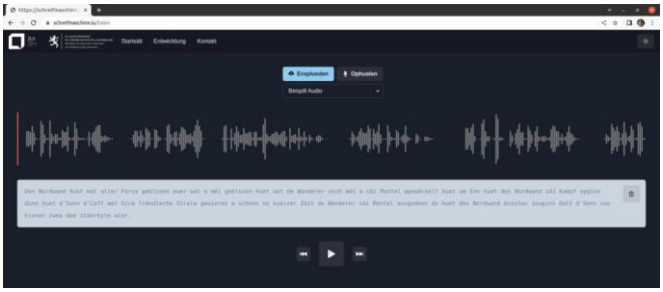
Coqui STT



Mon modèle ASR Coqui STT sur HuggingFace

La start-up Coqui AI n'est pas seulement actif dans le domaine de la synthèse vocale, mais également dans le domaine de la reconnaissance de la parole. Le projet afférent est désigné comme Coqui STT. Au printemps 2022, dans l'attente d'une correction de quelques bogues dans les modèles Coqui TTS, je m'étais focalisé temporairement sur la reconnaissance vocale du luxembourgeois moyennant le modèle Coqui STT. J'ai décrit en juin 2022 mes expériences dans mon récit « Mäi Computer versicht Lëtzebuergesch ze verstoen » sur mon site web et j'ai publié fin juillet 2022 un modèle STT multilingue « Mir verstinn och Lëtzebuergesch ! » sur la plateforme d'intelligence artificielle HuggingFace. Pour l'entraînement de ce modèle j'avais utilisé ma base de données multivoix de mon projet Coqui TTS. Il fallait seulement modifier le format des fichiers de transcription et supprimer les ponctuations dans le texte. Les performances du modèle Coqui STT sont toutefois largement inférieures aux modèles Wav2Vec et aux modèles présentés ci-après, ce qui m'a amené à abandonner le développement d'un modèle Coqui STT luxembourgeois.

schreifmaschinn.lu



Page web de reconnaissance de la parole schreifmaschinn.lu

À partir de l'automne 2022, Le Minh Nguyen a perfectionné son modèle avec l'appui de Sven Collette, linguiste informatique au ZLS. Après sa formation en neurosciences à Munich (Université Louis-et-Maximilien) et Lausanne (Ecole Polytechnique Fédérale), Sven Colette a obtenu un doctorat de philosophie en neurosciences à l'université Pierre et Marie Curie à Paris. Il a travaillé de 2012 à 2016 comme chercheur postdoctoral à l'institut de technologie de Californie (CALTECH) à Pasadena. Dans la suite il a été consultant et chercheur dans le domaine de l'intelligence artificielle, avant de joindre le ZLS au début de 2021.

Le duo de développement a remplacé le modèle original par la version Wav2Vec2-XLS-R et pour l'entraînement du luxembourgeois, la durée des enregistrements audio avec transcriptions a été étendue à seize heures. Un site web schreifmaschinn.lu a été développé pour rendre l'application de reconnaissance de la parole luxembourgeoise accessible à tout le monde. Les visiteurs peuvent dicter des textes qui sont affichés en temps réel sur l'écran du navigateur. Il est également possible de télécharger des fichiers audio. Le projet schreifmaschinn.lu constitue une vraie prouesse technique.

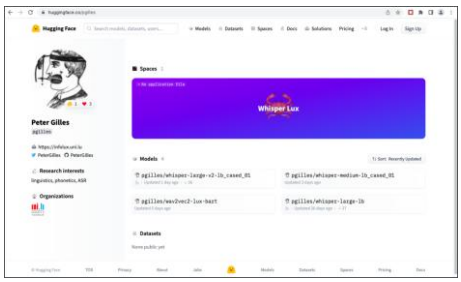
Sven Colette, Claude Meisch, Le Ming Nguyen,Luc Marteling (de gauche à droite)

L'application a été inaugurée le vendredi 9 décembre 2022 par le ZLS dans la salle de conférence à Clausen, en présence de Claude Meisch, ministre de l'Éducation Nationale, de l'Enfance et de la Jeunesse. Luc Marteling, directeur du ZLS, a souhaité la bienvenue aux nombreux invités. Le projet schreifmaschinn.lu a été présenté en longueur et en largeur dans tous les média du Luxembourg : quotidiens, hebdomadaires, magazins mensuels, radio et télévison. [6]

En fin d'année 2022, j'ai testé l'application schreifmaschinn.lu minutieusement d'une façon reproductible avec trois jeux d'enregistrements pour mesurer la qualité de reconnaissance des mots dictés. J'ai publié les résultats dans le rapport « Fënnef Stären fir d'Applikatioun schreifmaschinn.lu » publié sur mon blog.

Modèle Whisper

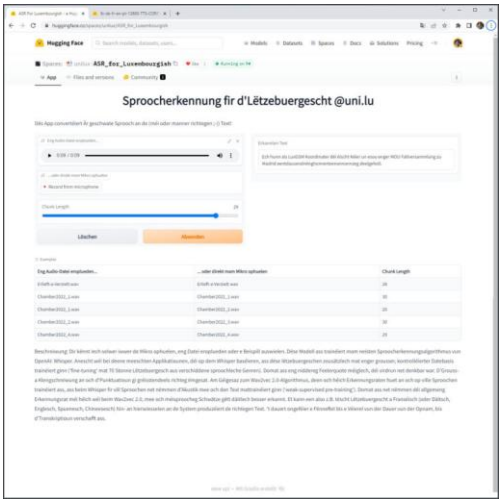
Le modèle de reconnaissance vocale pré-entraîné le plus récent s'appelle Whisper. Il a été présenté par la société OpenAI le 21 septembre 2022 et amélioré en décembre 2022 (version 2). OpenAI a été fondée en décembre 2015 comme organisation à but non lucratif avec l'objectif de promouvoir et de développer une intelligence artificielle à visage humain qui bénéficiera à toute l'humanité. Parmi les fondateurs, se trouve Elon Musk, entrepreneur et milliardaire emblématique. L'organisation a été convertie en entreprise à but lucratif plafonné en mars 2019. OpenAI est surtout connue pour ses applications récentes Dall-E2 et ChatGPT qui font ravage sur le web.



Premier espace Whisper de Peter Gilles sur HuggingFace

Whisper a été entraîné avec 680,000 heures d'enregistrements audio avec transcriptions, dont 83 % en anglais et 17 % avec 96 langues différentes. Lors de la présentation du projet schreifmaschinn.lu Peter Gilles m'avait informé qu'il s'est lancé dans l'adaptation de Whisper au Luxembourgeois. Peu après j'avais découvert les premières traces afférentes sur la plateforme HuggingFace.

L'application est maintenant opérationnelle sur HuggingFace et tourne sur une infrastructure GPU T4. À première vue, les performances de ce modèle sont géniales. Au rythme effréné actuel de l'évolution de l'intelligence artificielle, on va bientôt savoir si Whisper a le potentiel de détrôner le modèle Wav2Vec-XLS-R dans l'application schreifmaschinn.lu. De toute façon, ce sont les données luxembourgeoises audio, avec les transcriptions soigneusement corrigées et assemblées, ainsi que l'expérience gagnée jusqu'à présent, qui constituent la vraie valeur d'un projet de reconnaissance vocale, et non l'architecture du réseau neuronal à la base.



Application ASR Whisper uni.lu sur HuggingFace

2.1.9. Moyens de promotion du luxembourgeois

Le meilleur moyen de promotion de la langue luxembourgeoise consiste à l'employer dans la vie journalière et de donner un bon exemple. La croissance de l'utilisation de notre langue nationale dans la littérature, les médias, le théâtre, les films, les sites web, témoigne de l'intérêt grandissant des résidents au Grand-Duché pour le luxembourgeois. Je déplore de ne pas pouvoir relever ici tous les noms des auteurs, journalistes, régisseurs, producteurs et autres créateurs de contenus luxembourgeois, mais je dois me focaliser sur l'histoire des TIC dans ce livre. Je vais donc présenter ci-après quelques projets concernant l'utilisation de la langue luxembourgeoise qui sont plus apparentés aux nouvelles technologies.

Wikipedia Lëtzebuerg

J'estime qu'il n'existe pas d'internaute luxembourgeois qui ne connaît pas Wikipédia. Cette encyclopédie universelle et multilingue, créée par Jimmy Wales et Larry Sanger le 15 janvier 2001, a acquis une influence mondiale et figure parmi les dix sites web les plus visités dans le monde.

Wikipédia est une œuvre libre, c'est-à-dire chacun peut la diffuser à sa guise. Elle existe en plus de 300 langues dans une apparence unie. Pour financer l'infrastructure technique de Wikipédia, une fondation Wikimedia a été créée le 20 juin 2003. Le fonctionnement de Wikipédia repose sur le logiciel MediaWiki du wiki, une application du web qui permet la création, la modification, la gestion et l'illustration collaboratives de pages à l'intérieur d'un site web. Le premier wiki a été programmé en 1995 par Ward Cunningham et s'appelait WikiWikiWeb.

Page web d'accueil Wikipédia Lëtzebuerg

Le contenu de Wikipédia est généré et géré par les usagers. Tout lecteur de Wikipédia est un rédacteur ou correcteur potententiel. Un rédacteur est identifié par son adresse IP s'il est anonyme ou par son pseudonyme s'il s'est enregistré sur le site Wikipédia. Certains rédacteurs ont des privilèges dans l'utilisation du logiciel MediaWiki qui leur sont conférés en général par une entité linguistique locale. On distingue entre administrateurs, bureaucrates et arbitres. Ils peuvent se faire assister par des bots, des agents automatiques constitués par des programmes informatiques qui effectuent des tâches répétitives et fastidieuses pour un humain, par exemple la vérification d'hyperliens, la correction de fautes orthographiques ou le blocage de vandales.

Au Luxembourg, une association sans but lucratif Wikimedia Lëtzebuerg (WM-LU ; F11166) a été immatriculée le 26 novembre 2016 par neuf bénévoles. Les membres-fondateurs sont, par ordre alphabétique, René Beideler, Tom Diderich, Marc Espen, Jean-Louis Gindt, Claude Meisch, Gilles Peters, Robert Scheueren, Sandra Souza Morais, Joseph von Graes. Le siège social a été transféré en 2019 à Leudelange et l'association a été présentée dans le magazine « Gemengebuet Leideleng No 133 » en juillet 2019.

Mais les pionniers du Wikipédia luxembourgeois étaient bien actifs avant la création de leur association. La première copie de la page d'accueil « Wikipedia op lëtzebuergesch », captée par la Wayback Machine, date du 30 juillet 2004. En octobre 2004, on comptait déjà 1.370 articles publiés et il y avait trois administrateurs : Zinneke, Cornischong et Maradong. Les autres contributeurs étaient Briséis, Jim Hawk, Johnny Chicago, Karlethegreat, Kwisatz, RenderWandler et Thorben.

Parmi les nouveaux rédacteurs relevés en 2005 et en 2006 figurent « Les Meloures, Otets et Robby » qui sont les administrateurs actuels, à côté de l'administrateur de la première heure, Zinneke.

La société EducDesign avait enregistré le nom de domaine wikipedia.lu en 2004 auprès de RESTENA pour éviter qu'il tombe dans les mains de spéculateurs. Aujourd'hui, ce domaine est redirigé vers l'URL officiel « lb.wikipedia.org » de Wikipedia Lëtzebuerg.

Emission PISA sur RTL au sujet de Wikipedia en 2012

Les contributeurs à Wikipédia Luxembourg sont discrets, ce qui explique qu'ils sont rarement sur la une des journaux. On trouve un épisode sur RTL en décembre 2012 dans l'émission « PISA-de Wëssensmagazin », quelques mentions dans les média pour la publication du 50.000^{e} article en septembre 2017, un article « Den Här Wikipedia » alias René Beidweiler en avril 2019 dans le Tageblatt et un portrait de Jean-Louis Gindt intitulé « Wëssen fir jiddwereen » em février 2021 dans la Revue.

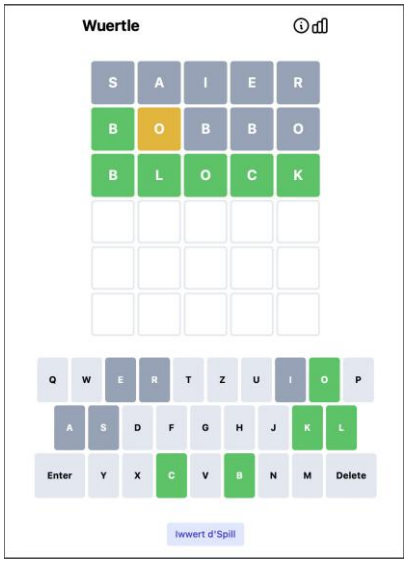
Avant la constitution de l'a.s.b.l., les membres ont déjà signé en décembre 2014 la charte du bénévolat et se sont enregistrés auprès de l'agence du bénévolat. L'« Aktioun Lëtzebuergesch » a décerné en 2021 la plaquette en argent Dicks-Rodange-Lentz à l'association Wikimedia Lëtzebuerg pour la promotion de la langue luxembourgeoise.

Comité de l'a.s.b.l. Wikimedia Lëtzebuerg en 2023

Un dernier fait curieux reste à signaler. Un cliché fort répandu sur le web dit que Wikipédia est dominé par des « hommes blancs âgés » et que les femmes sont minoritaires. Un regard sur la photo à gauche prise lors de l'assemblée générale en février 2023 montre que la situation n'est pas différente au Luxembourg. Cette remarque n'est pas péjorative, au contraire. Je fais moi-même partie de cette race.

À la fin de l'année 2023, le nombre d'articles luxembourgeois sur Wikipedia à dépassé 65.000.

Wuertle



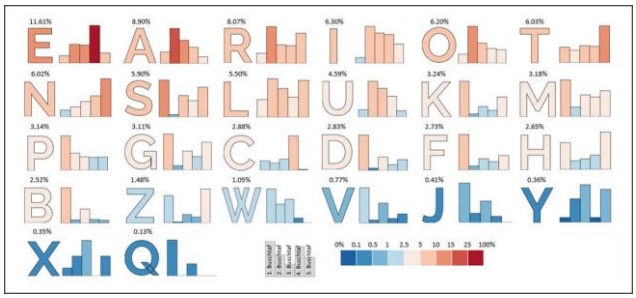
Application LOD Wuertle

L'informaticien Sven Clement, député du parti politique pirate (PPLU) dont il est le président et cofondateur, a publié le 21 janvier 2022 une petite application sur le web au nom de Wuertle, qui se réfère au jeu Wordle. Le défi consiste à deviner un mot de 5 lettres extrait chaque jour du LOD. Si on entre des lettres dans la première ligne du tableau Wuertle, les lettres incluses dans le mot cherché sont colorées en orange et les lettres incluses qui se trouvent à la position correcte sont affichées en vert. Avec ces indices, on peut entrer un nouveau mot dans la deuxième ligne. Le but est de trouver le mot correct avec un minimum d'essais. Ensuite, on peut partager ses exploits sur les réseaux sociaux et comparer ses performances avec celles d'autres joueurs.

Sven Collette a calculé la probabilité d'inclusion de chaque lettre de l'alphabet dans les mots luxembourgeois à 5 caractères qui figurent dans le LOD. Les résultats de cette contribution « E bësse Wuppes fir Wuertle » sur le site web du ZLS ont été repris dans un article « Tipps fir den Wuertle » dans la Revue No 9/2022. J'ai suivi les conseils de Sven Collette et soumis le mot « SAIER » dans la première ligne du jeu. Aucune des cinq lettres

n’est contenue dans le mot du jour, malgré la probabilité de 96 % calculé d’avoir au moins une lettre incluse.

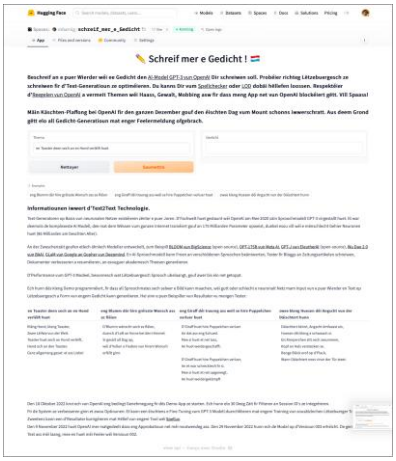
Comme deuxième mot, j’ai entré « BOBBO », un terme à éviter sur base des statistiques. Sans aucune chance théorique, j’ai marqué deux coups. Parfois, il faut se méfier des statistiques. Je me rappelle souvent l’exemple d’un film destiné aux enfants qui a manqué sa cible. Comme les grands-parents sont allés au cinéma avec leurs petits enfants, l’âge moyen des visiteurs a été évaluée à 40 ans.



E bësse Wuppes fir Wuertle moyennant des statistiques

Une semaine après la mise en ligne du jeu Wuertle, Peter Gilles a réalisé une version avec des phonèmes au lieu de caractères. Cette variante ajoute une complexité supplémentaire pour trouver le mot du jour.

Schreif mer e Gedicht



Schreif mer e Gedicht

En été 2022, j'ai procédé à mes premières expériences avec le modèle « NLP GPT-3 » d'OpenAI. C'était quelques mois avant le début de l'euphorie au sujet du projet ChatGPT qui est basé sur une version améliorée de GPT-3. J'avais découvert que le modèle « GPT-3 » comprend la langue luxembourgeoise et qu'il peut écrire des textes luxembourgeois. Je me suis amusé à demander à ce programme d'intelligence artificielle d'écrire des poèmes luxembourgeois sur des sujets spécifiques. Mon thème favori était un grille-pain qui tombe amoureux d'un chien.

Malgré quelques fautes orthographiques dans les textes produits, j'ai été impressionné, voir effrayé, par les résultats. Je me suis demandé si Blake Lemoine, qui est persuadé que le chatbot de Google qu'il devait évaluer est conscient, n'a pas eu raison ? Cet ingénieur a été licencié non pas à cause de sa croyance, mais parce qu'il a partagé des informations confidentielles de Google sur les réseaux sociaux.

J'ai programmé une application de démonstration « Schreif mer e Gedicht » sur la plateforme AI HuggingFace. L'utilisation de l'API GPT-3 (text-davinci-003) d'OpenAI est facturée au volume. À l'époque, il fallait faire approuver toute application GPT-3 par un ingénieur d'OpenAI, ce qui n'était pas chose facile. J'ai obtenu une approbation provisoire avec obligation d'ajouter une identification des utilisateurs dans le programme. Avant l'expiration de mon délai de carence, la société OpenAI a levé toutes les restrictions et mon projet a été validé pour usage anonyme. Un autre problème surgissait alors. Je m'attendais à une utilisation limitée par quelques intéressés à la langue luxembourgeoise, mais de nombreux étrangers ont rapidement compris que mon application se prête également pour la production des textes dans d'autres langues. De cette manière, le seuil mensuel de facturation que j'ai fixé sur mon compte OpenAI est dépassé les premiers jours du mois et la démo est bloquée pendant le reste du mois. Je n'ai pas encore une solution simple pour écarter les intrus.

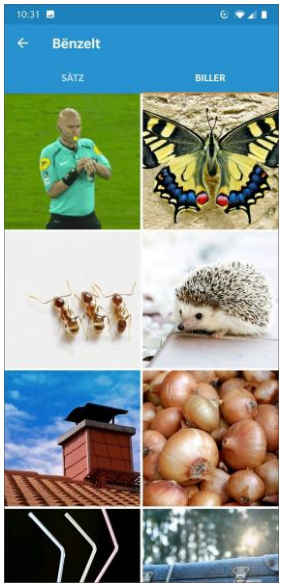
RTL Kultur Rubrik 11.2022

Il reste à signaler que RTL a présenté le projet en novembre 2022 sur son site web dans la rubrique Kultur News.

2.1.10. Projets de recherche linguistique

Il y a quelques années, l'université du Luxembourg a commencé à documenter la langue luxembourgeoise dans toute sa richesse moyennant des technologies d'analyse et d'infographie les plus modernes. Quelques projets représentatifs sont présentés ci-après.

Schnëssen pour la science



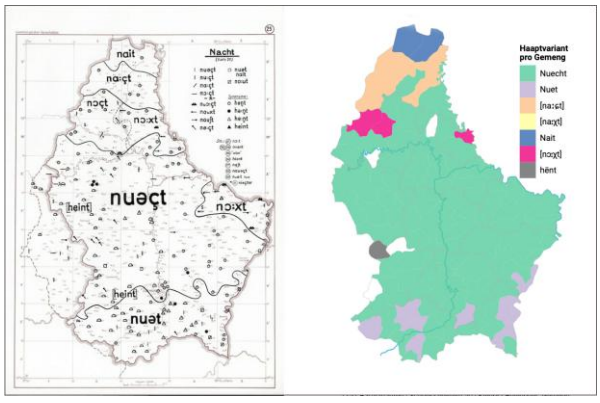
App Schnëssen

Le projet de recherche linguistique concernant les variantes locales du luxembourgeois dans le pays le plus connu est certainement l'application « Schnëssen – Är Sprooch fir d'Fuerschung ». Démarré en mars 2018, ce projet de « crowdsourcing » consistait à collecter des enregistrements vocaux moyennant une app développée pour smartphones iOS et Android. L'usager indiquait d'abord sa localisation géographique et son âge et enregistrait ensuite ses réponses à des questions multiples, par exemple le nom d'un objet sur une image ou la traduction d'un terme allemand ou français en luxembourgeois. Un exemple concret est le papillon pour lequel des expressions Pimpampel, Millermoler ou Päiperlék ont été enregistrées.

Environ 3.500 participants ont fourni plus que 250.000 contributions lors des 5 campagnes de promotion de l'application « Schnëssen ». L'analyse et l'interprétation des enregistrements ont été effectuées par les doctorantes Nathalie Entringer et Sara Martin, sous la direction de Peter Gilles et avec la collaboration de Christophe Purschke. Le lecteur intéressé trouve plus de détails dans la rubrique « Schnëssen-App » sur le portail infolux.lu de l'université.

Variatiounsatlas

Les résultats du projet « Schnëssen » ont servi à dresser des cartes géographiques du pays qui montrent la répartition des variantes d'expressions luxembourgeoises par région. Il s'agit d'un projet en cours qui comprend actuellement 725 cartes qui peuvent être sélectionnées et affichées sur la page web « Infolux / Variatiounsatlas ». Pour 173 termes, on peut comparer la situation actuelle avec celle documentée dans l'ancien « Luxemburgischer Sprachatlas » publié en 1963 pour ainsi découvrir l'évolution de la langue luxembourgeoise au courant des 60 dernières années.



Variatiounsatlas sur infolux.uni.lu

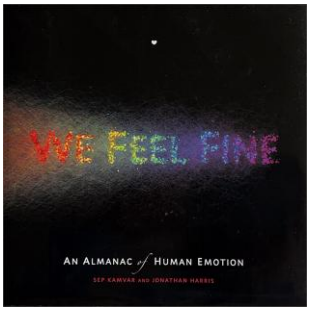
L'auteur de ce premier atlas de la langue luxembourgeoise est Robert Bruch, un linguiste luxembourgeois qui est surtout connu à l'étranger. Il est décédé à l'âge de 39 ans des suites d'un accident de voiture en 1959. Son dernier œuvre a été publié posthume en 1963 par le germaniste allemand Ludwig Erich Schmitt. Une version digitale du « Luxemburgischer Sprachatlas (LuxSA) a été créée en 2003 par Claudine Moulin, professeur de linguistique historique à l'université de Trèves. Elle est membre du conseil de gouvernance de l'université du Luxembourg depuis 2018.

Le nouvel atlas des variations locales de la langue luxembourgeoise est réalisé par l'équipe de Peter Gilles. La création des cartes est effectuée avec l'outil libre à source ouverte « shinydashboard ».

2.1.11. Reconnaissance automatique de sentiments

L'analyse de sentiments et d'attitudes présentes dans un texte est une nouvelle technique développée depuis les années 2000. Elle est utilisée par les sociologues pour sonder l'opinion publique (opinion mining) sur un sujet et pour caractériser les relations sociales sur le web. Elle est également employée en marketing pour déterminer comment un public cible accueille et perçoit une marque. L'analyse de sentiments demande une très bonne compréhension d'une langue, car il ne suffit pas de faire des statistiques sur la fréquence d'apparition de quelques mots liés aux sentiments, il faut les mettre dans le bon contexte.

We feel fine



Livre We feel fine

« We Feel Fine » est une œuvre visant l'exploration des sentiments de l'humanité en ligne. Elle comprend un logiciel qui collectait systématiquement pendant plusieurs années sur des blogues, toutes les quelques minutes, les phrases contenant le syntagme « I feel ». Les données recueillies sont affichées moyennant une application interactive sur le web en six mouvements : Madness, Murmurs, Montage, Mobs, Metrics et Mounds. Dans tous les cas, le visiteur peut sélectionner, selon divers critères, des phrases qui traitent des sentiments de la foule anonyme des blogueurs. Cela permet notamment de fournir une réponse à des questions du type : quelle est la ville la plus triste du monde ? Quelle est l'incidence du climat sur l'humeur des gens ? Y a-t-il une période de l'année où l'humanité est plus heureuse ? Comment se sentaient hier les habitants du Luxembourg ? Quelle est l'humeur générale des habitants de la planète aujourd'hui ? L'oeuvre a reçu de nombreux prix à travers le monde et elle est documentée dans un livre au même nom. L'application dispose d'une interface informatique (API) qui permet d'intégrer les données recueillies dans ses propres développements.

Le projet « We feel fine » a été développé à partir de 2005 par Jonathan Jennings Harris et Sep Kamvar. Jonathan Harris est un artiste multimédia de renommé mondiale. Il est diplômé en sciences informatiques de l'université de Princeton aux Etats-Unis et en conception interactive de La Fabrica en Italie. Il se considère comme un antrophologue du web et exprime son empathie pour les humains régulièrement par de nouveaux projets multimédia. Son talent a été récompensé par de nombreux prix décernés par des institutions prestigieuses. Sep Kamvar a étudié aux universités de Princeton et Stanford. Il était professeur des arts et sciences multimédia au MIT jusqu'en 2016. Dans la suite, il s'est entièrement consacré à la fondation et gestion de start-ups.

Sentiments à la Luxembourgeoise

J'ai assisté en novembre 2021 à la conférence « 15 Joer Luxemburgistik » qui a eu lieu dans le cadre du 15^{e} anniversaire de l'institut d'études linguistiques et littéraires luxembourgeoises du département des sciences humaines à l'université du Luxembourg. L'équipe de Peter Gilles présentait l'historique et les perspectives des projets de recherche dans le domaine de la langue luxembourgeoise. À cause des restrictions imposées par la pandémie COVID, les présentations et discussions se passaient surtout en ligne moyennant l'outil de vidéoconférence WebEx commercialisé par Cisco.

C'est surtout le dernier exposé avec le titre « Der Einfluss von liguistischen Merkmalen auf die maschinelle Erkennung von Sentiment » qui a retenu toute mon attention. La doctorante Daniela Gierschek a présenté les résultats de son projet de recherche concernant l'analyse de sentiments dans les commentaires soumis par les internautes sur le site web rtl.lu. Pour soutenir sa thèse, elle a pu se baser sur l'outil STRIPS développé au même institut. Les archives mises à disposition par RTL constituaient un atout considérable, car il est rare que des chercheurs aient accès à des données réelles d'une telle richesse pour entraîner un modèle d'intelligence artificielle.

TalkWalker

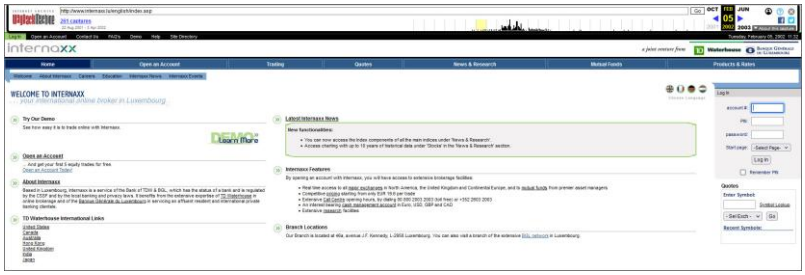
Présentation de la start-up Trendiction par Thibaut Britz

« Commercialiser un outil de recherche complémentaire à Google Search et Google Alerts pour permettre à des entreprises de visualiser l'opinion des internautes sur leurs produits et marques » était le rêve de Thibaut Britz, après l'obtention de son diplôme d'ingénieur en sciences informatiques à École Poytechnique Fédérale de Zurich en 2007. Pendant ses études, il avait déjà réalisé un moteur de recherche « blue.lu » pour indexer des sites web luxembourgeois. Le robot de recherche (crawler) y associé portait le nom de « Confuzzledbot ». Sur la Wayback-Machine, on trouve des traces de ces anciens projets.

En 2008, Thibaut Britz s'est mis ensemble avec son ami de longue date Christophe Folschette qui a obtenu son diplôme en informatique et économie à l'université technique de Munich en 2006. Christophe Folschette avait déjà quelques expériences avec la création de start-ups et avec des activités de consultance en Allemagne.

Le duo a été admis sous statut pré-commercial au Technoport d'Esch-sur-Alzette, l'incubateur du CRPHT, dirigé par Diégo de Biaso. Ils ont bénéficié d'un support pour finaliser leur plan d'affaires (business plan) et développer un premier prototype de leur outil de recherche et d'analyse d'attitudes sur le web. En mars 2009, ils ont créé la start-up Trendiction s.à r.l. avec un capital de 25.000 €. Thibaut Britz détenait trois quarts des parts sociales. La même année, la société a été admise définitivement au Technoport et a pu signer un contrat avec son premier client.

Deux ans après le démarrage des activités de développement de l'outil de collecte des données, appelé dans la suite « les yeux et les oreilles des plus grandes marques » par le magazine ITnation, les fondateurs de Trendiction avaient construit toute l'infrastructure des serveurs de collecte de données sur fonds propres, sans se verser le moindre salaire. Mais ils continuaient à croire dans leur idée et ils avaient la chance en 2010 de faire la connaissance de Robert Glaesener qui cherchait à l'époque un nouveau défi.



Page web d'accueil Internaxx sur la Wayback Machine

Diplômé de la HEC Paris et de la Harvard Business School, Robert Glaesener a fondé en 2000 la banque en ligne luxembourgeoise Internaxx, un supermarché numérique des fonds d'investissements et un service international de courtage en ligne. Après l'acquisition de 100% du capital d'Internaxx par le groupe TD Waterhouse en 2010, Robert Glaesener a quitté son ancienne pépite et décida d'investir dans la start-up Trendiction. Par ses études, ses expériences et son caractère, il était complémentaire aux deux fondateurs et il a pris la fonction de CEO en qualité de nouvel actionnaire principal. Thibaut Britz et Christophe Foschette assuraient les fonctions de directeur technique et de directeur commercial et ont pu gagner leur premier salaire. Fin 2010, la start-up était passée de 3 à 6 personnes.

En 2012, Thibaut Britz a décroché le sixième prix « Creative young entrepreneur Luxembourg (Cyel) », organisé par la Jeune chambre économique du Luxembourg.

Page web d'accueil Talkwalker en 2023

Grâce à l'ambition et le leadership du CEO et à la compétence technique des fondateurs, la start-up est devenue rentable et s'est transformée rapidement en PME, au nouveau nom de TalkWalker, avec 45 employés, puis en entreprise avec 240 personnes et finalement en groupe international avec une équipe de plus de 600 salariés, reconnue comme spécialiste mondiale de l'intelligence des données sociales.

En 2014, Talkwalker est devenu le troisième outil de ce type en Europe et le sixième dans le monde à faire partie du « Twitter Certified Products Program ». Toute cette croissance a été accompagnée par le perfectionnement et la diversification des produits et services proposés aux clients, de levées de fonds, de déménagements successifs dans des locaux plus spacieux, de présences dans d'autres pays, de partenariats, d'acquisitions de sociétés, de ventes de parts sociales, de réceptions de prix et récompenses et de changements de dirigeants. Pendant dix années, TalkWalker était synonyme de « success story luxembourgeoise ». Lors de l'inauguration du premier bureau aux États-Unis en 2015, le Premier ministre Xavier Bettel s'est même déplacé pour couper le ruban dans les locaux à New York, en marge de son passage aux Nations Unies. Parmi les centaines de clients, on citait des noms prestigieux comme Bennetton, Bonduelle, Crédit Agricole, Coca-Cola, Microsoft, Ogily, Orange, Publicis, Yves Rocher.

Mais le succès d'une entreprise est rarement éternel. À partir de 2016 TalkWalker a accumulé des pertes à tendance progressive. En été 2021, l'Américain Tod Nielsen prenait les rênes de l'entreprise et Robert Glaesener passa comme président au conseil d'administration. Dans un entretien avec le magazine Paperjam l'ex-CEO déclarait « le troisième étage de la fusée TalkWalker est prêt ». Le nouveau CEO procédait à de nombreux licenciements et une partie importante du personnel quittait volontairement la société. En 2022, le nombre de salariés avait reculé de 36 %. A peine une année après sa nomination, Tod Nielsen fut remplacé à son tour par Lookdeep Singh. Né en Inde et de formation technique, il vit au Luxembourg depuis 13 ans et a acquis la nationalité luxembourgeoise. Il a pour mission de restructurer les coûts, de revenir à la rentabilité et de se recentrer sur la technique. Ce processus s'est traduit par des suppressions supplémentaires de postes et la fermeture de bureaux à l'étranger pour assurer la durabilité de l'entreprise.

Thibaut Britz et Robert Glaesent ont quitté le navire pour se vouer à de nouvelles aventures, tandis que Christophe Folschette est resté à bord comme administrateur-gérant. En juin 2023, il présente la technologie TalkWalker au DMWF (Digital Marketing World Forum) à Londres. Thibaut Britz a fondé en février 2021 une nouvelle start-up Sedai s.à r.l. pour développer et exploiter des logiciels de trading automatisé.

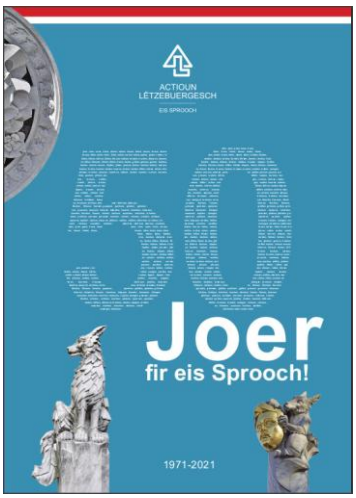
2.1.12. Apprentissage du luxembourgeois

L'apprentissage du luxembourgeois a pris de l'ampleur ces dernières années. L'offre et les méthodes d'apprentissage de la langue nationale pour les adultes se sont diversifiées. Les cours de langues classiques en présentiel sont toujours bien appréciés, mais les formations en ligne , accessibles aux quatre coins du monde, se sont multipliées. Avant de passer en revue la situation actuelle, il convient de jeter un regard en arrière.

E Liewe fir d'Sprooch

Pierre Reding, Alain Atten, Claude Meisch, Lex Roth, Marc Barthelemy (de gauche à droite) ©ZLS/Sophie Margue

« E Liewe fir d'Sprooch » est le titre d'une émission réalisée par l'animateur Jean-Claude Majerus au sujet de Lex Roth sur la radio 100 Komma 7 en décembre 2019. Mais ce titre s'applique également à un autre pionnier de la promotion de la langue luxembourgeoise, à savoir Alain Atten. Les deux promoteurs du luxembourgeois ont été récompensés pour l'ensemble de leurs œuvres par le prix national de mérite pour la langue nationale remis par Claude Meisch, ministre de l'Éducation nationale, lors d'une cérémonie officielle qui a eu lieu dans la salle des fêtes du lycée des garçons au Limpertsberg le 22 février 2023. Ce prix a été introduit par le règlement grand-ducal du 4 août 2022. Il est décerné annuellement à partir de l'année 2022. Comme le jury chargé de la sélection du lauréat n'a pas pu départager les mérites des deux candidats, le prix pour l'année 2022 a été attribué à l'aîné des deux pionniers, à savoir Lex Roth. Alain Atten a fêté son 85e anniversaire le jour de la remise du prix pour l'année 2023. A côté du prix, symbolisé par une sculpture d'un écureuil (Kaweechelchen), réalisée en impression 3D par des élèves du lycée des arts et métiers à Luxembourg, Alain Atten a eu droit a un gâteau d'anniversaire et à un « Gebuertsdagslidd » chanté par le groupe « Cojellico's Jangen ».



Anniversaire Actioun Lëtzebuerg

Lors de sa laudation sur les lauréats, Pierre Reding a souligné que Lex Roth et Alain Atten exerçaient il y a 60 ans les fonctions cumulées de commissaire, de conseil permanent, de centre de langues et d'institut de formation.

Lex Roth a exercé plusieurs fonctions au courant de sa carrière : enseignant, professeur attaché au ministère de la Culture, conseiller de gouvernement au ministère d'état, directeur du SIP (Service Information et Presse de l'Ètat) et membre de l'institut grand-ducal, section linguistique. Il est retraité depuis 1993.

En 1971, il a fondé l'association « Actioun Lëtzebuergesch » qu'il a présidé jusqu'en 1998 pour occuper dans la suite le poste de vice-président. Lors du 50e anniversaire de l'association en 2021 une publication « 50 Joer fir eis Sprooch » a été éditée par son comité.

Après Henri Rinnen en 1989, Lex Roth est l'unique autre enseignant du luxembourgeois qui a reçu une « plaquette Dicks-Rodange-Lentz » décernée par l'« Actioun Lëtzebuergesch ». C'était en 2003.

À partir de 1972 Lex Roth a animé des centaines d'émissions radio et TV et écrit encore plus d'articles dans la presse luxembourgeoise pour populariser le Luxembourgeois. Les rubriques les plus connues sont « Een Ament fir eis Sprooch », « Eis Sprooch mam Lex Roth », « Ronderëm eis Sprooch », « E Këppche fir eis Sprooch » et « Eng Klack fir eis Sprooch ». Il est en outre l'auteur de différents livres et de matériel didactique pour les écoles. Les résidents le connaissent comme « Monni Lex ».


KLACK SPROOCH

Eis Sprooch léieren

Den ‚Unzéckert'?

Eiser Lidder

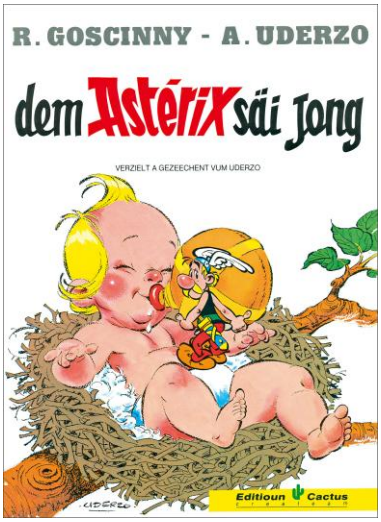
Eis Sprooch an d'Wahlen

Net léif!

Eng Klack fir eis Sprooch

Mais sur le plan littéraire, Lex Roth est surtout connu comme père de « Tutebattix », la version luxembourgeoise du barde radoteur qui fait partie des personnages de la série des albums Astérix et Obélix. La traduction a été effectuée par Lex Roth. On trouve les noms luxembourgeois de tous les acteurs de ces bandes dessinées sur le site web Wikipédia Lëtzebuerg.

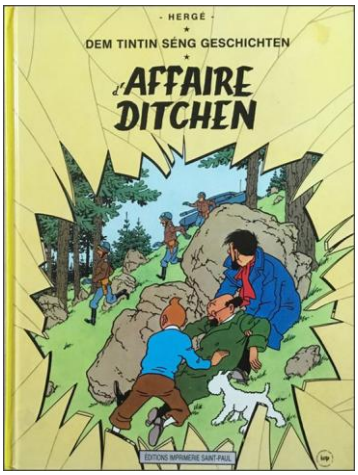
Lors de la présentation du premier album luxembourgeois « Dem Astérix säi Jong » à l'occasion du 20e anniversaire du groupe Cactus en mars 1987, Lex Roth a rencontré le dessinateur historique de la série, Albert Uderzo, qui participait aux festivités. Une amitié est née entre les deux hommes qui dure jusqu'à présent. L'idée de la traduction d'Astérix en luxembourgeois émanait du service marketing et publicité de Cactus (Createam), dirigé à l'époque par Jean Strock.



Astérix vum Lex Roth

Comme le succès du premier album d'Astérix en luxembourgeois n'était pas passé inaperçu auprès des éditeurs nationaux, Lex Roth et Gaston Zangerlé, responsable des éditions Saint-Paul, se sont mis d'accord pour traduire et publier une version luxembourgeoise de l'album « L'affaire Tournesol » de Tintin.

Cette première collaboration pour l'édition « d'Affaire Ditchen » était le début d'une production de dix aventures de Tintin et de huit autres livres d'Astérix en version luxembourgeoise. Le dernier album de Tintin « De bloe Lotus » a été publié en 1994, suivi par le dernier album « Den Astérix bei den Helveten » en 1996.



Tintin vum Lex Roth

En ce qui concerne Alain Atten, il est né en 1938 et il est donc cinq ans plus jeune que Lex Roth. Pendant sa carrière professionnelle, il était historien, linguiste, archiviste, auteur et animateur d'émissions radiophoniques. À partir des années 1960, Alain Atten a créé plusieurs mille contributions en rapport avec la langue luxembourgeoise : livres, cahiers locaux, articles de presse, histoires, récits, études, émissions TV et radio.

Sproochmates Alain Atten

Ses œuvres littéraires couvrent tous les genres : prose, lyrique, pièces de théâtre, scénarios de films. Comme Lex Roth, il a reçu de nombreux prix et récompenses pour ses activités de support de la langue luxembourgeoise et comme son ainé, il est menbre du conseil permanent de la langue luxembourgeoise (CPLL).

Pour la majorité des résidents, Alain Atten est toutefois le synonyme de « Sproochmates ».

« Sproochmates » est d'abord le nom d'une émission régulière sur RTL Radio qui a débuté à la fin des années 1970 et dont le nombre de diffusions a dépassé le seuil de 5.000 en 2015. C'était l'idée de Fernand Mathes qui a commencé en 1978 comme animateur radio auprès de RTL. Il a caractérisé Alain Atten comme dictionnaire luxembourgeois sur deux jambes.

« Sproochmates » est également le nom de deux livres avec CD's, publiés par la maison d'édition luxembourgeoise Schortgen en 2010 et 2013. Les illustrations de ces livres ont été dessinées par Roger Leiner.

Cours de langues

Les organisateurs de cours de langue luxembourgeoise sont nombreux et variés. Au niveau public, l'acteur le plus ancien est le centre de langues (CLL) qui a démarré ses activités en 1991. À la tête de cet organisme, se trouvait Alexis Werné, directeur du service de la formation des adultes (SFA) au ministère de l'Éducation nationale. Guy Bentner, directeur adjoint de ce service, était le chargé de direction du CLL.

Par la loi du 22 mai 2009, le CLL a été transformé en institut national des langues (INL), une administration publique placée sous l'autorité du ministre de l'Éducation nationale. La fonction de professeur de langue luxembourgeoise a été arrêtée par la même loi.

Bâtiment INL au Glacis

L'INL a pour principales missions de dispenser des cours de langues pour adultes et de certifier les compétences linguistiques par des diplômes et certificats. À côté du luxembourgeois, l'INL propose des cours pour les langues allemand, anglais, chinois, espagnol, français, italien, néerlandais et portugais. En septembre 2022, l'INL a lancé sa nouvelle plateforme d'apprentissage de la langue luxembourgeoise en ligne « llo.lu ».

Une contribution importante à la formation du luxembourgeois est également fournie par l'université du Luxembourg. Une série de dictées avec transcriptions est publiée sur le site web « infolux.lu » de l'université. Les oratrices et orateurs sont Caroline Doehmer, Nathalie Entringer, Sara Martin et Jemp Schuster.

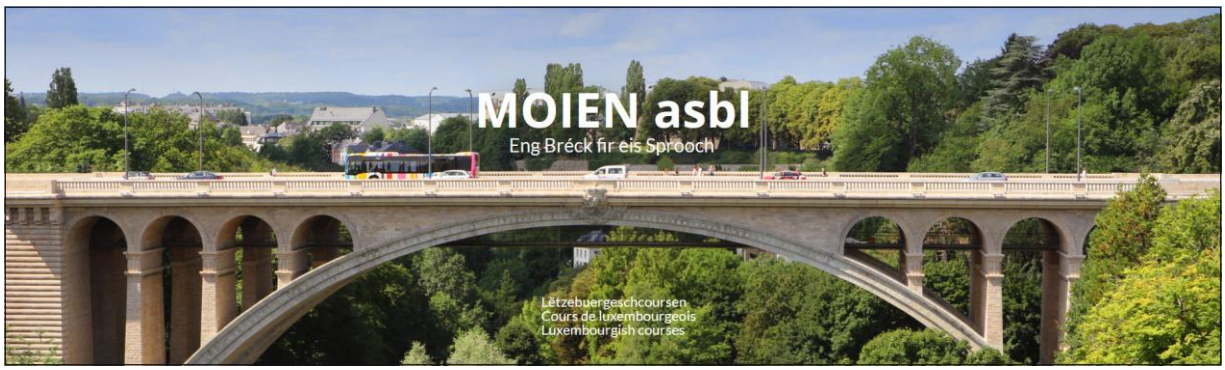
Il ne faut pas oublier les communes du pays et de la grande-région dont une majorité organise des cours luxembourgeois au niveau local. Le portail « Lifelong-Learning » présente une offre complète des cours proposés par les différents acteurs sur ses pages web.

On trouve toutefois les vrais pionniers concernant l'organisation de cours de formation pour la langue luxembourgeoise parmi les écoles privées : GrandJean (> 1949), Prolingua (> 1983), Inlingua (> 1993), Tower Training,(> 2001), languages.lu (> 2003), Cap Langues (> 2005), Berlitz (> 2007), Learn Luxembourgish (> 2010), SpeakUp (> 2019), ETIC Lëtzebuergesch. Les écoles privées se sont fédérées dans l'association pour l'enseignement du luxembourgeois.

Un premier projet de formation en ligne assistée pour le luxembourgeois fut lancé en 2004 par Daniela Clara Moraru sur son site web « languages.lu », en collaboration avec « lesfrontaliers.lu ». Originaire de Roumanie, elle continue à diriger son entreprise de formation linguistique à côté de nombreuses autres activités bénévoles. L'association « Actioun Lëtzebuergesch » lui a décerné en janvier 2023 la « plaquette Dicks-Rodange-Lentz » en bronze pour ses mérites dans la promotion de la langue luxembourgeoise.

Keynote speaker Clara Moraru

Après le secteur public et le secteur privé, c'est le tour des associations à mettre en vitrine. Je n'ai plus besoin de présenter l'« Actioun Lëtzebuergesch » qui est le doyen parmi ces organisations. Une association sans but lucratif qui a comme premier objectif l'enseignement et l'apprentissage de la langue luxembourgeoise et la création et gestion d'une plateforme d'échange et d'information entre formateurs s'appelle « Moien asbl - Eng Bréck fir eis Sprooch ».



Page web d'accueil moienasbl.lu

Elle a été constituée le 16 avril 2005 notamment par des formatrices et traductrices du luxembourgeois. L'association « Moien » publie un bulletin annuel, appelé « Infoblat », avec un résumé des activités de l'année.

Un autre pionnier parmi les organisateurs de cours de formation du luxembourgeois n'est pas vraiment une association, mais un duo de chevilles ouvrières qui gèrent le portail web « bonjour.lu ». Il s'agit de Jérôme Lulling introduit dans un chapitre précédent, et de sa tante Astrid Lulling qu'on n'a pas besoin de présenter. Elle fait partie des grandes personnalités du Grand-Duché.

Astrid et Jérôme Lulling

Jérôme Lulling est enseignant de Luxembourgeois auprès de la Ville de Luxembourg. Il a rédigé de nombreux exercices (applis www.exercice.lu) hébergés sur la plateforme en ligne suisse « LearningApps.org ». Jérôme Lulling est en outre l'auteur de nombreux livres, dictionnaires, leçons, podcasts, vidéos, jeux et autres supports didactiques que nous allons découvrir dans le chapitre suivant. En 2019, il a reçu la « plaquette Dicks-Rodange-Lentz » en argent pour ses mérites.

Last but not least il faut souligner que les associations spécialisées dans l'assistance des immigrés et dans l'insertion sociale favorisent également l'apprentissage du luxembourgeois. L'association de soutien aux travailleurs immigrés (ASTI) et le comité de liaison des associations issues de l'immigration (CLAE) sont deux exemples sur le plan national, le « café des langues » à Pütscheid est un exemple sur le plan local.

Supports didactiques : livres, sites web, vidéos, apps

Le support didactique le plus populaire pour apprendre le luxembourgeois est le livre. Le nombre de livres dédiés à l'apprentissage de la langue nationale a augmenté au fil des années. Il suffit de visiter le rayon des langues dans une librairie ou bibliothèque ou de consulter les pages web d'un magasin d'édition en ligne pour voir la variété et le volume des livres disponibles sur le marché. Le cadre de ma présente publication ne permet pas de citer les noms de tous les auteurs et contributeurs à la rédaction de ces livres de support à la formation. Je me limite donc à présenter quelques exemples.

Les trois tomes « Schwätzt Dir Lëtzebuergesch », publiés par l'INL depuis 2015, sont les manuels les plus vendus. Les dictionnaires bilingues LuxDico, assemblés par Jérôme Lulling et François Schanen, sont toujours très appréciés par le public. La première publication de la version français-luxembourgeois remonte à 2005. Une version allemand-luxembourgeois a été ajoutée quelques années plus tard. Aujourd'hui LuxDico est également accessible en ligne et disponible comme application pour smartphones.

Schwätzt Dir Lëtzebuergesch?

La situation était différente dans le passé. Il y a cent ans il n'existait qu'un seul livre d'apprentissage du luxembourgeois, à savoir « Das Luxemburgische und sein Schrifttum », rédigé en 1914 par Nik Welter. Né en 1871 à Mersch, Nik Welter a étudié les lettres et la philosophie à Louvain, Paris, Bonn et Berlin. Il était directeur général de l'instruction publique (équivalent d'un ministre de l'Éducation nationale de nos jours) entre 1918 et 1921. Nik Welter a écrit des poèmes, des pièces de théâtre, des romans et des études consacrées à l'histoire et à la théorie littéraires. Il a été le collaborateur de nombreuses revues littéraires luxembourgeoises et étrangères.


Das
Luxemburgische
und
sein Schrifttum
von Nikolaus Welter.
Neue durchgesehene Auflage.
Genehmigt von der Grossherzoglichen Unterrichts-Kommission.
Luxemburg. — Druck und Verlag von Gustave Soupert.
1923.

Nikolaus Welter

Le livre « Das Luxemburgische und sein Schrifttum » contient une orthographie connue sous le nom de « Welter-Engelmann Schreifweis ». René Engelmann est né en 1880 à Vianden. Il a étudié la linguistique à Paris, Berlin et Londres. Dans la suite, il a été chargé par le gouvernement luxembourgeois d'élaborer des règles pour l'orthographie chaotique de la langue luxembourgeoise. René Engelmann s'est suicidé en 1915. Bien que le livre d'apprentissage a été réimprimé plusieurs fois avec des corrections et utilisé dans les écoles jusqu'au début des années 1950, l'orthographie Welter-Engelmann n'a jamais été déclarée comme officielle.

La première orthographie luxembourgeoise officielle (OLO) a été introduite par arrêté ministériel du 5 juin 1946. Elle est connue comme « Margue-Feltes-Orthographie » en se référant aux noms des initiateurs, Nicolas Margue et Jean Feltes. Ce dernier est né en 1885 à Götzingen. Il suivait des études de philosophie et de philologie à Luxembourg, Paris, Munich et Londres. Après la deuxième guerre mondiale, il a été chargé par Nicolas Margue, ministre de l'Éducation nationale jusqu'en 1948, d'élaborer une nouvelle orthographie luxembourgeoise. Comme elle était basée sur la phonétique et se distinguait trop de l'allemand, la « ofizièl lezebuurger octografi » était très controversée et n'a jamais été acceptée par le peuple. On connaît la suite : la deuxième commission de dictionnaire a repris ses travaux en 1948 pour mieux faire.

À côté des livres, il y a les informations sur le web. Quelques références ont déjà été indiquées ci-avant. Une source très pertinente pour les internautes qui s'intéressent pour la langue luxembourgeoise constitue le « portail Lëtzebuerg » du site web « lb.wikipedia.org ». On y trouve dans la rubrique « D'Lëtzebuerger Sprooch » des données détaillées non seulement sur l'histoire et les auteurs luxembourgeois, mais également des règles d'écriture et des listes avec des proverbes et des expressions idiomatiques.


EIS SPROOCH
erausgi vun der Actioun Lëtzebuergesch
Nr. 3 / Nei Folleg / XII. Jor / 1973
Wat dran as:

Eis Sprooch No 3 1973

Une autre source bourrée d'informations pour les amateurs de la langue nationale est le site web www.actioun-letzebuergesch.lu. À côté des livres édités par l'« Actioun Lëtzebuergesch » à commander sur le site, on peut y télécharger les 37 numéros du magazine « eis sprooch », publiés entre 1972 et 1993.

Learn Luxembourgish

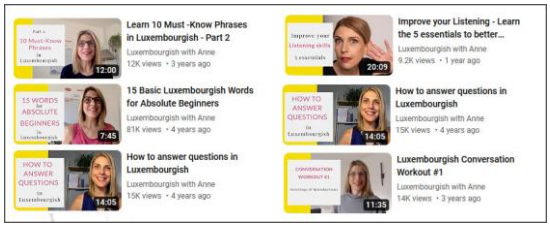
Sur Internet, on trouve également des sites web et des groupes sur les réseaux sociaux dédiés à l'apprentissage de la langue luxembourgeoise, créés et gérés par des particuliers.

« learnluxembourgish.com » est un premier exemple. La fondatrice est Liz Wenger qui a lancé ce projet en 2010 pour aider son mari canadien à apprendre le luxembourgeois. En 2011, elle a obtenu un certificat de formatrice du luxembourgeois par l'INL. En 2015, elle a écrit le premier livre d'apprentissage de la langue luxembourgeoise pour anglophones. Le livre peut être commandé en ligne sur son site web. Plus de 5.000 exemplaires ont été vendus jusqu'à présent dans le monde entier. Liz Wenger vit actuellement avec sa famille au Canada et continue à donner des cours luxembourgeois en vidéoconférence.

Un deuxième exemple est le site web « Lët's Learn Luxembourgish » géré par Tania Hoffmann depuis 2015. Elle est formatrice certifiée, traductrice, rédactrice et spécialiste du comportement des chiens. Elle est en outre le premier « Advanced Neurolanguage Coach » luxembourgeois, ce qui est symbolisé dans son logo. Tania Hoffmann a publié le livre « Verwiesselungsgefor » en 2015 et « Fester a Feierdeeg zu Lëtzebuerg » en 2022. Elle propose actuellement des cours luxembourgeois en vidéoconférence.

Logo Lët'z Learn

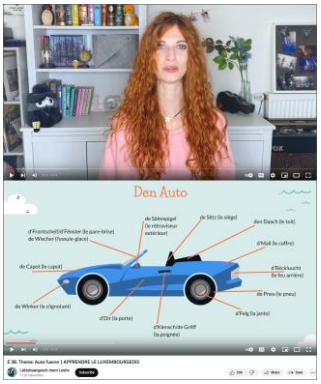
Si on entre les mots de recherche « lëtzebuergesch, luxembourgish, luxembourgeois ou luxemburgisch » sur Youtube, on trouve des centaines de vidéos pour apprendre la langue. On découvre quelques auteurs déjà connus comme Jérôme Lulling et Clara Moraru, mais également des nouveaux vidéastes comme Anne, Leslie et Jean-Paul.



Vidéos Luxembourgish with Anne sur Youtube

Anne Befort est formatrice en luxembourgeois passionnée, certifiée par l'INL. Elle a fondé en 2012 la « Babbelschoul » et le centre « Luxembourgish Language Coaching ». Sur sa chaîne Youtube « Luxembourgish with Anne » avec 13.700 abonnés, on trouve plus que 100 vidéos. La plus ancienne date de 2018, la plus récente a été mise en ligne il y a quelques mois. Chaque vidéo traite un sujet ou aspect spécifiques du luxembourgeois.

Leslie Schmit a étudié à l'université du Luxembourg. Elle est artiste et photographe et elle a été enseignante bénévole au centre de rencontre et d'information des jeunes (CRIJ) à Esch-sur-Alzette. Comme youtubeuse, elle a créé entre 2020 et 2022 une collection très riche de vidéos d'apprentissage du luxembourgeois, disponible sur sa chaîne Youtube « Lëtzebuergesch mam Leslie » avec 11.200 abonnés. Chaque édition est dédiée à un thème particulier, par exemple « rouler en voiture, vaccination, Noël, etc. » ou à une règle spécifique, par exemple « adjectifs, passé-composé, etc. ». Les illustrations dessinées pour présenter les différents termes associés aux sujets sélectionnés constituent elles seules déjà des petits chefs d'œuvre.



Lëtzebuergesch mam Leslie

Jean-Paul Piazzola est enseignant pour les langues luxembourgeoise, allemande et française. Il a fondé en 2015 le centre de formation « academia.lu ». Ses quelques dizaines de vidéos hébergées sur Youtube sont essentiellement des démonstrations de ses cours des différentes langues enseignées en présentiel, ce qui explique le faible nombre d'abonnés à sa chaîne Youtube.

Après les livres, sites web et vidéos, il nous reste à explorer les applications informatiques d'apprentissage du luxembourgeois. Un regard sur « AppStore » ou sur « GooglePlay », avec les mots de recherche appropriés, nous fait découvrir une liste de quelques apps dédiées à l'enseignement de la langue nationale. Parmi les programmes affichés, on retrouve des vieilles connaissances : LOD, LLO, LuxDico et Clara Muraro avec « 365 days luxembourgish »,


Alpaga

Deux applications dans la liste méritent d'être mises en évidence : BattaKlang Vocal et Aurelux.

La photo à gauche présente un alpaga. La question se pose quel est le lien entre ce mammifère et les nouvelles technologies ? La réponse est simple. La créatrice de l'application mobile « BattaKlang » est « Madame Paga » de Wormeldange-Haut qui élève des alpagas depuis 2011.

Son vrai nom est Béatrice Warichet. Jusqu'à fin 2015, elle était directrice et chef de projet auprès du cabinet d'audit Deloitte

Luxembourg. Aujourd'hui elle gère sa propre start-up « LaSauce s.à r.l. », immatriculé en 2013 et spécialisé dans le marketing digital et dans l'apprentissage ludique.

L'application mobile « BattaKlang » permet d'apprendre des mots du vocabulaire luxembourgeois par le jeu. Elle a été lancée en 2014 avec l'aide d'un développeur externe, d'abord en version muette, ensuite en version sonore.

App BattaKlang Vocal

Une application similaire pour l'apprentissage des mots en anglais, appelée « BattaKing », est également disponible sur « AppStore » et sur « GooglePlay ». Il faut désigner parmi les dessins affichés sur l'écran l'objet qui correspond au mot affiché ou prononcé. Fin novembre 2015, le nombre de téléchargements avait déjà dépassé le seuil de 10.000. Depuis le confinement Covid, Béatrice Warichet offre les achats intégrés pour les options des applications gratuitement.

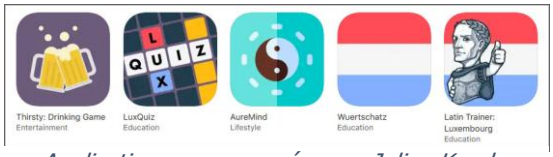
Wou ass de Kurf?

App Aurelux

« Aurelux » est également une application mobile pour apprendre le luxembourgeois d'une manière ludique avec des leçons, des questions et des jeux. Comme interface, on peut choisir parmi les langues suivantes : allemand, anglais, arabe, espagnol, français, italien, portugais et russe. L'application a été conçue par Aurélie Wagener et programmée par Julien Kessler en 2018.

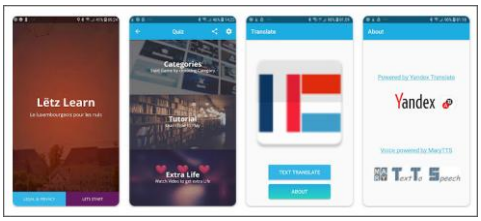
Aurélie Wagener est formatrice certifiée en luxembourgeois. Elle a fondé en juillet 2022 la société à responsabilité limitée simplifiée Aurelux avec objet principal l'apprentissage et la formation de la langue luxembourgeoise. La plateforme interactive « académie Aurelux » constitue un premier outil de support pour cette formation.

Julien Kessler est ingénieur en sécurité informatique auprès des CFL. À côté d'Aurelux, il a développé d'autres applications : Thirsty, AureMind, LatinTrainer, Wuertschatz et LuxQuiz.



Applications programmées par Julien Kessler

Les deux projets « BattaKlang » et « Aurelux » ont été présentés dans plusieurs médias au fil du temps, par exemple dans un petit reportage sur RTL TV en janvier 2019.



Le luxembourgeois pour les nuls

Un troisième projet d'apprentissage du luxembourgeois développé pour les smartphones Android doit être mise à jour pour fonctionner correctement sur les systèmes d'exploitation récents, mais il mérite d'être présenté pour son caractère ingénieux. Il s'agit de l'application « Lëtz Learn – Le luxembourgeois pour les nuls », programmée par Edouard Jimmy Kanku, alias « jimlux » en 2019. Il est technicien de support IT auprès d'une banque de la place. Une option intéressante de cette application est la possibilité d'écouter la prononciation luxembourgeoise d'un texte français, entré sur le clavier virtuel du smartphone. La traduction français-luxembourgeois est réalisée avec l'API de Yandex et le texte traduit est converti en parole avec l'API de synthèse vocale Marylux. Le code de l'application est disponible sur Github.