

La lexicographie informatisée: les vocabulaires nationaux dans un contexte européen

Dorothee Aquino-Weber, Yan Greub (éd.)

```

1 <xml>
2 <file name="GPSR_07_01_0001_tit1.tif"/>
3 <pb n="EAI:a"/>
4 <h><s n="EAI:a"/></h>
5 <file name="GPSR_07_01_0002_R0001.tif"/>
6 <pb n="EI"/>
7 <h><s n="EI"/></h>
8 <lb n="EI. 001"/><Z><Ap2>GLOSSAIRE</Ap2>
9 <lb n="EI. 002"/>DES
10 <lb n="EI. 003"/><Ap1>PATOIS DE LA SUISSE ROMANDE</Ap1></Z>
11 <file name="GPSR_07_01_0003_R0002.tif"/>
12 <pb n="EII"/>
13 <h><s n="EII"/></h>
14 <lb n="EII. 001"/><image/>
15 <lb n="EII. 002"/><P><AN>Soutenu par l'Académie suisse
16 <lb n="EII. 003"/>des sciences humaines et sociales
17 <lb n="EII. 004"/>www.assh.ch</AN></P>
18 <lb n="EII. 005"/><image/>
19 <lb n="EII. 006"/><P><AN>CONFÉRENCE INTERCANTONALE
20 <lb n="EII. 007"/>DE L'INSTRUCTION PUBLIQUE DE
21 <lb n="EII. 008"/>LA SUISSE ROMANDE ET DU TESSIN</AN></P>
22 <lb n="EII. 009"/><P> <AN>Université Unive
23 <lb n="EII. 010"/>de Neuchâtel</AN></P>
24 <file name="GPSR_07_01_0004_R0003.tif"/>
25 <pb n="EIII"/>
26 <h><s n="EIII"/></h>
27 <lb n="EIII. 001"/><Z><Ap2><b>GLOSSAIRE</b></Ap2>
28 <lb n="EIII. 002"/><b>DES</b></Ap2>
29 <lb n="EIII. 003"/><Ap2><b>PATOIS DE LA SUISSE ROMANDE</b></Ap2>
30 <lb n="EIII. 004"/><C>fondé par</C>
31 <lb n="EIII. 005"/><C>Louis GAUCHAT Jules JEANJAQUET Ernest TAPPOLET</C>
32 <lb n="EIII. 006"/><i>Ouvrage élaboré avec le concours de nombreux auxiliaires</i>
33 <lb n="EIII. 007"/><i>publié sous les auspices des Cantons romands</i>
34 <lb n="EIII. 008"/><i>avec l'appui financier de</i>
35 <lb n="EIII. 009"/><i>l'Académie suisse des sciences humaines et sociales</i>
36 <lb n="EIII. 010"/><Ap1>TOME VII
37 <lb n="EIII. 011"/><b>F</b></Ap1>
38 <lb n="EIII. 012"/><C>rédigé et publié par</C>
39 <lb n="EIII. 013"/><C>F. VOILLAT P.-H. LIARD H. GASSMANN H. CHEVALLEY</C>
40 <lb n="EIII. 014"/><C>A. HUBER E. FLÜCKIGER B. GROSS C. GODAT A. DIACONESCU</C>
41 <lb n="EIII. 015"/><Ap1><b>1</b><sup>re</sup></Ap1>
42 <lb n="EIII. 016"/><b>f-flyöka</b></Ap1>
43 <lb n="EIII. 017"/><C>librairie droz, Genève</C>
44 <lb n="EIII. 018"/><C>gessler. zahlen sa saint-blaise</C>
45 <lb n="EIII. 019"/>1989-2014</Z>
46 <file name="GPSR_07_01_0005_R0004.tif"/>

```



La lexicographie informatisée : les vocabulaires nationaux dans un contexte européen

Dorothee Aquino-Weber, Yan Greub (éd.)

Éditrice

Académie suisse des sciences humaines et sociales,
Laupenstrasse 7, Postfach, 3001 Bern
Telefon +41 (0)31 306 92 50, sagw@sagw.ch
www.assh.ch

Illustration de couverture

Mario Cafiso

Layout

Elodie Lopez (SAGW)

Impression et correctorat

Druck- und Werbebegleitung, 3098 Köniz

Première édition, 2020 (1400)

La brochure peut être commandée gratuitement au secrétariat de l'ASSH
ou à l'adresse www.sagw.ch/publikationen



Copyright: Ceci est une publication Open Access, distribuée sous les termes de la licence Creative Commons Attribution (<http://creativecommons.org/licenses/by/4.0/>). Le contenu de cette publication peut donc être utilisé, distribué et reproduit sous toute forme sans restriction, à condition que l'auteur et la source soient cités de manière adéquate.

Recommandation pour citer le texte:

Dorothee Aquino-Weber, Yan Greub (éd.) (2020)

La lexicographie informatisée: les vocabulaires nationaux dans un contexte européen
Académie suisse des sciences humaines et sociales.
Swiss Academies Reports 15(1).

ISSN (print): 2297 – 1815

ISSN (online): 2297 – 1823

DOI: <https://doi.org/10.5281/zenodo.3550494>

Inhalt

Zur Publikation Manuela Cimeli	7
Introduction Yan Greub	9
Traitement des données et mise en place d'une plateforme lexicographique numérique pour le Tesoro della Lingua Italiana delle Origini (TLIO) Salvatore Arcidiacono	17
L'état de la numérisation du LEI: un rapport Elton Prifti	27
1. Niveaux de numérisation	28
2. Notes sur les tentatives d'informatisation du LEI	29
3. État actuel de la numérisation	31
3.1 <i>Rétrodigitalisation</i>	32
3.2 <i>Numérisation des pièces publiées</i>	33
3.3 <i>Informatisation et réorganisation du système rédactionnel</i>	34
3.3.1 Principes	35
3.3.2 Principaux outils	36
De Frantext 1 à Frantext 2: la cure de jouvence d'une vieille dame Véronique Montémont	41
Frantext canal historique	42
Une modernisation nécessaire	44
Du neuf avec du vieux	48
La mue	52
<i>Retraitement du corpus</i>	52
<i>Langage de requêtes</i>	55
<i>Moteur de recherche</i>	60
Versione online del VSI e sviluppi futuri del progetto di informatizzazione Dafne Genasci e Dario Petrini	67

Dicziunari Rumantsch Grischun – Der lange Weg zur Retrodigitalisierung und zur Online-Publikation	79
Ursin Lutz	
Einleitung	79
Grundlagen	80
<i>Die ersten Schritte mit dem Computer</i>	80
<i>Herausforderung Sonderzeichen und deren Eingabe</i>	81
Projekt «Digitales Wörtermuseum»	84
Das neue Redaktionssystem	85
Das DRG-Online, ein Projekt in drei Etappen	88
<i>Retrodigitalisierung der DRG-Bände 1–13</i>	88
<i>TEI-konforme XML-Auszeichnung der elektronischen Abschrift</i>	91
<i>Erstellung eines Webportals als Benutzer-Frontend</i>	91
Herausforderungen	92
Lösungsansätze	94
Fertigstellung und Inbetriebnahme der ersten Version	95
<i>Festlicher Akt vom 7. Dezember 2018</i>	95
<i>Funktionalität und Funktionsweise der ersten Version</i>	95
Ausblick	99
<i>Fortlaufende Einpflege der publizierten Artikel</i>	99
<i>Erweiterung der Funktionalitäten</i>	100
Im Übergang zum digitalen Wörterbuch. Zum Stand der Digitalisierung des Schweizerischen Idiotikons: Rückblick und Ausblick	103
Tobias Roth	
Rückblick und Status quo	103
<i>Website und Lemmeregister</i>	103
<i>Digitales Faksimile</i>	104
<i>Automatische Texterkennung und Volltextsuche</i>	105
<i>Semantikregister</i>	107
<i>REST-API und Mobilversion</i>	109

Weitere digitale Projekte	111
<i>ortsnamen.ch</i>	111
<i>Materialien des Sprachatlas der deutschen Schweiz (SDS)</i>	111
<i>Schweizer Textkorpus</i>	112
<i>Aktiv archivierte Projekte</i>	112
Digitalisierungsstrategie	113
Übergang zum digitalen Wörterbuch	114

Rétrodigitalisation du Glossaire des patois de la Suisse romande : inauguration du portail web

Alexandre Huber	119
Historique	119
Exemples de recherches	120
<i>Recherches dans le sens patois-français</i>	121
<i>Recherches dans le sens français-patois</i>	124
<i>Recherches encyclopédiques</i>	127
<i>Recherches en onomastique</i>	129
<i>Opérateurs logiques</i>	131
Projet d'une base de données iconographiques	132

Zur Publikation

Manuela Cimeli

Die dialektalen und historischen Wortschätze unserer Landessprachen werden durch die vier Nationalen Wörterbücher erklärt: Das *Wörterbuch der Schweizerdeutschen Sprache (Schweizerisches Idiotikon)*, das *Glossaire des Patois de la Suisse romande*, das *Vocabolario dei dialetti della Svizzera italiana* sowie das *Dicziunari Rumantsch Grischun*.

Mit der 2009 ins Leben gerufenen Reihe zeigt die SAGW in loser Folge den Wert und den Nutzen der Dialektforschung wie auch ganz grundsätzlich die Relevanz unserer Dialekte für unser kulturelles und sprachliches Erbe sowie für die sprachliche und kulturelle Identität der Schweiz auf.

Als das Glossaire des patois de la Suisse romande (GPSR) am 12. September 2018 die neue Version seines Webportals einem grösseren Publikum präsentierte, waren auch Vertreter der anderen drei Nationalen Wörterbücher anwesend und zeigten ihrerseits auf, welchen Weg der Digitalisierung sie selbst für ihre Werke gewählt haben. Weitere Präsentationen erfolgten durch Vertreter des Tesoro della Lingua Italiana delle Origini (TLIO), des Lessico Etimologico Italiano (LEI) sowie von Frantext. Es ist interessant zu sehen, wie die Repräsentanten und Repräsentantinnen der verschiedenen Wörterbuch-Projekte die unterschiedlichen Herausforderungen der Digitalisierung angegangen sind bzw. angehen. Während die Veranstaltung im Herbst 2018 eine gute Gelegenheit zum Austausch der Experten und Expertinnen untereinander bot, gewährt der nun vorliegende zwölfte Band der Reihe «Sprachen und Kulturen», die Publikation «La lexicographie informatisée», allen Interessierten einen Einblick in die Wörterbuch-Entwicklung.

Sie können den vorliegenden Band sowie die übrigen Texte der Reihe kostenlos auf unserer Website herunterladen: www.sagw.ch/publikationen (Sprachen und Kulturen) oder www.assh.ch/publications (langues et cultures).

Introduction¹

Yan Greub

Le 12 septembre 2018, le Glossaire des patois de la Suisse romande (GPSR) ouvrait au public la nouvelle version de son portail web, qui ajoutait aux fonctionnalités existantes la possibilité de recherches plein texte. À la cérémonie officielle marquant cet événement, qui réunit dans la matinée des représentants des autorités, la presse et un large public, il nous avait semblé approprié d’associer, durant l’après-midi, un colloque durant lequel les responsables de plusieurs grands projets de lexicographie historique et variationnelle présenteraient leurs résultats et les problèmes auxquels ils avaient été confrontés ; ce volume regroupe, sous une forme un peu plus étendue, ces communications². Il les complète d’une part par une présentation du site web du Glossaire (il n’y en avait pas eu lors du colloque, car son équivalent avait trouvé place dans la cérémonie officielle du matin), et d’autre part par un exposé sur l’informatisation du *Dicziunari Rumantsch Grischun* (DRG), dont le site web n’existait pas encore au moment du colloque, et qui est maintenant pleinement opérationnel³ à l’adresse online.drg.ch.

La publication des grands dictionnaires historiques et polydialectaux sur internet est une question d’une forte actualité, au point que l’Union européenne a subventionné par plusieurs programmes récents un effort de coordination sur le web⁴. Par ailleurs, que ce soit ou non par une rencontre de hasard, plusieurs sites internet concernés ont connu récemment des transformations notables, qui leur permettront de répondre aux besoins – en constante évolution – de leur public et de le toucher plus largement.

Il est probable qu’en 2019 le public savant considère qu’une institution absente d’internet est insuffisamment active, et le Glossaire, comme les autres Vocabulaires nationaux sans doute, se sentait tenu d’offrir ce service et d’assurer

1 Ce numéro a été réalisé avec la collaboration de Maguelone Sauzet.

2 Cependant, l’exposé de Sabine Tittel, rédactrice au Dictionnaire étymologique de l’ancien français, sera publié dans un autre cadre et n’est donc pas repris ici.

3 Il a été inauguré le 7 décembre 2018.

4 On pense en particulier aux actions COST 1005 (Medioevo Europeo), qui n’était pas limitée à la lexicographie, et 1305 (European network of e-lexicography), qui n’était pas limitée aux dictionnaires historiques et variationnels.

cette visibilité. Mais en réalité, le besoin de disposer du Glossaire sous forme électronique existait indépendamment de cette impulsion extérieure, comme on va le voir.

Tout projet de publication se cherche un lectorat, et cela est d'autant plus vrai lorsqu'il s'étend sur des années, des dizaines d'années, voire plus d'un siècle, comme c'est le cas pour les dictionnaires présentés dans le cadre de notre journée. Mais la question est encore plus urgente (et à la fois délicate) pour les quatre Vocabulaires nationaux suisses : si l'on prend l'exemple du Glossaire, la destination de l'œuvre vers un large public a été un élément essentiel du projet de Louis Gauchat, son fondateur, et le système de classement alphabétique comme le système de transcription phonétique dépendent de la volonté de rendre le Glossaire lisible par les non-spécialistes. On ne pouvait pas attendre moins, d'ailleurs, d'une œuvre orientée vers un but patrimonial, et qui reposait sur une collaboration effective avec la population romande. La part étroite prise par les cantons romands dans la création du Glossaire, puis dans son subventionnement et son administration, n'a fait que renforcer le sentiment d'un devoir de publicité. Si les moyens utilisés au XX^e siècle pour garantir l'accès au Glossaire à un lectorat large et divers ont été une diffusion dans les bibliothèques publiques et auprès d'abonnés individuels, ainsi qu'une certaine attention à la lisibilité du dictionnaire, un accès facilité sur internet est aujourd'hui une composante essentielle du service que nous devons au public. Nous avons la chance qu'un accord avec notre éditeur (Droz, à Genève) nous permette de donner un accès gratuit au Glossaire, avec un peu de retard sur la publication papier.

Les trois autres Vocabulaires nationaux suisses, qui sont placés dans la même situation, ont choisi des approches semblables. Le *Vocabolario dei dialetti della Svizzera italiana* (VSI), par exemple, est intégré dans une série de publications qui présentent le lexique de la Suisse italienne sous des angles différents, et multiplie ainsi les moyens d'accès à sa connaissance, tandis que la rédaction de l'*Idiotikon*, qui est présent en ligne depuis 2010, présente quelques-unes de ses analyses dans une émission radiophonique régulière.

Si la volonté de répondre aux besoins du public a eu une importance essentielle dans le processus d'informatisation de nos dictionnaires, elle n'a pas été le seul facteur à l'œuvre, car les besoins des rédacteurs rendaient eux aussi nécessaire une informatisation, et ont pu dans certains cas la déterminer entièrement. On

pourra lire dans ce volume comment des conditions différentes ont abouti, dans plusieurs projets lexicographiques, à l'élaboration de systèmes informatiques complexes, qui ont pu fournir la base des instruments de recherche et de lecture mis à la disposition du public.

Ce public, d'ailleurs, est diversement composé. Les projets d'informatisation peuvent bien avoir pour ambition d'élargir leur lectorat parmi les non-spécialistes (et en particulier, pour les Vocabulaires nationaux, parmi les patoisants), mais nous devons aussi tenir compte du fait que les usagers les plus assidus sont les spécialistes: linguistes en général, et spécialistes des domaines décrits en particulier. La lexicographie sous forme informatisée s'adresse à eux autrement qu'au grand public. Une catégorie encore différente est celle des premiers utilisateurs: les rédacteurs eux-mêmes, qui étaient bien placés pour créer des outils de recherche et de lecture adaptés à leurs besoins.

L'informatisation des dictionnaires historiques et polydialectaux est une problématique très actuelle, mais elle n'est pas neuve pour autant. Le portail web du Glossaire repose directement sur des travaux qui ont débuté il y a vingt ans: Frantext est la mise à la disposition de tous d'une base de données qui ont été réunies à partir des années 1960, et l'informatisation de plusieurs entreprises lexicographiques a connu une longue histoire. Les communications rassemblées dans ce volume font allusion à ces histoires, qui ont parfois été, malgré beaucoup de bonne volonté et d'intelligence, pleines de détours et de difficultés. La possibilité qui nous est offerte ici de comparer les problèmes qu'ont identifiés plusieurs entreprises lexicographiques et les solutions qu'elles leur ont données démontre la grande diversité des cas d'espèce: si le texte des dictionnaires a été le plus souvent mis en ligne, tout le reste change, comme changent aussi les outils informatiques utilisés par les rédactions. On verra aussi que les contributeurs ont choisi de mettre l'accent sur des aspects différents du développement informatique de leurs dictionnaires.

Quoi qu'il en soit, la longue histoire de ces développements informatiques est celle d'une réponse à des questions qui s'étaient posées plus tôt encore, et qui touchaient l'accès aux données: les exigences particulières à un vocabulaire polydialectal faisaient que le choix des entrées (la macrostructure du dictionnaire) et de leur forme avait donné lieu à des réflexions délicates, dont on trouve une trace dans l'introduction du premier volume du Glossaire. Mais pour plu-

sieurs dictionnaires, l'informatisation découle d'un autre moyen d'accès aux données : les index de fin de volume, pour la réalisation desquels des systèmes informatiques ont été mis en place, avant d'être étendus à des fonctions plus vastes.

Pour mettre en lumière et en perspective les outils informatiques proposés par le Glossaire, nous avons voulu les comparer à ceux des ouvrages lexicographiques les plus voisins. En font partie en premier lieu, naturellement, les trois autres Vocabulaires nationaux, qui sont liés au Glossaire par une relative unité de conception et des buts communs. À l'intérieur de la linguistique historique romane, la plus importante entreprise lexicographique actuelle est le *Lessico etimologico italiano* (LEI), et son nouveau responsable, Elton Prifti, vient précisément de contribuer à mettre en place un nouveau processus rédactionnel, lourdement assisté par des outils informatiques. Les deux instituts de recherche scientifique chargés chez nos voisins français et italien de tâches lexicographiques, l'Opera del vocabolario italiano (OVI) et le laboratoire ATILF, successeur de l'Institut national de la langue française, mettent à la disposition du public, sur internet, de très importants dictionnaires, ainsi que d'autres outils de recherche, qui sont aussi des instruments de travail du Glossaire. Avec un propos, des objets et des moyens différents des nôtres, ils nous présentent donc des parallèles significatifs et des exemples à observer.

Le *Tesoro della Lingua Italiana delle Origini* (TLIO), réalisé à l'institut Opera del vocabolario italiano (du Centre national de la recherche italien), à Florence, a dès son origine utilisé des outils informatiques : toute la rédaction a été réalisée informatiquement, ce qui a permis de publier les articles progressivement (la publication est en cours depuis 1997) et de donner un caractère évolutif au dictionnaire. La communication de Salvatore Arcidiacono présente le nouveau système de rédaction dont il est en train de terminer l'élaboration. C'est donc un aspect de l'informatisation orienté spécialement vers les collaborateurs du dictionnaire. Le TLIO présente une situation remarquable, car ses logiciels gèrent à la fois une base textuelle considérable, un dictionnaire qui l'exploite, et des projets lexicographiques associés et dépendants de lui.

C'est aussi un aspect qui intéresse particulièrement Elton Prifti, directeur du LEI, puisque ce dictionnaire vient de mettre en place un système de rédaction assistée par ordinateur. Elton Prifti intègre cependant le traitement de cette question dans un modèle plus général des niveaux d'informatisation d'un dictionnaire. Sa communication expose l'histoire, mais surtout l'actualité de l'ensemble des problèmes informatiques auxquels est confronté le LEI : ceux-ci touchent aussi bien la publication de l'œuvre que la numérisation de la documentation (qui sert à la fois à faciliter le travail rédactionnel et à sauvegarder les documents) et le processus de rédaction lui-même. L'informatisation dont il est question ici est à la fois celle qui s'adresse aux utilisateurs professionnels (le LEI est destiné à être employé par un public spécialisé, italianiste ou romaniste, linguiste ou philologue, spécialiste des dialectes ou des périodes anciennes de la langue) et aux rédacteurs eux-mêmes.

Frantext n'est pas un dictionnaire mais une base textuelle. Cette base descend cependant directement des outils informatiques extrêmement précoces qui ont servi à réaliser le *Trésor de la langue française*, le grand dictionnaire français du XX^e siècle. Une réécriture récente de l'outil de recherche a induit une modification des modes d'interrogation, et en partie de l'organisation de la base. Plusieurs projets lexicographiques menés au laboratoire ATILF sont liés au même outil de recherche, et pourraient donc être amenés à subir eux aussi des modifications. Véronique Montémont, qui a été plusieurs années en charge de Frantext, présente l'histoire de l'élaboration des nouveaux modes d'interrogation et fournit une aide détaillée à l'utilisation des outils riches et complexes dont l'utilisateur dispose désormais, grâce à des exemples. Elle s'interroge enfin sur la position de la base de données Frantext dans un contexte complètement transformé par l'apparition de bases conçues différemment, mais parfois beaucoup plus étendues, à commencer par celles qu'offre Google. Frantext, qui n'est accessible que par abonnement, s'adresse en priorité, par conséquent, à un public qui travaille dans une université ou un institut de recherche ; l'autre produit dérivé du *Trésor de la langue française*, le *Trésor de la langue française*

informatisé (TLFi), est au contraire utilisé par un public considérable, puisqu'il est consulté chaque jour par des centaines de milliers de lecteurs. Les fonctionnalités de l'interface de recherche seront nécessairement différentes dans ce cas.

La présentation de Dafne Genasci et Dario Petrini porte sur le VSI online, tel qu'il existe actuellement et tel qu'il sera développé. On comprend ici que la mise à la disposition du public d'une recherche plein texte est nettement moins coûteuse que la construction d'un outil de recherche par champs déterminés : aussi précise et contrainte que soit la forme rédactionnelle des dictionnaires pris en considération ici, l'intégration dans un simple formulaire de leurs indications étymologiques demande une analyse, et ne peut être réalisée automatiquement.

Ursin Lutz donne des informations détaillées sur l'histoire de l'informatisation de la rédaction du DRG, et montre que celle-ci a été marquée par des pertes considérables. C'est une situation qu'auront connue plusieurs dictionnaires utilisant des logiciels propriétaires, et à laquelle le DRG a répondu en faisant élaborer un système propre au dictionnaire et en stockant toutes ses données en format XML. Lors du processus de rétrodigitalisation des volumes précédents du dictionnaire, il n'a pas été possible de récupérer les fichiers numériques existants, et le dictionnaire a donc dû procéder à une (double) saisie de l'ensemble du texte. La présentation d'Ursin Lutz nous fait entrer dans le monde concret des incompatibilités informatiques auxquelles sont confrontés les auteurs de dictionnaire, et surtout de l'instabilité consubstantielle des logiciels. Même s'il ne le dit pas, on ne peut pas être sûr que l'avenir soit tout à fait exempt de ce type de déboires.

Le *Schweizerisches Idiotikon* se trouve dans une situation nettement différente de celle des autres Vocabulaires nationaux, puisque 50 000 utilisateurs consultent son site chaque mois. Cela n'empêche pas que plusieurs des problèmes auxquels il a été confronté et des solutions qu'il y a trouvées soient semblables. Tobias Roth présente l'histoire de ces solutions et la stratégie de développement qui a été adoptée. Il insiste justement sur le fait que l'*Idiotikon*, malgré tout le travail de rétroconversion et malgré les collaborations auxquelles

il est associé, n'est pas véritablement un dictionnaire électronique, mais plutôt un dictionnaire présenté sous forme numérique ; on peut d'ailleurs dire la même chose de la plupart des dictionnaires présentés ici, et en particulier des Vocabulaires nationaux. Cela n'empêche pas l'introduction de fonctionnalités (et en particulier de modes d'interrogation) complexes.

Il est inutile de s'étendre longuement sur le cas du *Glossaire des patois de la Suisse romande*, dont il a déjà été question. Après un rapide historique, Alexandre Huber montre dans sa communication que les méthodes de recherche que le portail web met désormais à la disposition de tous peuvent servir directement au grand public, et envisage plusieurs cas de figure pour suggérer dans chaque cas les manipulations appropriées de l'interface de consultation.

Des sept exposés rassemblés ici ressortent certaines conclusions qui sont peut-être applicables généralement. Tout d'abord, les coûts de l'informatisation sont très importants, en argent et en temps. Comme l'informatisation sert à créer un moyen d'accès à une connaissance scientifique inatteignable sinon, cet argent est bien dépensé, d'autant plus que l'accès du grand public à nos matériaux et à nos analyses en est facilité. Dans le cas des Vocabulaires nationaux de la Suisse, comme dans celui de projets nationaux tels que le *Trésor de la langue française* ou le TLIO, le citoyen, qui paie (à travers ses impôts) pour la création ou le maintien d'une connaissance, et qui possède une connaissance directe d'une partie au moins de l'objet étudié (les langues nationales), a un droit éminent à accéder à cette connaissance, et il n'y a donc rien d'exagéré à lui consacrer des efforts significatifs. Il n'en reste pas moins que la force de travail consacrée à ces tâches est soustraite à la production de connaissance elle-même, et doit donc être utilisée avec efficacité.

De ce point de vue, les efforts d'informatisation qui sont spécifiquement destinés à rendre le travail rédactionnel plus efficient⁵ doivent être évalués avec une attention particulière: il paraît tout à fait clair qu'ils ont abouti à une meilleure utilisation des dictionnaires eux-mêmes par leurs rédacteurs, et donc à un gain en qualité. Le gain de temps, quant à lui, est sans doute réel, mais là où la com-

5 Indépendamment, par conséquent, des mises en ligne des dictionnaires, qui servent certes aux rédactions, mais ne leur sont pas spécifiquement destinées.

paraison est possible⁶, il n'occasionne pas non plus une révolution du rythme de la publication. Dans ces conditions, le rapport entre le coût (en temps de travail) d'une réforme et ses résultats attendus doit être estimé avec soin.

Si l'informatisation des dictionnaires a pu entraîner un travail de Sisyphe du fait de la caducité de certains programmes informatiques, les solutions choisies actuellement sont relativement plus stables. La stabilité des systèmes informatiques est cependant loin d'être garantie, et dépend de la capacité des institutions lexicographiques à garantir leur entretien et leur mise à jour régulière. La question de l'équilibre entre publication papier et publication électronique reste posée.

6 Elle ne l'est pas pour le TLF ou le TLIO, dont toute la rédaction s'est faite avec le même type d'assistance informatique, et elle ne l'est pas encore pour le LEI, dont le système de rédaction informatisée vient seulement d'être mis en place.

Traitement des données et mise en place d'une plateforme lexicographique numérique pour le Tesoro della Lingua Italiana delle Origini (TLIO)

Salvatore Arcidiacono

Le *Tesoro della Lingua Italiana delle Origini* (TLIO) est un dictionnaire historique de l'italien ancien, basé sur toute la documentation existante, depuis le premier document (qui date de 960) jusqu'à la fin du XIV^e siècle. Le TLIO est élaboré à l'Istituto Opera del Vocabolario Italiano (OVI) du Consiglio Nazionale delle Ricerche (CNR), et il représente la première branche chronologique d'un grand vocabulaire historique de la langue italienne. Dès son commencement, dans les années 60 du siècle dernier, le projet a adopté les technologies numériques, en se servant de manière constante des nouvelles acquisitions et des avancements méthodologiques apportés par l'informatique. Depuis 1997, le TLIO est publié en accès libre exclusivement sur le site web de l'OVI¹, au fur et à mesure du progrès de sa rédaction², sans attendre la fin des travaux lexicographiques et sans respect de l'ordre alphabétique, en suivant une approche qu'on pourrait dire évolutive: « La struttura modulare della voce e il metodo di pubblicazione elettronica in rete consentono successivi sviluppi sistematici senza imporre la modifica delle parti non interessate » (Beltrami 1998: 277)³.

Cette modularité est ancrée dans une microstructure qui définit de manière rigoureuse l'organisation des informations, de telle sorte qu'on a pu récemment mettre en place l'adaptation du dictionnaire (rédigé à l'origine en format Word Office) aux nouveaux formats standard de marquage numérique, tels que XML/TEI⁴. Pluto (Piattaforma Lessicografica Unica del Tesoro delle Origini) est un nouveau système lexicographique en cours de développement à l'OVI, qui vise à permettre l'élaboration, la gestion et la publication du TLIO et des ressources qui s'y rattachent, dans le cadre du nouvel environnement technologique.

1 <http://www.oivi.cnr.it>.

2 Le TLIO se compose de 41 500 articles, qui représentent plus du 70 % du total prévu, dont 36 000 déjà publiés sur le site (<http://tlio.oivi.cnr.it>).

3 Une version mise à jour des Norme est disponible sur le site web de l'OVI: <http://tlio.oivi.cnr.it/TLIO>. Pour une définition de « lexicographie évolutive », voir Martin (2007).

4 Le schéma XML et le logiciel de conversion ont été présentés par Andrea Boccellari (2012) et Andrea Boccellari et Domenico Iorio-Fili (2013).

La plate-forme se sert de Lexicad, une bibliothèque de logiciels conçue pour la réalisation de dictionnaires *web-based*, qui permet de définir en termes abstraits les données et les procédures d'un système lexicographique informatisé. On encourage ainsi un cercle vertueux d'échange et de réutilisation de modèles et outils et, dans le même temps, on offre une grande flexibilité au développeur, ce qui lui permet d'adapter le système aux besoins spécifiques du projet lexicographique (Arcidiacono 2019).

Pluto intègre, dans une architecture unique, les systèmes informatiques des divers logiciels impliqués dans l'élaboration du TLIO (à savoir : bases de données bibliographiques, documentation philologique, outils pour la gestion du processus de rédaction, etc.), et permet aussi à chaque ressource de partager les données avec toutes les autres, en les optimisant et en augmentant leur potentiel d'analyse.



Fig. 1. Logo de la plateforme Pluto

Du point de vue de l'éditeur, le logiciel constitue un point d'accès unique simplifié, conçu pour une gestion centralisée du projet lexicographique. Le panneau de contrôle est marqué par une approche collaborative, avec un système flexible qui permet de définir clairement les rôles, les autorisations d'accès et les responsabilités des utilisateurs et d'établir avec précision les opérations que chaque collaborateur est autorisé à effectuer et les données auxquelles il pourra avoir accès.

Les entreprises lexicographiques comportent souvent des activités dont la durée s'étend sur une très longue période, ainsi que des flux de travail relativement complexes. Pour organiser l'équipe en vue de la réalisation des objectifs fixés, Pluto comprend un dispositif de monitoring informatisé des activités, qui enregistre un historique des opérations effectuées. Au début du processus d'élaboration, l'article lexicographique n'est qu'un mot-vedette accompagné par une catégorie grammaticale et éventuellement par des notes de prérévision. La première étape consiste à attribuer l'article à un rédacteur, à qui le système confère les droits d'intervention et de modification du fichier. Le processus se déroule ensuite tout au long d'une série d'étapes de révision, jusqu'à la publication finale.

Le masque de saisie des entrées représente le cœur du système. Cette interface est conçue pour la création de données semi-structurées en format standard, sans que l'utilisateur soit expert du langage XML. Le masque de saisie diminue autant que possible l'effort nécessaire au formatage de l'entrée, en facilitant le travail du rédacteur par des listes déroulantes, des champs à choix multiple et des suggestions diverses. Dans le masque de saisie, la microstructure du dictionnaire est subdivisée en éléments minimaux, qui concernent des ensembles d'informations homogènes. Chaque élément est conçu pour faciliter et accélérer le travail du rédacteur, en utilisant tous les moyens nécessaires pour maîtriser la complexité des données et réduire au minimum (ou même éliminer complètement) les tâches mécaniques et répétitives. Certains formulaires peuvent être remplis automatiquement par le logiciel sur la base des données collectées auparavant. Dans bien des cas, chaque élément est accompagné par des applications spécifiques, grâce auxquelles le travail est allégé et accéléré. L'automatisation contribue aussi à réduire la probabilité d'erreurs, notamment dans toutes les sections du dictionnaire qui concernent de données numériques ou listes ordonnées. De plus, Pluto permet de définir un ensemble de procédures de contrôle d'exactitude qui exécutent des vérifications en temps réel lors de la saisie de données et qui peuvent être réunies en procédures complexes pour confirmer la régularité formelle de la base de données tout entière.

TLIO - Maschera di redazione

Novità • Gestione • Amministratore

Rinvia: Lemma: Diaamb: C.p.:

- ~
- s.m.
- agg.
- v
- s.f.
- adv.
- agg./s.m.
- s.m./agg.
- escl.
- agg./adv./s.m.
- s.f./pl.
- s.m./pl.
- indef.
- prep.
- conj.
- poss.
- escl./s.m.
- adv./prep.
- adv./indef.
- s.f./s.m.

Inserimento da XML, contesto

0.2 - Etimo

Da

0.2 - Note etimo

0.4 - Firma grafiche: per spazio, es. "incoib=, a=coib=di, a=coib=che")

Fig. 2. Masque de saisie

Un système lexicographique numérique doit maîtriser la gestion des données textuelles relatives aux sources et contribuer à l'intégration entre dictionnaire et corpus. Cependant, Pluto ne peut pas remplacer les logiciels sur lesquels les corpus ont été développés et pour lesquels ils sont souvent optimisés ou adaptés. Il faut considérer aussi qu'un projet lexicographique peut s'appuyer sur plusieurs sources d'information, ayant chacune ses exigences techniques particulières. Tout en maintenant son indépendance par rapport aux logiciels de gestion des corpus, Pluto peut contrôler les contextes d'occurrence des items lexicaux, grâce à la saisie dans la base de données d'extraits de textes comportant toutes les informations supplémentaires (métadonnées, références à des ressources externes, etc.).

Le lien entre plate-forme lexicographique et corpus est possible grâce à un système d'interconnexion entre les applications. Les dernières versions de GATTO et GattoWeb, les logiciels de gestion des corpus développés par l'OVI pour la rédaction du TLIO⁵ et pour d'autres projets⁶, offrent une fonction d'exportation des résultats en format XML pour Pluto, mais la compatibilité parfaite est possible, en modifiant les scripts d'importation, avec un large éventail de moteurs d'interrogation. Après l'importation, Pluto stocke le fichier en vue des futures consultations et en indexe le contenu. Le panneau permet de travailler sur la documentation dans la même interface de rédaction, sans qu'il y ait besoin de revenir au logiciel d'interrogation : le rédacteur peut consulter, gérer, modifier, annoter, interroger les contextes (avec un panneau de contrôle dédié) et lier les exemples aux signifiés avec un clic.

La nouvelle structuration des données offre au TLIO la possibilité de dynamiser les pages publiées sur la version antérieure du site du dictionnaire (par exemple tous les index vont devenir beaucoup plus interactifs) et de développer les potentialités de ses bases de données, avec de nombreux parcours de lecture. Dans la version bêta, par exemple, le lecteur peut sélectionner une langue, lister tous les étymons enregistrés pour cette langue, trier par nombre d'occurrences, et lister tous les mots du TLIO qui proviennent du même étymon.

5 Le Corpus TLIO (<http://tlioweb.ovi.cnr.it>) est le corpus sur lequel le dictionnaire est construit. Il se compose de 2324 textes, soit 22 567 996 mots, et permet de faire des recherches sur des formes ou des lemmes. Le Corpus OVI (<http://gattoweb.ovi.cnr.it>) contient la collection complète de textes italiens anciens que l'OVI rend accessibles à la consultation publique. Il permet de faire des recherches sur des formes dans une base de données comportant 2446 textes pour un total de 23 874 376 mots.

6 Voir, par exemple, le Corpus DiVo, qui contient les traductions vernaculaires du Moyen Âge (<http://divoweb.ovi.cnr.it/>), le Corpus LirIO (<http://lirioweb.ovi.cnr.it/>) une base de données consacrée à la lyrique, ou le Corpus ARTESIA (<http://artesia.ovi.cnr.it/>) qui contient une collection de textes en sicilien vernaculaire, du premier document du XIV^e siècle jusqu'au milieu du XVI^e siècle.

La nouvelle architecture des informations a favorisé la mise en place de fonctions d'interrogation améliorées, ainsi que de nouvelles possibilités d'agrégation des résultats. Par exemple, dans Pluto la recherche par définitions, précédemment confiée à un seul champ de recherche, dispose maintenant d'une page de recherche détaillée où, parmi les nouvelles fonctions, se distinguent la recherche par date (rendue possible par l'intégration de la bibliographie citée) et la recherche par niveau hiérarchique de la définition (ce qui permet d'exclure les sous-définitions à partir d'un niveau déterminé, ceci vers le bas).

En plus de fournir un accès plus facile et plus rapide aux données figurant aussi dans un dictionnaire papier, un logiciel lexicographique devrait être capable de les analyser, de les combiner et de les regrouper à l'aide d'algorithmes pour extraire des informations implicites et générer de nouvelles connaissances. La version bêta de Pluto a commencé à produire de nouvelles informations et à enrichir les entrées. À l'aide de procédures d'analyse simples, la nouvelle page de consultation des entrées affiche une icône (signe «+») qui permet d'accéder à des informations supplémentaires qui sont élaborées et collectées par le logiciel. On a déjà évoqué précédemment la possibilité de lister tous les lemmes qui dérivent du même étymon. De même, Pluto peut signaler les différents articles dans lesquels sont classées des formes homographes. La récupération d'informations implicites est une expérimentation en cours : l'augmentation des données importées dans la plate-forme et l'intégration progressive de nouvelles informations dans la base de données permettront de définir d'autres algorithmes susceptibles d'enrichir la documentation fournie au lecteur.

La conception s'est concentrée sur les possibilités de réutiliser des logiciels et des données dans d'autres projets lexicographiques et de traitement de texte. Pluto est un système facilement adaptable à différents domaines d'application. Le développement de la première version du programme a déjà permis la mise en œuvre parallèle de PlutoVD, une version modifiée développée pour la préparation du Vocabolario Dantesco, un dictionnaire sur l'œuvre de Dante, projet commun de l'OVI et de l'Accademia della Crusca⁷.

La plate-forme est aussi équipée de modules spécifiques permettant de gérer l'échange de données entre systèmes utilisant Lexicad. Par conséquent, pour le démarrage d'un nouveau projet, l'adoption de systèmes Lexicad / Pluto et de

7 <http://www.vocabolariodantesco.it>.

The screenshot shows the TLIO (Treccani Online) interface. At the top, there is a search bar and navigation links: "Indice", "Ricerche", "Tutto sul TLIO", and "Cerca". The main heading is "abitativa s.f.". Below this, there is a navigation menu with options: "Lista forme", "Nota etim.", "Prona am.", "Distrib. geogr.", "Note ling.", "Note", "Lista derivazioni", "Riduzione", and "Tutti / Stampa". The "Nota etim." tab is selected. The content area displays the etymology of "abitativa", starting with "0.2 Da abitare I." and "Etim. marcati in questa voce: abitare [1] [derivazione da voci del TLIO (o comunque volgari)].". It lists other forms like "abitamento s.m.", "abitante s.m.", "abitante² agg.", "abitare² s.m.", "abitazione s.f.", "abitato² agg.", "abbate² s.m.", "abitato³ agg.", "abitatore s.m.", "abitazione s.f.", "abitabile agg.", "abitata s.f.", and "abitare s.m.". A section titled "Tabella degli etimi" is also visible. At the bottom, there is a reference to "1 La capacità di dare abitazione a qno." and a footnote [1] citing Jacopo Alighieri's "Dottrinale" (1340).

Fig. 3. Interface utilisateur

GATTO / GattoWeb offre non seulement une suite de logiciels, mais également l'accès direct à une base de données de départ déjà prêtes. Ces données peuvent être transférées très facilement d'un dictionnaire à un autre et ce sera Pluto lui-même qui se chargera de la sélection et du traitement avant le transfert.

Parallèlement aux opérations traditionnelles d'importation et d'exportation des informations, deux systèmes peuvent partager facilement des informations en temps réel, sans dupliquer les données. Le paradigme conceptuel de l'informatique en nuage, qui a favorisé la fourniture de ressources informatiques distantes, que ce soient des ressources hardware, logicielles ou des données, a encouragé certaines initiatives lexicographiques à rendre disponibles des fonctionnalités ou des données, au moyen d'une série d'API (acronyme d'Application Programming Interface) qui permettent à d'autres développeurs d'utiliser leurs contenus et leurs procédures de traitement des données selon des formules gratuites ou payantes⁸. Pluto développe le potentiel du dictionnaire en tant que service et dispose de fonctionnalités dédiées au dialogue avec d'autres applications, dûment autorisées, avec lesquelles il peut échanger les données. Un autre dictionnaire pourra confier au TLIO la tâche de remplir, automatiquement et

8 Voir, par exemple, les API d'Oxford University Press (<https://developer.oxforddictionaries.com/>).

toujours avec des contenus actualisés, les points de ses entrées non directement liés à ses objectifs principaux. De la même manière, à l'aide d'un simple script, tout texte au format électronique peut afficher, pour chaque mot, les définitions correspondantes du TLIO, en les obtenant en temps réel grâce aux services offerts par Pluto.

La mise à jour et la migration de données du TLIO sont actuellement en cours. L'importation implique une phase de conversion, avec un analyseur développé par Andrea Boccellari et Domenico Iorio-Fili⁹, et une phase d'importation d'un fichier XML. Pendant l'importation, Pluto analyse les données, vérifie leur exactitude formelle, et tente d'extraire de nouvelles informations par inférence. Le processus est presque complètement automatisé, sauf pour certains points dans lesquels l'analyse algorithmique ne peut pas remplacer l'interprétation humaine. Au fur et à mesure que la quantité d'informations introduites dans le système augmente, de nouvelles perspectives d'analyse et de recherche s'ouvrent et apportent de nouvelles possibilités de synergie avec des projets similaires.

9 Andrea Boccellari s'occupe aujourd'hui de ce transfert.

Bibliographie

Arcidiacono, Salvatore (2019): «Dizionari e database lessicali on-line verso un orizzonte condiviso: modelli, pratiche e strumenti», in: Consiglio nazionale delle Ricerche (éd.): *Bollettino dell'Opera del Vocabolario Italiano* (2018), vol. 23, Alessandria: Edizioni dell'Orso, p. 369-378.

Beltrami, Pietro G. (1998): «Norme per la redazione del Tesoro della Lingua Italiana delle Origini», in: *Bollettino dell'Opera del Vocabolario Italiano*, III, p. 277–300 [<http://tlio.ovi.cnr.it/TLIO/>].

Boccellari, Andrea (2012): «Il sistema di redazione e pubblicazione web del TLIO», in: *Dizionari e ricerca filologica. Atti della Giornata di Studi in memoria di Valentina Pollidori, Firenze, Villa Reale di Castello, 26 ottobre 2010*, Alessandria: Edizioni dell'Orso, p. 57-64.

Boccellari, Andrea & Domenico Iorio-Fili (2013): «Il supporto dell'informatica al Vocabolario», in: Larson, Pär, Paolo Squillacioti & Giulio Vaccaro (éd.): «*Diverse voci fanno dolci note*». *L'opera del Vocabolario Italiano per Pietro G. Beltrami*, Alessandria: Edizioni dell'Orso, p. 15-30.

Martin, Robert (2007): *Le Dictionnaire du Moyen Français (DMF) (1330–1500). Seconde version: DMF 2, Présentation* [<http://atilf.atilf.fr/gsouvey/dmf2/Pre-sentationDMF2.pdf>].

L'état de la numérisation du LEI : un rapport¹

Elton Prifti

Bien que les débuts du *Lessico Etimologico Italiano* (LEI), remontant aux années 1960, correspondent à peu près à la période des premières tentatives d'utilisation des technologies de l'information pour l'analyse linguistique, Max Pfister a naturellement conçu le LEI, tant au niveau méthodologique qu'opérationnel, en suivant l'illustre exemple du *Französisches Etymologisches Wörterbuch* (FEW), qui était, jusqu'à la fin de sa première édition en 2002, une entreprise entièrement analogique. Compte tenu des progrès technologiques constants et consistants, ainsi que de l'expansion rapide de l'application de l'informatique dans les domaines les plus divers de la recherche scientifique, y compris la linguistique et surtout la lexicographie, il ne fallut pas longtemps pour que l'éventualité d'une utilisation des technologies informatiques pour la rédaction soit envisagée. En mai 1981 déjà, Max Pfister lui-même participa à l'une des premières conférences sur la linguistique informatique, organisée à Pise et consacrée à l'utilisation des ordinateurs dans la création et la publication de dictionnaires. Il y présenta pour la première fois quelques réflexions significatives sur la possibilité de créer des index lexicaux et morphologiques dans le cadre du LEI avec les outils numériques (Pfister 1983). Il y a près de quarante ans, Antonio Lupis et le physicien Flavio Waldner développèrent le programme Lexicon, « capace di creare concordanze automatiche e formattate secondo le esigenze, anche formali, degli schedari del LEI » (Lupis 2002 : 93). Ce programme fut présenté à deux reprises au moins (Lupis/Panunzio/Waldner 1984, cité dans Lupis 2012 : 126) mais, pour diverses raisons, il ne fut jamais adapté aux besoins du LEI, ni utilisé dans le processus de rédaction. Le volume dédié à Max Pfister pour son sixantième anniversaire (Glessgen/Holtus/Kramer 1992) contient un article consacré à la question de l'informatisation du LEI (Linciano 1992). Bien que de taille réduite, il est le seul de ce type à apparaître dans le volume et propose des idées et des éléments nouveaux, qui ne furent suivis qu'en partie. Pour Linciano (Linciano 1992 : 123), « l'obiettivo generale dell'intera opera di informatizzazione è di razionalizzare e rendere più efficienti i diversi momenti del lavoro

1 Une version italienne de cet article a été transmise en octobre 2018 aux rédacteurs des actes du Congrès international « Italiano antico, italiano plurale. Testi e lessico del Medioevo nel mondo diptale », organisé à Florence en 2018 à l'occasion des 40 000 premiers articles du TLIO. Le contenu présenté ici reflète les progrès réalisés jusqu'en octobre 2018.

di compilazione del Lessico, riservando ad un elaboratore opportunamente istruito le parti esecutive quali gestione degli archivi, trattamento delle informazioni di base, ecc. ». De plus, Linciano aspirait à « la creazione di una rete informatica, mediante la quale i collaboratori del LEI (tedeschi e italiani) potessero comunicare istantaneamente tra loro con la sede centrale di Saarbrücken accedendo, quindi, alla consultazione delle informazioni redazionali centralizzate e aggiornate in < tempo reale > » (Linciano 1992 : 124). Par la suite, et à plusieurs reprises, d'autres auteurs liés au LEI en tant que rédacteurs ou relecteurs ont exprimé leurs points de vue sur divers aspects de son informatisation, comme, par exemple, dans diverses publications consacrées à Max Pfister, surtout dans les actes du colloque organisé à l'occasion de son 70^e anniversaire, et publiés sous la direction de Wolfgang Schweickard (Schweickard 2006) ; cf. aussi Lupis 2012, Giuliani/Vinciguerra 2018 et autres.

1. Niveaux de numérisation

D'un point de vue général, il est possible d'identifier trois niveaux de numérisation, applicables au LEI, qui se distinguent les uns des autres par leurs objets et leurs objectifs, d'une part, et par la profondeur de la numérisation, d'autre part.

Le **premier niveau** consiste en l'utilisation de moyens numériques ayant pour objectif la préservation, le traitement et la consultation du matériel lexicographique. Il s'agit tout d'abord de la numérisation dite de rétrodigitalisation des articles publiés, produits de manière analogique, ainsi que du matériel lexicographique sous forme imprimée et, en particulier, des fiches. Le traitement des pages imprimées à l'aide de programmes de reconnaissance optique de caractères (OCR) permet d'utiliser le matériel plus efficacement, bien que dans une mesure limitée.

Le **deuxième niveau** correspond à la numérisation avec traitement informatique approfondi des données lexicographiques : autrement dit, la numérisation des articles publiés. Cela permet de rechercher, sur la base d'un nombre relativement restreint de critères prédéfinis, qui peuvent être combinés entre eux, des éléments individuels de contenu qui sont tirés du texte écrit des articles.

Le **troisième niveau** consiste en l’informatisation de la méthode de travail, c’est-à-dire du processus rédactionnel. Cela permet à la fois une numérisation des données de meilleure qualité et, surtout, une augmentation du rendement du travail rédactionnel.

Dans la plupart des projets de lexicographie historique numérique en cours, la numérisation se situe au deuxième niveau. L’évolution rapide des technologies de l’information au cours des dernières années ainsi que les nouvelles perspectives qu’elles ouvrent ont permis et rendu nécessaire un changement radical dans le processus de numérisation, qui n’est plus seulement axé sur la présentation des données, mais aussi – et surtout – sur l’optimisation de la méthode de travail. Il en résulte une amélioration immédiate et substantielle de la qualité, puisqu’elle garantit l’informatisation généralisée du processus rédactionnel, ce qui génère également une longue série d’avantages techniques.

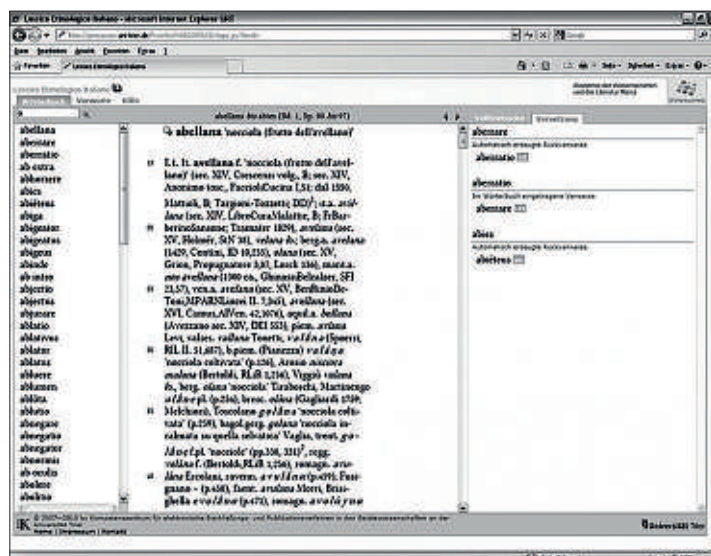
2. Notes sur les tentatives d’informatisation du LEI

Au cours des cinquante années d’histoire du LEI, plusieurs initiatives ont été prises en vue de l’informatisation, et ont été orientées vers des niveaux individuels de numérisation parmi ceux décrits ci-dessus. On peut en distinguer principalement trois, qui ont donné des résultats concrets. L’initiative la plus éloignée dans le temps et pourtant la plus moderne compte tenu de l’évolution des technologies de l’information à l’époque, est celle, principalement axée sur l’informatisation du processus rédactionnel, entreprise par Antonio Lupis dans les années 90. Le projet consistait en la création d’un environnement numérique unique contenant les entrées de la bibliographie du LEI précédemment créée par Michele Linciano (Linciano 1992), « disponibili per ulteriori trattamenti, come, ad esempio, il trasferimento automatico di ogni singolo «valore» (bibliografico, localizzativo, cronologico ecc.) da solo o in combinazione con altri, verso un *file* di testo o un altro archivio digitale » (Lupis 2012 : 127). Il s’agit du programme auxiliaire ItaCa (v. fig. 1), qui est encore partiellement utilisé aujourd’hui.

La deuxième initiative, entreprise il y a une quinzaine d’années, consistait à numériser les neuf premiers volumes du LEI (A-CALIDUS), c’est-à-dire à rétro-digitaliser la publication par une procédure de *double keying* et à procéder à son marquage XML (*Extensible Markup Language*). La réalisation a été confiée



Fig. 1. Capture d'écran du programme ItaCa src

Fig. 2. Capture d'écran de la page LEI-online
<https://kompetenzentrum.uni-trier.de>

par la commission au Kompetenzzentrum – aujourd’hui Center for Digital Humanities – de l’Université de Trèves. En raison de la qualité peu satisfaisante du résultat, le projet a été abandonné et la version en ligne du LEI (v. fig. 2) a été retirée d’internet.

L’initiative la plus récente est en cours et a débuté il y a environ trois ans. Elle est maintenant menée par une équipe d’une dizaine de personnes, comprenant des informaticiens et des rédacteurs, et consiste à la fois en la numérisation des articles publiés et en l’informatisation des processus rédactionnels. Quelques informations générales sont présentées ci-dessous.

3. État actuel de la numérisation

Le projet de numérisation du LEI s’articule actuellement autour de cinq axes principaux :

1. La numérisation capillaire et complète des parties écrites de façon analogique, publiées ou non (*A-E, Germanismi*) ;
2. Le développement d’un système rédactionnel entièrement informatisé, capable d’accélérer considérablement le processus rédactionnel, notamment par l’automatisation des différentes opérations ;
3. La rédaction de nouveaux articles (à partir de la lettre G) de manière entièrement numérique, en utilisant le nouveau système rédactionnel ;
4. La mise à jour et la correction automatisée non seulement des sigles, mais aussi des parties écrites de manière analogique ;
5. La publication en ligne du LEI, accompagnée de multiples fonctions de recherche, et, concernant les articles écrits de manière entièrement numérique, enrichie par la visualisation informatisée des contextes ainsi que des fiches rétrodigitalisées.

Initialement, l'objectif était double : le premier aspect, et le plus important, était de rétrodigitaliser l'ensemble du fichier du LEI, après un contrôle et un premier arrangement manuel du matériel. Le deuxième objectif, beaucoup moins exigeant, était de rétronumériser et de traiter par OCR les différents articles déjà publiés par le LEI.

Au cours de la phase de rétrodigitalisation (cf. 3.1), ce nouveau projet de grande ampleur de numérisation du LEI a progressivement pris forme, en devenant la solution la plus adéquate pour achever le projet LEI dans les délais impartis. Cela permettra non seulement de maintenir – et, au moins en partie, peut-être même d'améliorer – la qualité, mais aussi de moderniser le LEI, en le rendant plus facilement accessible, actualisable et apte à être corrigé. Il est également important de continuer à investir dans l'élargissement et la formation de l'équipe du LEI de même que dans l'extension de l'infrastructure rédactionnelle du projet. Dans cette optique, il a été possible de consolider progressivement les nouveaux centres du LEI à l'Université de Mannheim (depuis février 2015), à l'Université de Vienne (depuis novembre 2018) et à l'Université pour étrangers de Sienne (à partir de janvier 2019).

3.1 Rétrodigitalisation

Le processus de numérisation du matériel papier (tant le fichier que les parties déjà publiées du LEI) s'est déroulé entre février 2015 et juin 2017 à l'Université de Mannheim grâce à une collaboration étroite et particulièrement fructueuse avec l'Université pour étrangers de Sienne. Les documents ont été numérisés en haute résolution et en couleurs, et environ 7,5 millions de fiches ont été contrôlées, puis regroupées par étymon en environ 20 000 documents PDF disponibles, avec les articles déjà publiés, sur un portail (v. fig. 3) dont l'accès est réservé à l'équipe du LEI. Le matériel rétrodigitalisé remplit plusieurs fonctions importantes : en plus d'être une copie de sauvegarde de l'ensemble du fichier, il constitue la base du traitement partiellement automatisé des fiches pour les articles de la partie F-Z du LEI, traitement qui est brièvement présenté ci-dessous (3.3.2).

| lei.romphil.de/ - [GFsearch=ge](#)

Lessico Etimologico Italiano

DIGITALE

A (pubbl.) B (pubbl.) C (pubbl.) D (pubbl.) E (pubbl.) GERM. (pubbl.)

A B C D E F G H I L M N O P Q R S T U V Z ANGL. GERM. SPAN.

gange	gemma	genista	geniura	gerae
ge-	gemmula	genialis	genus	gerae
geb-	gemo	genitivus	geo-	gerius
gebula	gemonium	genitor	geographia	gerulus
geg	gemulus	genitu	geomantia	gerundium
geg-	gen-	genitura	geometres	gerwo
gegg	gena	genitus	georgius	ges
gegemar	genarilis	genius	gesp-	gesk
gehonna	genea	genna	ger-	gesse
gei	generabilis	gennoo	geranion	gesta
getala	generalia	gennoos	gerbe	gestare
getare	generalitas	genova	gerbh	gestallo
getala	generare	gens	gerbidu	gesticulari
getalidium	generatio	gentiana	gerdius	gesticulatio
getiu	generosus	gentile	geremia	gestio
getja	generu	gentilis	gerenie	gestire
getolophyllis	genesis	gentius	gerene	gestor

[CERCA](#) | [Torna](#) | [Novità](#) | [Sul LEI](#) | [Volumi](#) | [Bibliografia](#) | [Elenco](#) | [Forè](#) | [TUD](#) | [OVI](#) | [Immagini](#) | [Cronaca...](#) | [prilija@uni-mannheim.de](#)

Fig. 3. Capture d'écran de l'archive numérique du LEI <http://lei.romphil.de>

3.2 Numérisation des pièces publiées

À ce jour (mars 2019), 5098 articles ont été publiés en format papier, soit un total de plus de 15 000 pages imprimées au format A4. La publication en ligne de ce matériel, offrant de multiples fonctions de recherche, assure une consultation détaillée, fonctionnelle et efficace. Elle constitue également l'un des principaux objectifs du *LEI digitale*. Le traitement numérique des pièces en question est prioritaire et consiste en la transformation largement automatisée du contenu des articles en une base de données relationnelle, dont la structure est identique à celle des pièces à produire dans le futur de manière entièrement numérique (cf. 3.3.1).

Les défis à relever, tant sur le plan technologique que sur celui du contenu, sont nombreux. Les deux tiers du matériel de base environ consistent en images de pages imprimées, qui doivent être transformées en texte calquant le formatage du texte original, lequel servira pour le balisage automatique ultérieur. Bien que la systématisation graphophonétique correcte du patrimoine documentaire dialectal varié, avec des graphèmes accompagnés de combinaisons de diacritiques, reste compliquée, elle demeure néanmoins réalisable au niveau de la représentation. En ce qui concerne le contenu, il suffit de mentionner, à titre d'exemple, la gestion des mises à jour – quoique sporadiques – qui ont eu lieu au fil des décennies dans les sigles et la datation des sources.

3.3 Informatisation et réorganisation du système rédactionnel

Afin d'atteindre les objectifs mentionnés ci-dessus, il est essentiel de procéder à une informatisation approfondie et poussée aussi bien des données que de la longue chaîne éditoriale. Il est important de noter que le système rédactionnel informatisé du LEI ne peut être emprunté à d'autres projets de lexicographie historique en raison non seulement de ses grandes dimensions, mais aussi de la complexité de ses processus rédactionnels et de ses caractéristiques spécifiques.

Il est nécessaire de concevoir un système parfaitement adapté aux besoins rédactionnels du LEI, autrement dit de lui « tailler un costume sur mesure ». La conception et le développement d'un tel système rédactionnel ont débuté en novembre 2016, après l'examen et l'évaluation en profondeur des systèmes rédactionnels de plusieurs autres projets de lexicographie historique, utilisant à des degrés divers des méthodes et des outils informatiques spécifiques. La phase finale du projet, principalement axée sur des questions techniques, a été réalisée en coopération avec des spécialistes en informatique. En 2018, le noyau du groupe d'informaticiens était composé, par ordre alphabétique, de Frank Dopatka, spécialiste du *web-development* et de *game-engineering*, professeur à la faculté d'informatique de la Hochschule Mannheim ; de Bernd Freisleben, professeur d'informatique au département de mathématiques et d'informatique de la Philipps-Universität de Marburg, spécialisé dans le développement de méthodes et d'approches informatiques liées aux applications ; et d'Oliver Hummel, spécialiste en *big data management* et en *software-engineering*, professeur à la faculté d'informatique de la Hochschule Mannheim. Cette équipe, soutenue

par ses propres collaborateurs et par des membres de l'équipe du LEI, identifie les meilleures solutions et met en œuvre l'infrastructure informatique du projet de numérisation. Nous présentons ci-dessous un bref aperçu des principes de base et des solutions, mais aussi des principales applications du *LEI digitale*.

3.3.1 Principes

Le système rédactionnel numérique du LEI se base sur un modèle relationnel de données, ce qui signifie que l'accent est mis sur les « objets », et plus précisément sur les graphes, et non sur certains éléments du texte, comme dans le cas des autres projets de lexicographie historique numérique. L'élément de base d'un article du LEI est constitué par l'ensemble des informations sémantiques, formelles et bibliographiques, qui sont hiérarchiquement liées les unes aux autres et réunies dans une chaîne, qui, en informatique, représente un graphe. C'est la composition de ces informations qui détermine la structuration de la base de données relationnelle du LEI. La rédaction d'un nouvel article utilisant le système rédactionnel numérique et la numérisation d'un article déjà écrit selon la méthode analogique sont toutes deux basées sur le même principe. La seule différence structurelle entre les graphiques produits par les deux méthodes d'édition consiste en l'accompagnement du graphe par le contexte respectif, s'il provient des bases de données GDLI, TLIO et/ou OVI, ou par l'image de la fiche, si le graphe est issu du traitement de cette dernière.

Un article du LEI consiste donc en des graphes hiérarchiquement interreliés au sein de groupes de graphiques, qui à leur tour sont également interreliés. Cette dernière corrélation constitue également la structure de l'article. À la fin du processus rédactionnel, le commentaire de l'article est traité de manière presque entièrement manuelle.

L'automatisation de la production de graphes permet d'introduire – dans les conditions actuelles – deux nouveaux profils de collaborateurs au LEI : l'assistant de rédaction et le prérédacteur. La tâche principale de l'assistant de rédaction est de produire des graphes incomplets à l'aide d'applications spéciales développées spécifiquement pour le LEI et, sous la direction du rédacteur responsable, de traiter les principales sources, qu'il s'agisse du GRADIT, du GDLI, du TLIO, de l'OVI ou des fiches. La poursuite de la réalisation, en particulier au niveau sémantique, ain-

si que la vérification et la correction éventuelle des graphes sont effectuées par le prérédacteur. Le matériel est ensuite pris en charge par le rédacteur responsable, dont la tâche est de terminer l'article. Il intervient sur les graphes, en vérifiant et corrigeant les données, en créant de nouveaux graphes sur la base d'autres sources spécifiques, liées à l'article unique, et définit surtout la corrélation entre tous les graphes et les groupes de graphes, structurant ainsi l'article. Enfin, il écrit le commentaire correspondant.

Selon les premiers sondages réalisés, ce mode d'organisation du travail devrait au moins permettre de doubler la productivité de chaque rédacteur. Une telle conception du processus rédactionnel permet de réaliser le travail d'abord pour la préparation partielle de l'ensemble des articles d'une lettre, puis de passer à la dernière étape, centrée sur chaque article de la lettre, au cours de laquelle la rédaction des différents articles sera terminée.

3.3.2 Principaux outils

Pour que le processus de rédaction informatisé du LEI soit réellement efficace et puisque le LEI est particulièrement complexe, il est nécessaire de recourir à des applications multiples, capables d'automatiser, en totalité ou en partie, les différentes étapes rédactionnelles. Vingt-six applications de volume et de complexité variables ont été conçues, dessinées, et ont déjà en partie été réalisées à ce jour. Les quatre plus importantes sont présentées ci-dessous.

L'application *Bibliografia generale detagliata e commentata* (BiG) est utilisée pour compléter de manière entièrement ou partiellement automatisée les coordonnées bibliographiques, locales et chronologiques de chaque graphe. L'application est basée sur une large base de données, construite à partir des bases de données ItaCa et de diverses autres sources.

La première étape du traitement des fiches consiste en la transformation entièrement automatisée du sigle en informations bibliographiques, géographiques et chronologiques contenues dans le graphe correspondant. L'application *Elaborazione automatizzata delle informazioni areali, bibliografiche e cronologiche delle schede* (ABC) est basée sur une logique *deep learning* et interagit avec l'application BiG.

L'*Interfaccia redazionale online* (IReO) correspond à la plate-forme de travail centrale dont l'accès est limité aux seuls membres de l'équipe du LEI et sur laquelle s'effectue le traitement des graphes, leur corrélation au sein de groupes de graphes et la connexion de ces derniers dans l'article, la rédaction de l'article, ainsi que la communication entre les collaborateurs, tant au sein des différents groupes de rédaction responsables des lettres individuelles qu'au sein de toute l'équipe du LEI. Les utilisateurs d'IReO ont la possibilité d'utiliser plusieurs fonctions de recherche dans la base de données du LEI.

La *Piattaforma pubblica online* du LEI permettra de consulter les articles publiés. Elle sera dotée de diverses options, parmi lesquelles les différents degrés de détail dans la représentation du contenu des articles individuels, les différentes fonctions de recherche, ainsi que la transformation du contenu de l'article en format imprimé classique du *LEI analogico*. Le produit final du *LEI digitale* sera exactement identique au *LEI analogico* d'un point de vue visuel, cependant il sera réalisé techniquement d'une manière totalement différente et contiendra toute une série d'informations supplémentaires et très utiles.

Il y a huit ans, Antonio Lupis écrivait : « L'apporto dell'informatica è (...) oggi decisivo in ogni impresa lessicografica. Eppure alcune iniziative (...) sono per così dire dubitosamente a metà del guado, impossibilitate per loro natura a cambiare il passato, e dal passato troppo appesantite per modificare vantaggiosamente il futuro. Una condizione simile è in una certa misura avvertibile anche nel LEI » (Lupis 2012 : 124). Les perspectives qu'offrent aujourd'hui l'informatique et les supports numériques à notre disposition, au même titre que le nouveau principe de base de la numérisation décrit ci-dessus, permettent non seulement de faire passer le LEI dans un monde entièrement numérique, mais aussi d'accélérer le processus rédactionnel, les mises à jour et l'interaction du LEI avec d'autres projets, en maintenant une qualité élevée et en respectant et préservant son esprit, comme l'aurait voulu son illustre créateur, Max Pfister.

Bibliographie

FEW = Wartburg, Walther von et al. (éd.) (1922–2002): *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes*, 25 vols., Bonn et al. : Klopp et al.

GDLI = Battaglia, Salvatore (éd.) (1961–2002): *Grande dizionario della lingua italiana*, 21 vols., Torino: UTET.

Giuliani, Mariafrancesca & Antonio Vinciguerra (2018): «Conservare innovando: per una riflessione critica sulle architetture e sul repertorio del LEI al servizio dell'indagine geo- e storico-linguistica», in: Retali-Medori, Stella (éd.): *Actes du colloque de lexicographie dialectale et étymologique en l'honneur de Francesco Domenico Falcucci (Corte – Rogliano 28-30 Octobre 2015)*, Alessandria: Edizioni dell'Orso, p. 293–310.

Glessgen, Martin-Dietrich, Günter Holtus & Johannes Kramer (éd.) (1992): *Etymologie und Wortgeschichte des Italienischen. LEI. Genesi e dimensioni di un vocabolario etimologico*, Wiesbaden: Dr. Ludwig Reichert.

GRADIT = De Mauro, Tullio (éd.) (1999–2007): *Grande dizionario italiano dell'uso*, 8 vols., Torino: UTET.

LEI = Prifti, Elton & Wolfgang Schweickard (éd.) (1979-): *LEI. Lessico Etimologico Italiano. Fondato da Max Pfister*, Wiesbaden: Reichert Verlag.

Linciano, Michele (1992): «Il supporto dell'elaboratore per la bibliografia e per la redazione degli articoli, ossia il LEI e gli Utensili Informatici (= L.U.I.)», in: Glessgen, Martin-Dietrich, Günter Holtus & Johannes Kramer (éds.): *Etymologie und Wortgeschichte des Italienischen. LEI. Genesi e dimensioni di un vocabolario etimologico*, Wiesbaden: Dr. Ludwig Reichert, p. 123–130.

Lupis, Antonio (2002): «Vent'anni dopo. Il romanzo del LEI di Max Pfister alla lente della storia e dell'avanzamento tecnologico, con qualche proposta di giunte ai dizionari storici italiani», in: Glessgen, Martin-Dietrich, Wolfgang Schweickard, Günter Holtus & Johannes Kramer (éd.): *Ex traditione innovatio. Miscellanea in honorem Max Pfister septuagenarii oblata*, Darmstadt: Wissenschaftliche Buchgesellschaft, 2, p. 91–101.

Lupis, Antonio (2012): «Trent'anni dopo (e vent'anni prima): due nuovi approdi digitali per la barca del LEI», in: Lubello, Sergio & Wolfgang Schweickard (éd.): *Le nuove frontiere del LEI. Miscellanea di studi in onore di Max Pfister in occasione del suo 80° compleanno*, Wiesbaden: Reichert-Verlag, p. 125–146.

Lupis, Antonio, Saverio Panunzio & Flavio Waldner (1984): «Il progetto Lexicon: metodi e prospettive», in: AA. VV., *Scienze del linguaggio e insegnamento delle lingue e delle letterature (Atti del Convegno Internazionale di Bari, maggio 1982)*, Bari: Adriatica, 2, p. 197–218.

OVI = *Corpus OVI dell'italiano antico*, Firenze: Opera del Vocabolario Italiano. [<http://gattoweb.ovi.cnr.it>].

Pfister, Max (1983): «Présentation du LEI (Lessico Etimologico Italiano). Possibilités d'établir des index lexicaux et morphologiques par ordinateur», in: Zampolli, Antonio & Amedeo Cappelli (éd.): *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries. Proceedings of the European Science Foundation Workshop (Pisa, May 1981)*, Pisa: Giardini, p.187–199.

Schweickard, Wolfgang (éd.) (2006): *Nuovi media e lessicografia storica. Atti del colloquio in occasione del settantesimo compleanno di Max Pfister*, Tübingen: Niemeyer.

TLIO = *Tesoro della Lingua Italiana delle Origini* (1988-): Firenze: Opera del Vocabolario Italiano [<http://tlio.ovi.cnr.it/TLIO/>].

De Frantext 1 à Frantext 2 : la cure de jouvence d'une vieille dame

Véronique Montémont

La question de la production d'un grand dictionnaire de la langue française du XX^e siècle s'est posée avec force en France à la fin des années 1950. Elle a été débattue en novembre 1957, lors d'une conférence de Paul Imbs, professeur à la Faculté des lettres et directeur du Centre de philologie romane de Strasbourg, à l'occasion du colloque « Lexicologie et lexicographie françaises et romanes ». L'alternative qui s'offrait était la suivante : fallait-il moderniser le Littré, dictionnaire de la langue du XIX^e siècle tombé dans le domaine public, ou au contraire recommencer sur de nouveaux frais pour offrir un « exemple-type de lexicographie scientifique moderne » (Imbs 1961 : 285) ? C'est la deuxième solution qui a été retenue, et c'est elle qui a constitué le point de départ de l'aventure du *Trésor de la Langue Française*. Il faut rappeler que le contexte de l'époque invitait à ce type d'ambition intellectuelle : la cybernétique, qui essaimait aux États-Unis depuis la fin des années cinquante, avait familiarisé le monde scientifique avec l'idée de machines pouvant servir de relais à l'intelligence humaine, et jeté les bases de ce qui allait devenir l'informatique. Le structuralisme émergent, de son côté, préparait le terrain du rapprochement entre les sciences humaines et des méthodes empruntant la rigueur et le formalisme des sciences exactes. Le projet du futur *Trésor de la Langue Française* présentait une parfaite convergence avec ces manières nouvelles d'envisager le savoir. Ses objectifs, définis par Paul Imbs, étaient multiples : bâtir un dictionnaire de référence du français, le doter d'une dimension historique (chaque mot bénéficierait de sa rubrique étymologique) et linguistique (avec une description du mot en contexte). Un corpus d'exemples particulièrement riche viendrait étayer les définitions, et permettrait d'offrir une analyse concrète des usages du mot dans la langue¹.

Dans les faits, la réalisation du dictionnaire s'est traduite par la création d'un centre de recherche, installé à Nancy en 1960, qui deviendra rapidement l'INaLF (Institut national de la langue française) – aujourd'hui ATILF (Analyse et Traitement informatique de la Langue française). Compte tenu de la taille du

1 Sur la genèse du dictionnaire, voir Buchi & Pierrel 2009.

projet, il est d'emblée décidé de constituer à partir de 1965 une base de données informatisée de textes français, une idée tout à fait pionnière pour l'époque. Elle sera composée d'un noyau de mille textes, comportant pour 80 % des textes littéraires de création, principalement du XX^e siècle, et pour les 20 % restants de textes dits « scientifiques et techniques », une catégorie qui regroupe aussi bien des ouvrages d'économie que des traités de cynégétique ou d'œnologie. La sélection est opérée par des scientifiques et des documentalistes ; les textes sont ensuite « mécanographiés », puis reportés sur des cartes perforées, du moins durant les premières années, où les dispositifs de reconnaissance optique des caractères n'ont pas encore été inventés. Par la suite, les textes seront directement scannés et feront l'objet d'un balisage en XML-TEI. Ce corpus informatisé doit permettre d'établir une nomenclature, notamment grâce au calcul des fréquences des mots. Certains termes considérés comme rares, voire réputés à emploi unique (les *hapax*), peuvent en effet se révéler plus représentés qu'on ne pensait ; d'autres seront en revanche écartés en raison de fréquences trop faibles ou marginales. En plus du calcul de fréquences, absolues et relatives, de l'examen de la répartition du vocabulaire dans les corpus respectivement littéraires, scientifiques et techniques, cette base informatique autorise l'extraction rapide et facile de contextes ainsi qu'un examen des exemples en situation. Cette donnée est essentielle à la qualité et à la réussite du dictionnaire : la multiplicité des exemples – plus de 430 000 figurent dans le TLF – autorise une approche lexicographique plus réaliste, proche de l'usage, et encourage l'élaboration de définitions exhaustives où tous les sens, même les plus marginaux, seront représentés.

Frantext canal historique

La base qui contient ces textes n'a d'abord aucune existence autonome : elle n'est qu'un produit dérivé du dictionnaire, destiné à être consulté en interne par les collaborateurs du TLF. En 1984, une première tentative d'autonomisation a lieu, sous la forme de la production d'un CD-Rom, Discotext : ce support englobe un certain nombre de textes de la base, ainsi qu'un moteur de recherche, élaboré par Jacques Dendien, l'informaticien du TLF. Celui-ci permet de fouiller le corpus, mais non de le lire intégralement. À cette époque, les problématiques juridiques

liées aux textes numérisés sont moins sensibles et l'on peut encore se permettre de diffuser une telle ressource, qui n'affiche, pour ses utilisateurs, que des extraits d'environ 350 signes, dans le cadre du droit de citation.

En 1994, la parution du dernier volume du TLF signe la fin d'un vaste chantier lexicographique qui aura duré près de trente ans : le corpus compte alors environ 2000 items de plus qu'à sa création, soit 3000 références en tout. Il s'est en effet enrichi au fil des années d'œuvres contemporaines, pour refléter autant que possible l'état actuel de la langue, mais il a également été complété, en parallèle, avec les versions numérisées de textes réalisées dans le cadre d'autres projets de recherche, notamment sur le théâtre du XVIII^e siècle. Se pose alors la question de l'avenir de la base : certes, elle a perdu sa raison d'être historique, mais elle constitue un outil de travail particulièrement précieux pour les lexicographes, qui semble justifier une diffusion plus large au sein de la communauté scientifique. C'est pourquoi lorsque le TLFi, à la fin des années 1990, fait l'objet d'une rétroconversion destinée à le rendre disponible en ligne, il est décidé que la base bénéficiera du même mode de diffusion. Dans son cas, l'opération est favorisée par le fait que les données numériques existent déjà, tout comme l'outil de recherche² qui sert à les explorer.

Rebaptisée Frantext, la base est ouverte et mise en ligne en 1998, toujours sous l'égide du CNRS, avec pour principe de proposer un corpus couplé à un moteur de recherches, ainsi que l'affichage d'extraits en contexte. Il ne s'agit pas d'une liseuse (il est impossible de lire plus de quatre lignes consécutives), ni d'une plateforme de téléchargement de textes libres de droits, à l'instar de Gutenberg : le fait que le corpus est centré sur des textes qui se trouvent pour la plupart sous droits d'auteur n'encourage de toute façon pas cette utilisation. De plus, dans le souci compréhensible de protéger leurs données, en tout cas pour les textes modernes, les éditeurs édictent plusieurs conditions préalables à cette mise en ligne : un affichage d'extraits toujours aussi réduit en taille de contexte, un accès réservé à la communauté scientifique, un mot de passe obligatoire pour pouvoir identifier chaque utilisateur ou institution et un abonnement payant. L'objectif n'étant pas de faire du profit, la tarification demeure néanmoins accessible et raisonnable³. La base est ouverte et devient rapidement un outil de recherche fondamental dans la communauté linguistique, en France, mais aussi

2 Le moteur de recherche Stella, développé par Jacques Dendien.

3 À ce jour, il est de 370 euros annuels pour une institution et de 35 euros pour un particulier.

à l'étranger, où elle est relayée, en particulier en Amérique du Nord, par l'ARTFL (American and French Research on the Treasury of the French Language) à Chicago.

Rappelons qu'à l'époque, les autres grands réservoirs textuels sont en cours de développement ; soit ils sont fondés sur le bénévolat (Gutenberg), soit ils obéissent à une vocation patrimoniale tournée vers les textes anciens (Gallica), ou encore ils résultent d'initiatives privées, officiellement légales et gratuites, mais subsidiairement destinées à accroître la valeur d'entreprises cotées en bourse (Google Books). La base Frantext, service public qui couple textes récents, corpus échantillonné et outil de recherche performant garde donc un caractère unique, qu'elle n'a pas perdu à ce jour.

Une modernisation nécessaire

Frantext a donc d'abord été la petite sœur du TLF, dans l'ombre duquel elle a grandi ; dans le paysage encore balbutiant des humanités numériques de la fin des années 1990 et du début des années 2000, elle était perçue comme unique et sa mise en ligne (et en lumière) sur le web pouvait augurer de beaux développements. Mais la croissance de la petite sœur, peu à peu, a été négligée : au fur et à mesure que les années passaient, elle a subi un vieillissement accéléré qui a fait d'elle une vieille dame certes vénérable, mais dont les rides et les douleurs articulaires étaient de plus en plus visibles. Le fait était particulièrement perceptible dans ses interfaces vieillissantes et son ergonomie peu souple :



Fig. 1. Message d'erreur de Frantext 1

The screenshot shows a search interface with four main sections:

- Sélection par auteur**: A text input field for author selection.
- Sélection par titre**: A text input field for title selection.
- Sélection par date**: Radio buttons for 'Date indifférente', 'A la date précise DATE1', 'Avant DATE1', 'Après DATE1', and 'Entre DATE1 et DATE2'. Below are input fields for 'DATE1: 0' and 'DATE2: 2000'.
- Sélection par genre**: A grid of checkboxes for literary genres: correspondance, éloquence, mémoires, pamphlet, poésie, récit de voyage, roman, théâtre, traité, and essai.

Buttons at the bottom are 'ENREGISTRER LA SÉLECTION' and 'EFFACER LE FORMULAIRE'.

Fig. 2. Masque de saisie de Frantext 1

raire avait bien été paramétrée, mais la taxinomie sur laquelle elle était fondée laissait quelque peu à désirer, mettant sur le même plan « romans » (avec ses milliers d'items) et des catégories aux contours beaucoup plus flous comme « récits de voyage », quantitativement très peu représentées.

Mais ces problèmes de surface en masquaient d'autres plus graves: le traitement automatisé des textes et le balisage des textes, en particulier, étaient problématiques. Il est en effet assez rare de disposer d'une archive électronique vieille de presque cinquante ans, et cet étalement diachronique avait entraîné plusieurs solutions de continuité dans les normes utilisées pour le balisage en XML-TEI, un langage d'annotation destiné à rendre une ressource textuelle lisible et exploitable sur plusieurs types de support. Les premiers textes mécanographiés avaient ainsi utilisé des subterfuges, comme l'aposition d'une astérisque devant les noms propres, une technique abandonnée en cours de route. Les traitements de plusieurs caractères diacritiques, comme les traits d'union, tirets cadratins et semi-cadratins, souvent confondus, étaient discordants. Le balisage XML n'était pas irréprochable partout – et il ne l'est d'ailleurs toujours pas complètement, mais du moins a-t-on pu faire quelques campagnes de nettoyage.

Sélectionner un corpus, par exemple, ne laissait le choix qu'entre la sélection de toute l'œuvre d'un auteur ou bien un seul titre, sans possibilité intermédiaire. Une option de sélection d'entrée par genre littéraire

■ Liste des genres	
1	poèmes en prose
481	poésie
4	poésie didactique
4	presse
1	rapport
12	récit
2	récit autobiographique
55	récit de voyage
1	récit personnel
1	récits personnels
1238	roman
1	roman arthurien
1	roman autobiographique
1	roman fantastique
1	roman jeunesse
3	roman policier
680	théâtre
1	théologie
52	traité
730	traité ou essai

Fig. 3. Sélection par genres littéraires sur Frantext 1

Frantext utilisait par ailleurs des fonctionnalités de recherche spécifiques, dites « expressions de séquences » dont la syntaxe d'utilisation était fondée sur l'usage de l'esperluette (&m, &c, &l, &r). Ces fonctions permettaient de rechercher respectivement des lemmes (substantifs, verbes), des listes de mots, ou des « grammaires » – c'est-à-dire, dans la terminologie Frantext, des séquences de recherche combinées. On pouvait ainsi, dans le corpus, rechercher des paradigmes entiers – toutes les formes du verbe « aller » – ou établir des listes et les faire cooccurrer avec certains substantifs et des adjectifs. On pouvait aussi additionner certains ensembles, examiner leur voisinage, soustraire ou exclure les listes d'éléments en utilisant les opérateurs *et*, *ou*, *sauf*. Nous avons ainsi pu, par exemple, rechercher les traces du corps dans l'œuvre de Raymond Queneau en une seule passe, en mobilisant une « grammaire » (ou séquence de recherche)

```
|&mmenton |&mmiche |&mmoignon |&mmollet |&mmotte |&mmoustache |&mnaseau |&mmuscle |&mnaze |&mnnerf |&mnneurone |&mnnez |&mnichon |&mnombri |&mnnuque |&moeil | occiput | oigne | &momoplate | &mongle | &morgane | &morifice | &mpalais | &mgénital | &mpectoral | &mpeau | phalle | &mphysiologie | &mpied | &mpince | plexus | &mpoignet | &mpoing | &mpoil | &mpoitrine | &mpostérieur | postère | &mpore | pouatine | &mpouce | &mpoumon | &mpulmonaire | pulmoneux | pulmoniques | pulmoniales | pulmونيens | pubis | &mpudenda | &mrein | &mrotule | scrotum | &msein | &msexe | &msquelette | &msourcil | &mspermatozoïde | &mstomacal | tarin | &mtempe | &mtesticule | &mthorax | &mthoracique | thymus | &mtibia | &mthoracique | &mtrombine | &mtroufignon | &mtympan | &mlymphatique | &mventre | &mverge | &mverruqueux | &mvisage | &mviscère | &mviscéral | &mzygomatique
```

Fig. 4. Un extrait de la règle « organes » de la grammaire « Corps_Queneau »

qui comportait plus de 450 éléments. Divisée en six « règles » (autrement dit sous-ensembles), elle regroupait aussi bien des lemmes, des formes non fléchies, de diverses natures grammaticales, que des néologismes... Dans l'exemple qui suit, le sigle &m désigne toutes les flexions d'un substantif et d'un adjectif, &c toutes les flexions d'un verbe, la barre verticale « | » un « ou » inclusif. Il va sans dire qu'il fallait taper le tout à la main...

Les possibilités étaient variées, remarquables et par leur degré de granularité et par leur vitesse d'exécution. Mais elles requéraient une pratique intensive de l'outil, voire des formations dispensées directement par les personnels administrant la base au sein du laboratoire. Plusieurs universités se sont du reste résolues à rédiger des tutoriels volumineux à destination de leurs étudiants pour tenter de rendre Frantext plus accessible. Pendant ce temps, les autres outils de recherche en pleine émergence, notamment TXM, adoptaient des usages plus démocratiques : emploi des expressions régulières, que nous n'avions pu introduire qu'à la marge dans les améliorations de Frantext 1, ainsi que de la syntaxe CQL (Corpus Query Language) dont nous détaillerons les caractéristiques un peu plus bas. La poignée d'utilisateurs experts et enthousiastes des débuts cédait la place à une nouvelle génération de chercheurs, demandeurs de simplicité et d'efficacité dans des environnements graphiques plus conformes à l'esthétique contemporaine. Frantext devenait peu à peu un outil non standardisé, voire obsolète, dont la prise en main trop complexe, ajoutée à celle de l'accès sur abonnement à l'heure de la grande gratuité d'internet⁴, était jugée dissuasive.

Enfin, une opération d'étiquetage grammatical en parties de discours (l'autre nom en linguistique des classes grammaticales, comme celles des verbes, adjectifs ou pronoms) avait été réalisée sur 1900 textes de la base. Malheureusement, elle n'avait pas été poursuivie au fur et à mesure des enrichissements, et le départ à la retraite des concepteurs de cet outil d'étiquetage grammatical, Jacques Maucourt et Marc Papin, l'a de fait rendu inutilisable dès lors que plus personne n'était en mesure d'appliquer le traitement qu'ils avaient élaboré. Frantext hébergeait en conséquence deux bases, l'une catégorisée et l'autre non, ce qui rendait certaines recherches impossibles à mener dans l'ensemble du corpus. Certes, un fléchisseur, intégré au moteur de recherche, permettait bien de rechercher des formes *ou* des lemmes dans le Frantext dit « intégral » avec ses 3000 textes : par exemple un substantif au singulier *et* au pluriel, ou toutes les formes conjuguées d'un verbe donné. Mais on ne pouvait pas l'interroger par étiquette grammaticale, comme le font souvent les spécialistes de syntaxe, pour chercher une séquence du type : « Nom + Adjectif + Verbe » – et ce alors que l'opération était possible dans la partie de la base catégorisée.

4 Celle-ci s'est rapidement révélée irréaliste pour les textes sous droits, la musique ou la diffusion de tout autre bien culturel, sauf à entrer dans des logiques de piratage assumées.

Dans l'intervalle, d'autres répertoires, comme Wikisource et Google Livres, dans sa déclinaison européenne, se développaient ; en parallèle, de nombreux outils de textométrie, destinés à faire des recherches et des comptages dans un corpus de textes, étaient élaborés ou perfectionnés dans plusieurs laboratoires : Lexico 3 à Paris ou Hyperbase à Nice. On citera en particulier TXM, projet lancé à Lyon en 2007, dont l'ambition, atteinte aujourd'hui, était de fédérer les acquis des meilleures avancées en textométrie et de s'imposer comme l'outil de référence de la fouille de données textuelles en France. En cumulant les deux types de ressources, les textes numériques téléchargés sur certaines plateformes et des logiciels de fouille textuelle, gratuits ou payants, les chercheurs pouvaient donc désormais effectuer des recherches sur corpus sans passer par Frantext, qui avait de fait perdu sa position hégémonique.

D'autres difficultés, cette fois juridiques, sont venues ralentir le fonctionnement de la base : le milieu éditorial, effaré par l'exemple de l'industrie musicale, laquelle voyait son économie s'effondrer inexorablement sous les coups de boutoir du piratage digital, était de plus en plus réticent à autoriser une exploitation électronique de son fonds, aussi partielle, restreinte et protégée soit-elle. Seule la négociation d'un accord par le CNRS avec le Syndicat National de l'Édition dans les années 2010 a permis d'éviter la fermeture de Frantext.

Du neuf avec du vieux

Le départ à la retraite de l'informaticien créateur du moteur de recherche et de l'interface de Frantext, Jacques Dendien, a entraîné une redistribution des responsabilités. Gilles Souvay, ingénieur linguiste spécialisé dans le traitement des états anciens de la langue, et moi-même, enseignante-chercheuse en langue et littérature françaises, nous sommes partagé la gestion de la base : l'un était responsable du développement et des fonctionnalités, l'autre de la gestion patrimoniale et des enrichissements. Dans les faits, nous travaillions la plupart du temps de conserve. Ayant participé à de nombreux projets de recherche nationaux et internationaux, notamment le *Dictionnaire du Moyen Français*⁵, Gilles Souvay avait à sa disposition de nombreuses versions numériques de textes anciens. S'est alors posée la question de la diachronie et du cœur historique de Frantext : puisque le dictionnaire était terminé, nous n'avions plus d'obligation de maintenir son centre de gravité du côté de la langue du XX^e siècle, et plus

5 <http://www.atilf.fr/dmf/>.

marginalement du XIX^e siècle, qui constituait l'ordinaire des enrichissements. Au contraire, nous pouvions même envisager d'effectuer un « rattrapage » historique, du Moyen Âge au XVIII^e siècle, qui nous a permis de faire démarrer le corpus au X^e siècle, avec des textes qui présentaient la caractéristique non négligeable d'être libres de droits.

Par ailleurs, après des décennies d'enrichissements à la fois spécialisés et majoritairement tournés vers la littérature de création, la poésie et le théâtre, il nous a paru qu'il était temps de revenir soit vers des textes scientifiques et techniques, soit vers des textes au caractère linguistiquement plus varié, dont les auteurs n'étaient pas tous écrivains de métier. Nous nous sommes donc tournés, en partenariat avec l'équipe « Genèse et Autobiographie » de l'ITEM-CNRS, vers l'ajout d'écrits personnels, tels que les journaux personnels, témoignages, récits, autobiographies, correspondances, pour environ 500 titres. Au total, environ 2000 textes ont été ajoutés en dix ans.

Ce décalage temporel a eu un corollaire intéressant : les outils de recherche ont dû être adaptés à la période considérée. Le proto-Frantsert fondait ses requêtes sur la nomenclature du *Trésor de la Langue Française*, celle de la langue du XX^e

Base Frantsert intégral

Le corpus est échelonné du X^e au XXI^e et se répartit comme suit.

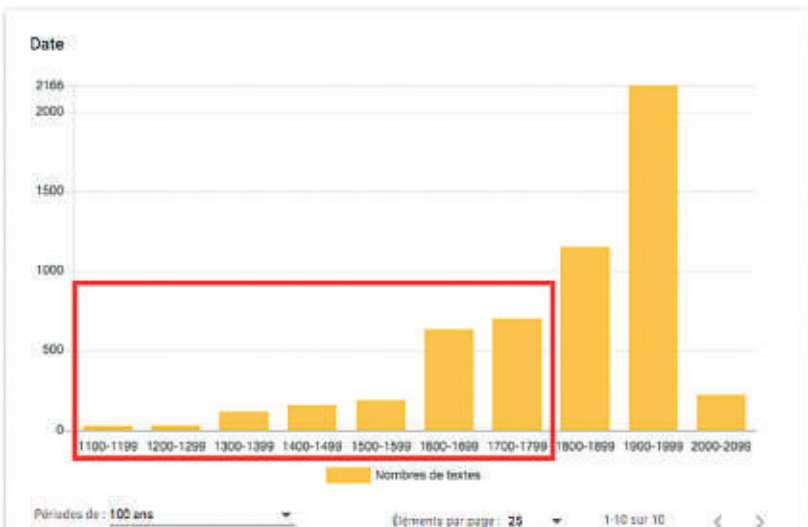


Fig. 5. Échelonnage du corpus

siècle, traitée par un fléchisseur appelé Morphalou. Il s'agit d'un programme qui crée des liens entre un lemme et ses formes : par exemple, le lemme *vert* était rattaché à quatre formes différentes, *vert*, *verte*, *verts*, *vertes*. Gilles Souvay a introduit d'autres lexiques dans Frantext, l'un médiéval, l'autre centré sur le français préclassique. Ces lexiques prennent en compte la variation graphique, en d'autres termes le fait qu'un mot, en fonction de l'époque à laquelle il est employé, tolère des orthographes différentes : ainsi, *enfant* peut aussi être orthographié au Moyen Âge *enffant*, et au pluriel *enffants* ou *enffanz*. Le lexique médiéval, en plus des formes connues de la langue moderne, ajoute donc l'ensemble des variations observées dans le corpus des textes anciens et mémorisées par un outil spécifique, LGeRM (Lexiques, Grammaires et Règles Morphologiques) ; celui-ci permet aussi d'archaïser des formes modernes pour arriver à une couverture de la langue de l'époque presque complète. Disposer de tels lexiques, paramétrables à chaque recherche, permettait donc une fouille des données beaucoup plus fine, absente des autres outils textométriques disponibles sur le marché.

Enfin, certaines fonctionnalités ont été redessinées et rendues plus ergonomiques, comme le formulaire de sélection de corpus : non seulement l'utilisateur pouvait cette fois sélectionner les œuvres en utilisant un jeu de descripteurs plus étendu, mais en plus il pouvait combiner autant de critères qu'il le souhaitait. Une fois constituée, la liste des items du corpus pouvait être téléchargée, puis réimportée à chaque session de consultation, de sorte à redémarrer sur une base de travail identique. Mais le corpus pouvait aussi être amendé à n'importe quel moment de la session de travail, resauvegardé, etc.

Corpus de travail

Formulaire Multicritères Auteurs Date Genre littéraire Corpus Mes corpus

■ Recherche dans les éléments bibliographiques

1 l'auteur : contient : [] liste + -

2 le titre : contient : [] liste + -

Options

- sensible à la casse
- sensible aux diacritiques
- sous-chaîne
- bibliographie détaillée

■ Corpus de travail

Nombre de textes : -

Nombre de mots : -

Sélectionner tous les textes

Lancer la recherche

Fig. 6. Formulaire de sélection du corpus amélioré de Frantext 1

Gilles Souvay a de même procédé à de multiples modifications d'ergonomie visant à simplifier l'usage de l'outil: remplacement de la saisie manuelle des fonctions &m ou &l par des boutons radio explicites (« rechercher un lemme », « rechercher une liste »), menus déroulants, onglets, aide présente sur chaque page, affichage KWIC, possibilité de mémoriser des corpus, des requêtes, des listes sur le serveur... Le problème de cette méthode de travail, qui, sur le plan de son fonctionnement, nous donnait toute satisfaction, est que nous étions contraints de procéder par ajouts et rustines ; l'informaticien devait intervenir sur un programme qu'il n'avait pas écrit lui-même, et qui lui aussi devenait doucement obsolète en termes de langage de programmation, nous faisant craindre à chaque manipulation une panne irrémédiable. Nous faisons, sommes toute, de

la chirurgie esthétique sur notre bonne vieille base, une situation qui ne pouvait perdurer. Le changement, rendu obligatoire pour diverses raisons, du moteur de recherches historique de la base, Stella, a donc entraîné la refonte en profondeur de la base Frantext : un chantier de longue haleine, qui a duré presque dix ans.

La mue

La version 2.0. de Frantext a vu le jour en juillet 2018 et elle a été le fruit d'un travail impliquant plusieurs équipes du laboratoire. Les exigences de rénovation portaient sur à peu près tous les aspects, de l'étiquetage grammatical des textes en parties de discours à la refonte du moteur de recherche et des interfaces. Toutefois, deux points semblaient importants quant à l'ergonomie : d'abord ne pas introduire de fracture avec le public habituel des utilisateurs, en leur permettant de retrouver aisément des routines de travail ; ensuite permettre un accès plus démocratique à des fonctionnalités complexes et enrichies, de sorte à mieux donner à percevoir l'étendue et la qualité des recherches possibles à partir de Frantext.

Retraitement du corpus

La première opération a consisté à découper en unités lexicales (« tokeniser ») un corpus de près de 300 millions de mots. Rappelons que ce découpage a pour objet d'éviter la séparation artificielle d'éléments qui sémantiquement ne font qu'un, comme « aujourd'hui » ou « parce que » et qu'il implique la prise en compte de nombreux cas particuliers, comme des néologismes composés avec des traits d'union ; leur traitement est rendu plus complexe encore par l'encodage XML, qui peut introduire des caractères invisibles, mais parasites et empêcher certains regroupements automatiques. La deuxième opération d'envergure était l'étiquetage grammatical, qui consiste à caractériser la nature grammaticale de chaque unité du texte sans exception (verbe, substantif, adverbe, etc.). Il existe plusieurs logiciels capables d'accomplir cette opération de façon automatisée, mais ils sont en général entraînés sur du corpus journalistique, qui présente la particularité (commode) d'offrir une syntaxe et un registre de langue

relativement homogènes. Dans le cas de Frantext, la situation était tout autre car fallait entraîner le logiciel d'apprentissage sur du texte littéraire qui brassait des genres variés, parfois loin des usages de la langue vernaculaire : poésie, théâtre, essai, roman... Le corpus a donc d'abord été divisé en deux périodes : pré-1850 et post-1850, date à laquelle la langue écrite est globalement stabilisée et où la variation graphique disparaît. Sur le corpus post-1850, qui comprend environ 3000 textes, il m'a été demandé de constituer un échantillon de 100 extraits de 20 000 caractères, qui devaient refléter la diversité de Frantext. Je m'y suis attelée, en respectant l'équilibre global des genres et des périodes dans la base, mais aussi en tentant de situer les textes sur une échelle de difficulté lexicale et syntaxique progressive comportant quatre paliers : pour prendre un exemple simple, *La Chamade* de Françoise Sagan était classé 1 tandis que *Aromates chasseurs* de René Char atteignait le niveau 4. J'ai ensuite composé dix sous-groupes panachant genres et dates, partant de textes majoritairement simples et grand public (pas de lexique spécialisé, une syntaxe standard) pour aller vers des formes de moins en moins normées : lexique et syntaxe propres à la poésie (Saint-John Perse, *Amers*), syntaxe et lexique populaires, incluant l'usage de l'argot (Céline, *Voyage au bout de la nuit*), syntaxe atypique par la longueur des phrases (Proust, *Du côté de chez Swann*). L'objectif était de permettre à l'outil d'étiquetage d'être d'abord entraîné sur un corpus accessible, de capitaliser une première couche d'apprentissage, puis d'attaquer le deuxième sous-groupe, d'un niveau de difficulté un peu plus grand, etc.

Fig. 7. Recherche par unité lexicale de Frantext 2

Une linguiste, Sandrine Ollinger, a piloté le chantier d'étiquetage. Celui-ci a été la source de difficultés nombreuses, lesquelles sont d'ailleurs loin d'être toutes résolues. Après de nombreux tests comparatifs, Talismane, un logiciel libre développé par Asaf Urieli, a été choisi pour procéder à l'étiquetage. Une liste d'étiquettes grammaticales a été arrêtée : elle est assez restreinte (25 étiquettes) et on a renoncé à descendre à des niveaux profonds, par exemple l'indication du temps des verbes conjugués, de peur de créer plus de bruit (information parasite) que d'informations fiables. Le résultat, visible dans l'interface, est qu'il existe maintenant trois entrées pour faire une recherche : par la forme, le lemme, ou la catégorie grammaticale.

La partie pré-1850, elle, a été traitée par Gilles Souvay, mais sans moyens humains supplémentaires et sur la base d'un corpus autrement plus difficile du point de vue de l'irrégularité graphique. Malgré l'usage d'outils performants, comme LGeRM, l'étiquetage des textes anciens, notamment dans la levée d'ambiguïtés, reste de notre point de vue encore largement insuffisant. Il devra nécessiter plusieurs aménagements spécifiques, comme la possibilité d'apposer deux étiquettes différentes (par exemple ADV + ADJ) sur le même mot dans le cas d'un composé, un élément fréquent dans le texte médiéval.



Fig. 8. Étiquetage LGeRM

De manière générale, l'étiquetage de la partie médiévale et préclassique du corpus nécessitera un retraitement en profondeur dans les années à venir pour mieux intégrer la spécificité de ces textes faiblement normés linguistiquement, et qui, en conséquence, répondent assez mal aux tentatives d'étiquetage automatisé.

Une fois l'ensemble de ce corpus de 300 millions de mots, dans ses parties anciennes et modernes, étiqueté, la tâche était loin d'être terminée puisqu'il fallait procéder à son réhabillage en XML-TEI, une tâche assurée par Bertrand Gaiffe : autre étape épineuse, notamment en raison du grand nombre d'annotations du texte, entre étiquettes grammaticales et balises XML, qui entraient parfois en conflit les unes avec les autres.

Langage de requêtes

Le corpus étant traité et enrichi de ses différentes annotations, il a fallu repenser la syntaxe de recherche, c'est-à-dire les codes que l'utilisateur devait saisir pour accéder à telle ou telle fonction et formuler ses requêtes. Le nouvel outil devait en particulier mieux intégrer l'usage des expressions régulières, un langage de recherche largement employé chez les linguistes et qui permet de se servir :

- d'opérateurs de choix : « jour | nuit » cherche *jour* ou *nuit* ; « s.it » cherche toute chaîne de caractères avec un élément non défini en deuxième position, ce qui ramène *suit*, *sait* ou *soit*, « [bp]eau » cherche *beau* ou *peau* (à l'exclusion de toute autre consonne initiale) tandis que « .+ailles » recherche tous les mots suffixés en *-ailles*.

	Texte	Contexte gauche	Pivot	Contexte droit
1	S863	humeur maligne dont le venin nous dévore déjà les	entrailles	? Le Fils de Dieu reconnaît que Pilate a reçu d'en
2	S863	écume. Ô homme, que penses-tu faire, et pourquoi te	travailles	-tu vainement ? - Mais je saurai bien m'affermir et
3	M977	chevet, et le soir, au coin du feu, je te raconterai mes	batailles	. Et le cruel partit. Ni les remontrances, ni les
4	M977	beau et triste comme ces vastes salles du palais de	Versailles	, qu'on admire en les traversant, mais où l'on sent
5	M977	de la rentrée du marquis dans ses terres, et les	fiançailles	de Raoul et d'Hélène. Cette triple solennité

Fig. 9. Résultats d'une recherche sur le suffixe *-ailles* avec Frantext 2

- de quantificateurs « nu(it)? » cherche zéro ou une fois le caractère qui précède (ramenant *nu*, *nuit*) ; « (ha){2} » cherche exactement *n* occurrences de l'expression précédant les accolades (*haha*) ; « cré+e » cherche une ou plusieurs fois le caractère ou groupe qui précède (*créé*, *créée*).

- de classes de caractère : « \d » cherche un chiffre (0, 1, 2, 3) ; „ \W » tout caractère hormis un caractère alphanumérique, par exemple les ponctuations (., -, ?, ;, : , «...), etc.

Le tableau qui suit récapitule une partie des expressions régulières utilisées dans Frantext : il permet de voir la variété des opérations possibles, en particulier grâce à l'usage de l'élément optionnel (« joker »), matérialisé par un point d'interrogation, et qui rend certains éléments facultatifs.

Expression	Description	Exemple de résultats
[word="libertés?"]	Dernier caractère facultatif	liberté, libérés
"âgé?e?s?"	Trois derniers caractères facultatifs	âge, âgées, etc.
"nation.*"	Suffixe de 0 ou plusieurs caractères	nation, nationalisme, etc.
".+able"	Préfixe de 1 ou plusieurs caractères	table, véritable, etc.
"..." ou ".{3,3}"	Mot de 3 caractères exactement	que, est, les, etc.
".."	Un point	,
"[tsf]able"	Mot débutant par t, s ou f	table, sable, fable
"guerre paix"	guerre ou paix	guerre, paix
"[re ap sur]prendre"	Variantes de préfixe	reprendre, apprendre, surprendre
"\d" "janvier"	Un chiffre suivi de janvier	1 janvier, 2 janvier, etc.

Fig. 10. Liste récapitulative des expressions régulières

Le langage de requêtes initialement propre à Frantext avec ses « expressions de séquence » et ses esperluettes constituait la deuxième partie du problème, en raison de son caractère spécifique, et partant marginal. Il a donc été en partie remplacé par CQL (Corpus Query Language). Il s'agit d'une syntaxe attachée à un outil, Sketch Engine, qui permet de chercher des motifs grammaticaux ou lexicaux complexes⁶, et qui est lui aussi aujourd'hui assez largement répandu dans la communauté des linguistes. Ce langage permet en particulier de tenir compte de l'information « catégorie grammaticale » : on peut chercher non seulement des mots (opérateur [word]), mais aussi des lemmes (opérateur [lemma]) ou des classes grammaticales entières (opérateur [pos], abréviation de *part of speech*, ou *partie de discours*, synonyme de « classe grammaticale »).

Expression	Description	Exemple de résultats
[word="bonheur"]	La forme graphique	bonheur
[lemma="aimer"]	Toutes les formes (conjuguées ou non) du verbe	aime, aimer, aimait, etc.
[pos="VINF"]	Tous les verbes à l'infinitif.	être, faire, avoir, etc.

Fig. 11. Exemples de requêtes exprimées en CQL

La puissance de ce langage est augmentée par la possibilité d'y introduire les opérateurs booléens *et*, *ou* et *sauf*, ainsi que de le combiner avec celui des expressions régulières. Il est ainsi possible de rechercher le mot *porte* mais uniquement dans un emploi substantif en formulant la requête: [word=«porte» & pos=«NC»].

5	M977	monter les degrés du perron et franchir le pas de sa	porte	, Raoul suivit sa mère avec un mouvement d'humeur que	
6	M977	dans l'antichambre et cherchait à voir, par la	porte	entr'ouverte, madame la baronne et son fils. Depuis l'	
7	M977	sur ma tête. à ces mots, il marcha résolument vers la	porte	; mais, épuisé par l'effort de dignité qu'il venait	
8	M977	, sans fasto et sans bruit. Stamply le reçut à la	porte	du parc et lui présenta tout d'abord, en guise de clefs,	

Fig. 12. Résultats d'une recherche complexe: le substantif « porte »

L'objectif était donc d'utiliser des procédures de recherche normées, courantes et présentes dans d'autres outils – CQL est également employé par TXM –, de sorte que l'utilisateur un tant soit peu chevronné puisse recycler certains de ses automatismes de linguiste sans avoir à tout réapprendre. Cependant, certaines fonctions historiques avancées de Frantext, comme les listes, le choix des lexiques et les fameuses « grammaires », ces fonctions de recherches combinées, n'existant pas dans CQP, il a fallu recréer plusieurs fonctionnalités spécifiques pour pouvoir les réinsérer dans les requêtes: attributs *liste*, *règle* (pour les « grammaires »), ou encore *lexique*. Là encore, ces fonctionnalités s'ajoutent aux autres et peuvent être combinées, décuplant d'autant la puissance de l'outil

de recherche. Ainsi, la requête &lexicon(« préclassique », « française ») ramènera toutes les flexions préclassiques du lemme « française » : sans surprise *française* et *françaises* mais aussi *françoise(s)* et *françoise(s)*.

Un autre exemple intéressant est celui de listes de mots, qui peuvent être établies en utilisant les possibilités offertes par la combinaison de la syntaxe CQL et des expressions régulières. Imaginons que je veuille examiner la qualification d'un animal, le chat, par une couleur : qui remportera le grand match « chat noir » *versus* « chat blanc » ? Je peux créer une liste de différentes couleurs, en tenant compte du lemme, du lexique, de la catégorie grammaticale, ou en y insérant des expressions régulières pour déclarer certains éléments facultatifs, comme les désinences de pluriel... Je la baptise « couleurs_v » pour la distinguer de la liste « couleurs » présente par défaut, que j'entreprends d'enrichir.

← Liste : couleurs_v
Pas de description SAUVEGARDER

Mot
 Expression régulière
 Flexion
 Expression CQL

Lemme * moderne Catégorie

AJOUTER

	Mot	Sensible à la casse	Sensible aux diacritiques	Expression régulière	Supprimer
1	<input type="text" value=" lemme='rouges?'%c "/>				×
2	<input type="text" value="bleu?s?"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	×
3	<input type="text" value="verte?s?"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	×
4	<input type="text" value="blanc(he)?s?"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	×
5	<input type="text" value="noir?s?"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	×

Fig. 13. Liste de mots

Je vais ensuite faire cooccurrer cette liste avec le nom de l'animal souhaité : ici le lemme *chat*, au masculin et au féminin. Je prends en compte la présence d'un joker entre le substantif et l'adjectif, qui pourrait être un adverbe intensificateur du type *très*, *si* ou *tout* ou encore un substantif désignant la race, par exemple *angora* ou *persan*.

The image shows a search interface with three main panels and a search button.

- Mot 1:** Forme: CHAT|CHATTE; Lemme: CHAT|CHATTE; Catégorie: (dropdown); Options: N'est pas, Mot facultatif, Sensible à la casse, Sensible aux diacritiques; Buttons: RÉINITIALISER, SUPPRIMER.
- Joker:** Distance minimale: 0; Distance maximale: 1; Sauf la catégorie: (dropdown); Buttons: RÉINITIALISER, SUPPRIMER.
- Liste:** Liste: couleurs_v; Description: Pas de description; Buttons: RÉINITIALISER, SUPPRIMER.

A blue button labeled "RECHERCHER" is located below the panels.

Fig. 14. Recherche intégrant une liste de mots

Résultats 1/100 37 résultats en 113ms SAUVEGARDER

Essentiel de données Contour: 100 Exporter Statistiques Vue < >

	Texte	Contexte gauche	Pivot	Contexte droit	Actions
1	P927	. La Butte est à ce moment une pépinière. Il y a le	Chat-Noir	, ancêtre de nos cabarets de Montmartre que Bruant	🔍 ✕
2	E325	, mademoiselle Mallarmé est spirituelle; le	chat est noir	, De chat, d'ailleurs, est une chatte, et répond au	🔍 ✕
3	E333	. Le manège dura bien deux jours, au bout desquels un	chat tout noir	, minuscule et terrorisé, commença de venir à un	🔍 ✕
4	R828	? - Oh non! dit-elle en riant, c'est mon chat, un gros	chat tout noir	. Tu ne veux pas goûter? Olivier aurait bien voulu.	🔍 ✕
5	S294	au talent de Paul Delmet qui venait de se révéler. Du	Chat-Noir	au Café-Concert, du plus élégant salon au plus	🔍 ✕
6	E326	. On n'a jamais su son âge. C'était un gros et beau	chat tout noir	, avec quelques poils blancs au cou et au ventre.	🔍 ✕
7	E326	; mademoiselle Mallarmé est spirituelle; le	chat est noir	, De chat, d'ailleurs, est une chatte, et répond au	🔍 ✕
8	E283	cherchait à se souvenir de la couleur de la	chatte. Blanche	avec des points gris ? Avec des points roux ? Il - Chat	🔍 ✕
9	E196	fumante sur la table du salon japonais. Un	chat encore blanc	sorti d'un vieux Walt Disney sauta du fauteuil où	🔍 ✕
10	E382	aux tendres flocons. De mon livre dépasse le	chat botté Rouge	rouge rouge est mon chaperon. Est-ce la légende.	🔍 ✕
11	R927	, des délieries stupéfiantes et suprêmes de	Chat-Noir	. En somme, la rencontre de ce Motel aurait eu sur lui	🔍 ✕
12	K386	de bois et de saumons, les frimousses des petits	chats de Rose	, encore trop jeunes pour préférer, comme leur	🔍 ✕

Fig. 15. Résultat d'une recherche intégrant une liste de mots

Mais le chat a plus d'un tour dans son sac, et parfois plus d'une couleur dans son pelage. Qu'à cela ne tienne : il suffit d'ajouter à la requête un deuxième joker et un deuxième appel de la liste « couleur », ce qui en CQP se traduit par :













1	E331	une photo de moi à huit ans, dans le jardin, un	chat noir et blanc	dans les bras, gros pull, bouffe ronde, cheveux	 
2	S525	la nouvelle venue berçait comme un poupon le gros	chat blanc et roux	qu'Élisabeth connaissait déjà. Cet animal	 
3	S735	? et à Maman ? C'est alors que je cassai la tête de	chat bleue et rose	où on mettait pour moi les sous-neufs, comme de l'	 
4	S037	sans complaisance. Je me souviens aussi qu'un	chat blanc et noir	descendait la travée de gauche, celle qui sépare	 
5	E382	tendres flocons. De mon livre dépose le	chat botté rouge rouge rouge	est mon chaperon. Est ce la légende, est	 
6	S225	pour elle sur la table de nuit. Il y avait une	chatte noire et blanche	qui ne me quittait pas et allait prévenir mes	 

Fig. 16. Résultat d'une recherche complexe intégrant une liste de mots

```
((lemma=>CHAT|CHATTE)%c) ([{0,1}) (&liste(«couleurs-v»)) ([{0,1})
(&liste(«couleurs-v»))
```

Et le tour est joué !

Moteur de recherche

La dernière opération a consisté à récrire le moteur de recherche, une tâche effectuée par l'équipe Soutien technique à la recherche. Ce nouveau moteur, baptisé Allegro, devait être en mesure de prendre en compte les informations multiples que nous avons évoquées (étiquettes grammaticales, lexiques), de recréer l'ensemble des anciennes fonctionnalités du Frantext historique et d'en ajouter de nouvelles, le tout en restant lisible et ergonomique aux yeux des utilisateurs. Forte gageure... Cette phase a été ardue, et pas seulement pour des raisons techniques. La logique d'un informaticien programmeur et celle d'un utilisateur, dont je me suis faite souvent la diabolique avocate, ne sont pas toujours les mêmes : là où l'un va souvent privilégier une forme d'implicite et de complexité, pour plus d'efficace et une performance rapide, l'autre va plaider la simplification, l'explication méthodique, voire la redondance des informations, génératrices de lenteur. J'ai ainsi insisté pour insérer sur les pages des principales fonctions de recherche quelques lignes d'explications relatives à la fonctionnalité déployée, plutôt que de les reporter d'emblée dans la documentation, au demeurant fort riche, proposée dans la rubrique du même nom.

Il est aussi apparu que chacun ne mesurait pas toujours non plus la difficulté de la tâche de l'autre : le caractère parfaitement opérationnel et logique d'un formulaire qui a été long à programmer et fonctionne sans accroc, du point de vue strictement technique, va parfois devoir être entièrement revu, car il est peu commode à utiliser. Le formulaire de constitution de corpus, par exemple, a dû être repris plusieurs fois, en raison du caractère contre-intuitif de plusieurs opérations : celles-ci, quoique cohérentes, ne pouvaient que dérouter et gêner l'utilisateur dans cette opération préalable à la consultation de la base. Beaucoup de débats ont tourné autour de l'ergonomie et du degré de guidance que nous devons à nos actuels et futurs utilisateurs face à cette nouvelle version. En tout état de cause, nous nous sommes mis d'accord pour déterminer trois modes de recherche :

Fig. 17. Recherche simple

- La **recherche simple**, avec ce que l'on appelle un « guichet unique », directement inspiré de grandes bases de données comme Gallica ou Gutenberg : une seule barre de saisie, dans laquelle l'utilisateur entre un mot ou une suite de mots.

• La **recherche assistée**. Celle-ci permet d'utiliser plus de fonctionnalités tout en guidant de manière pédagogique l'utilisateur. Celui-ci entre les différents éléments de sa recherche dans des « boîtes » prévues à cet effet, chaque boîte lui permettant de choisir entre forme, lemme et catégorie, d'activer la flexion médiévale ou moyen français s'il le souhaite, ou de placer un élément facultatif entre deux mots recherchés. Le moteur traduit ensuite cette requête en langage CQL.

The image shows two side-by-side search boxes labeled 'Mot 1' and 'Mot 2'. Each box contains the following elements:

- Forme:** A text input field.
- Lemme:** A text input field containing 'CHAT' for Mot 1 and 'persan' for Mot 2.
- Catégorie:** A dropdown menu.
- Options:** Four checkboxes: 'N'est pas', 'Mot facultatif', 'Sensible à la casse', and 'Sensible aux diacritiques' (checked).
- Buttons:** 'RÉINITIALISER' and 'SUPPRIMER' in red text.

A blue button labeled 'RECHERCHER' is located below the 'Mot 1' box. A blue circle with a white plus sign is positioned between the two boxes.

Fig. 18. Recherche assistée

• La **recherche avancée** permet directement de formuler des requêtes en CQL. Dès le départ, nous avons prévu de faire de Frantext 2 une passerelle vers l'utilisation des fonctions complexes peu exploitées dans Frantext 1 malgré leur puissance. Une idée pédagogique consistait donc à afficher la formulation de la requête en CQL en haut de chaque page de résultat, exactement comme on faisait du « petit latin » à la faculté en lisant le texte original sur une page et sa traduction en regard. Le mode « avancé » permet ensuite de récupérer cette expression CQL, de la copier et la mémoriser pour une réutilisation ultérieure, ou de la modifier directement, ce qui va parfois plus vite que de remplir les cases une à une. Peu à peu l'utilisateur se familiarise avec les opérateurs et les codes, et il apprend à les manier.

🔍 ((lemma="CHAT"%c)) ((lemma="persan"%c))

Fig. 19. Recherche avancée

La nouvelle version remplit ainsi l'objectif qu'elle s'était fixé : devenir un outil capable de couvrir toute une gamme de besoins, des plus simples aux plus experts, et permettre à une communauté plus large d'accéder facilement à des fonctionnalités de recherche poussées. Celles-ci trouvent leur première application dans des recherches à caractère linguistique, mais peuvent tout aussi bien être utiles à tout chercheur, quelle que soit sa discipline, désireux de localiser des thèmes récurrents ou des isotopies dans un corpus de textes récents.

La base de données Frantext, qui fut pionnière dans le paysage des humanités numériques, a dû profondément se reconfigurer, vingt ans après son lancement sur le web, pour conserver sa pertinence. Elle l'a fait en capitalisant sur ses points forts, à savoir poursuivre le couplage entre un corpus échantillonné de 5000 textes contenant plus d'une moitié de textes postérieurs à 1900 et un outil de recherche performant. Mais sa mue a également été l'occasion de varier et de densifier son corpus, notamment en diachronie, d'ajouter différentes strates d'annotations (comme l'étiquetage grammatical), d'enrichir certains outils (les lexiques), d'intégrer d'autres langages d'interrogation (les expressions régulières, CQP), ou de permettre de meilleures visualisations statistiques, ce qui fournit autant de points d'entrée supplémentaires pour mener à bien des recherches nouvelles. L'autre aspect délicat de cette refonte a tenu au fait qu'il ne fallait pas se couper des utilisateurs historiques, tout en rompant avec certaines routines, et proposer de nouvelles fonctionnalités, techniquement élaborées, mais visibles et simples à utiliser.

Évidemment, un tel chantier ne va pas sans difficultés : celles-ci ont jalonné toutes les étapes de la réfection de la base et ont fait apparaître un certain nombre d'attentes contradictoires autour de cette ressource. Certains souhaitaient un haut degré de technicité, pendant que d'autres récusaient un usage trop expert ou trop tourné vers les linguistes à l'exclusion des autres disciplines. Les plus révolutionnaires auraient volontiers arasé toute trace du Frantext historique, tandis que les plus impliqués dans l'utilisation quasi quotidienne de la base faisaient de la continuité un gage essentiel de la réussite de la transition.

Tenter de concilier tous ces points de vue fut complexe, mais a permis de rendre la base plus lisible et plus visible auprès d'une communauté scientifique élargie : celle des linguistes, bien sûr, qui en sont les utilisateurs premiers, mais aussi celle des chercheurs en littérature, sociologie, histoire, et toute autre discipline qui peut avoir intérêt à disposer d'un corpus raisonnable en taille interrogeable avec un tel niveau de précision. En effet, même si (et parce que...) le nombre

de textes en stock est immense dans d'autres ressources, pas toujours facile de trier dans les milliers de résultats retournés par certaines requêtes dans Google Livres ou même Gallica...

Évidemment, il reste beaucoup à faire, en termes d'aménagement des interfaces, qui manquent encore d'intuitivité et d'amélioration de la didactisation ; l'organisation de la documentation, pour l'heure touffue, ainsi que l'élaboration de tutoriels vidéo, font ainsi partie du cahier des charges de l'avenir proche. Le véritable point noir est celui de l'étiquetage grammatical, en particulier des textes anciens, qui nécessitera encore plusieurs années de travail, envisagées par étapes, pour atteindre un niveau de qualité congruent à nos ambitions. Par ailleurs d'autres développements, comme l'annotation des noms propres et l'annotation en dépendance syntaxique, ou encore des formes plus complètes de visualisation statistique des résultats, sont envisagés, là encore sur un terme moyen ou long.

Mais malgré les imperfections qui subsistent, on peut se réjouir que dans un paysage devenu archi-concurrentiel du côté des outils, et de plus en plus sensible d'un point de vue juridique, la base Frantext ait pu demeurer un acteur majeur de la recherche pour tous les scientifiques impliqués dans le domaine de la langue française, et plus largement de la recherche sur des textes écrits en français.

Bibliographie

Buchi, Éva & Jean-Marie Pierrel (2009): «Research and Resource Enhancement in French Lexicography: the ATILF Laboratory Computerised Resources», in: Bruti, Silvia, Roberta Cella & Marina Foschi Albert (éd.): *Perspectives on Lexicography in Italy and Europe*, Newcastle-upon-Tyne: Cambridge Scholars Publishing, p. 79-117.

Imbs, Paul (éd.) (1961): *Lexicologie et lexicographie françaises et romanes. Orientations et exigences actuelles (12-16 novembre 1957)*, Paris: Éditions du CNRS.

Versione online del VSI e sviluppi futuri del progetto di informatizzazione

Dafne Genasci e Dario Petrinì

Il *Vocabolario dei dialetti della Svizzera italiana* (VSI) nasce all'inizio del Novecento e si affianca a simili imprese dedicate alle altre tre regioni linguistiche della Confederazione, ossia lo *Schweizerisches Idiotikon*, il *Glossaire des patois de la Suisse romande* e il *Dicziunari Rumantsch Grischun*. Istituito con lo scopo di documentare, conservare e analizzare il patrimonio dialettale della Svizzera italiana, il VSI è un'opera di carattere enciclopedico, che unisce l'interesse linguistico (e più in particolare lessicale) a quello etnografico e folclorico. Esso si compone attualmente di 8 volumi. Ogni voce del VSI presenta a lemma il termine dialettale, a cui seguono i significati che può assumere e le diverse pronunce locali; a queste informazioni si aggiungono locuzioni, modi di dire, proverbi, filastrocche o altre espressioni in cui il termine dialettale compare e, a seconda della natura del referente, vengono fornite informazioni riguardanti usi, tecniche di lavorazione o di sfruttamento, tradizioni e credenze ad esso inerenti; infine, in conclusione dell'articolo, viene indagata l'origine del termine.

Nel 2016 è stata pubblicata la versione online del VSI¹, consistente in un database che raccoglie tutti gli articoli del VSI finora usciti a stampa. Buona parte degli articoli, a partire dal lemma *brütt maa*, era già in formato digitale (QuarkXpress), perché redatta con il computer; quanto agli articoli precedenti, si è optato per una retrodigitalizzazione dei volumi cartacei, affidandosi a un programma OCR (Optical Character Recognition), ossia un programma di riconoscimento ottico dei caratteri (nella fattispecie Abby FineReader). Come ci si attendeva, la parte di testo scansionato conteneva svariati errori, dovuti soprattutto alla presenza dei caratteri fonetici, che sono stati risolti perlopiù manualmente e con ricerche mirate per tipologia di errore. Questi materiali sono poi stati consegnati a un informatico che li ha elaborati per allestire un database

1 Consultabile all'indirizzo <www.vsi-online.ch>, è attualmente accessibile ai soli abbonati del VSI attraverso una password nominale, ma si prevede che venga in futuro aperta a tutti (in conformità alla politica di accesso all'informazione adottata dall'Accademia svizzera di scienze umane e sociali). La visualizzazione del VSI online è stata adattata anche per tablet e smartphone.

organizzato per lemmi: il prodotto che si presenta all'utente non è dunque un semplice PDF. Ogni volta che viene pubblicato un nuovo fascicolo del VSI, esso viene caricato nel database.

All'interno di questo database si possono effettuare due tipi di ricerca: per «lemma» (che permette di cercare solo fra i termini a lemma) o per «testo» (che consente invece di cercare all'interno di tutti gli articoli del VSI, vale a dire una ricerca full text). In entrambe le ricerche non è necessario inserire accenti o caratteri speciali, poiché il database è costruito su due livelli: un livello di visualizzazione, che riporta tutti gli accenti, i diacritici, i caratteri fonetici ecc. (ad esempio il termine dialettale *dirüpéri* «acquazzone, bufera»); l'altro che presenta invece un testo normalizzato, ossia senza accenti, senza diacritici o caratteri fonetici (il termine citato appare dunque come *diruperi*) e che consente all'utente di fare delle ricerche senza dover inserire segni speciali. Per questo motivo, digitando *cova* nel campo di ricerca per «lemma», si ottengono i seguenti risultati: *cová*¹ «covare», *cová*² «attaccatura della coda, parte terminale della colonna vertebrale», *cóva* «coda» e *cöva* «covone». È possibile cercare anche solo una parte della parola, con l'aiuto dell'asterisco: inserendo **cova** nel campo «lemma», oltre ai risultati già citati, si troveranno anche *brüsacóva*, *covaia*, *covaróss*, *covaróssa*, *covatass* e *covazza*. Quando si apre un articolo del VSI online, il testo appare a schermo essenzialmente come nel cartaceo, con tutti i caratteri speciali. La principale differenza fra le due versioni è la suddivisione dell'articolo in sezioni, ciascuna introdotta da un titoletto: «Lemma e significato», «Varianti», «Trattazione», «Etimologia», «Bibliografia e note» e «Autore»:

e
Vocabolario dei dialetti della Svizzera italiana
online

Lemmi Testo Elenco lemmi Elenco figure Abbreviazioni Bibliografia

☰ ← →

Cerca nella pagina

A+ A- 100% 200%



Fig. 111

Lemma e significato

CLAUD (kláut) s.m. Scompartimento, scaffale, cassetto.

Varianti

Var.: *claud* (Castasegna), *cláudar* (Bondo), *clódar* (SopraP.).

Trattazione

Indica diversi elementi del mobilio tradizionale: a Bondo il cassetto che si apre sul davanti del mobilio, a Castasegna ciascuno degli scaffali posti nella parte superiore della credenza, chiusa da antine, nella SopraP. il piccolo scompartimento posto a sinistra in alto della cassapanca per la biancheria: il suo coperchio, se sollevato, mantiene aperto anche quello del mobile.

Etimologia

Stampa fa derivare la var. *clódar* direttamente dal lat. CLAUDERE 'chiudere' [1]; per l'esito di CLAU- quale si osserva nella forma *cláudar* cfr. → *cladizián*; la var. di Castasegna perfeziona la sostantivizzazione con la caduta del suff. verb. Meno probabile la discendenza dal lat. CALATHU(M) 'cesto; secchio' avanzata da Salvioni, anch'egli prevedendo comunque il concetto di CLAUDERE o di sue forme participiali [2]. Cfr. anche il rom. *claster*, nel suo senso specifico di 'separazione, scompartimento della cassapanca' [3].

Bibliografia e note

Bibl.: [1] STAMPA, Bergell 58. [2] SALVIONI, R 43. 577. [3] DIRG 3.200.

Autore

Moretti

Fig. 1. Esempio di un articolo nella versione online del VSI

Queste sezioni permettono di fare delle ricerche più mirate, come verrà illustrato in seguito. Le immagini, che nel cartaceo sono integrate nel testo, appaiono qui sotto forma di icone cliccabili sulla sinistra. Inoltre, dall'articolo visualizzato è possibile spostarsi avanti e indietro all'interno dei risultati di ricerca (nel caso ve ne sia più di uno) per mezzo di frecce, o ancora ritornare direttamente alla lista dei risultati della ricerca.

Più ampie sono le possibilità di ricerca nel campo «testo» (fig. 2), quali ad esempio la ricerca per temi: l'utente potrebbe essere interessato a consultare tutti gli articoli del VSI che hanno toccato l'argomento «ferragosto». Dalla ricerca emerge che se ne parla ad esempio sotto il lemma *agost* «agosto», nel paragrafo relativo alle festività del mese; o ancora sotto il lemma *Denedaa* «Natale», dove si trova il proverbio *Dinadaa pulénta e saa, mèzz aúst pulénta e crust, san Martígn pulénta e vign* «a Natale polenta e sale, a ferragosto polenta e croste, a San Martino (11 novembre) polenta e vino». La ricerca può essere condotta su tutte le sezioni oppure essere affinata focalizzandosi solo su una o alcune delle sezioni che compongono ogni articolo del VSI.



Fig. 2. La maschera di ricerca per testo nel VSI online

Un potenziale utente alla ricerca dei lemmi dialettali derivati dal latino *CAMPANA*, cercando nel campo «testo», otterrebbe 378 occorrenze in 115 lemmi diversi, un numero piuttosto alto di risultati che potrebbe effettivamente scoraggiarlo a proseguire nella ricerca; questo perché il sistema trova indistintamente tutti gli articoli in cui compare la parola *campana*, come il lemma *cavra* «capra» che a Poschiavo significa «mozzo, armatura di legno della campana». Limitando invece la ricerca alla sezione «Etimologia», otterrebbe un risultato più accessibile, ovvero 44 occorrenze in 28 lemmi diversi; sebbene il sistema non sia in grado di distinguere fra *CAMPANA* in latino e *campana* in italiano o in dialetto, la ricerca diventerebbe molto meno onerosa. È stata inserita anche la possibilità di distinguere fra maiuscole e minuscole: nel VSI può infatti rivelarsi utile distinguere Monti, autore di un usatissimo vocabolario dialettale di Como (Monti 1845), da *monti*, plurale dell'appellativo *monte*.

Scorrendo la pagina «Elenco dei lemmi», si ha una panoramica di tutte le voci pubblicate finora e cliccando successivamente sulla singola voce si accede al relativo articolo. Sul sito si trovano inoltre gli elenchi delle immagini, delle fonti bibliografiche e delle abbreviazioni usate nel VSI.

A partire dal dicembre 2017 si è cominciato a sviluppare una nuova parte del progetto online, nella forma di un database che si integrerà con quanto è già presente sul sito. La sua realizzazione renderà possibili ricerche linguistiche più specifiche e mirate rispetto alla ricerca full text, nella fattispecie riguardanti l'etimologia e la formazione delle parole, considerate come fondamentali chiavi di accesso ai materiali offerti dal VSI. Punto di partenza per impostare ed elaborare il nuovo database sono gli indici etimologici e morfologici già allestiti per la versione cartacea. L'operazione di trasferimento nel database non è di tipo meccanico, dal momento che gli indici che chiudono i volumi fin qui pubblicati sono stati redatti con criteri in parte diversi fra loro; per consentire l'inserimento dei dati all'interno di uno schema unitario è stato dunque necessario uniformare tali criteri. Inoltre va ricordato che, a partire dalla lettera C, alcuni indici sono stati soppressi, fra i quali l'indice riguardante la formazione delle parole: occorrerà pertanto recuperare i singoli dati partendo direttamente dalle sezioni di commento negli articoli del VSI².

2 L'operazione, fra l'altro, vede il VSI sulla stessa via percorsa dal GPSR nell'allestire la sua banca dati.

Il nuovo database, tuttora in fase embrionale e concepito per una prima schedatura dei dati, è stato elaborato su FileMaker. Vi si trovano, da un lato, le informazioni relative all'etimologia del termine dialettale (base o basi che vi concorrono), specificata secondo certe caratteristiche (origine dal lessico di una determinata lingua, o da un nome di luogo/di persona, oppure infantile, onomatopeica, fonosimbolica) e il grado di certezza (etimologia sconosciuta, incerta, respinta); dall'altro, i dati relativi alla sua struttura (prefissi e suffissi) e al tipo di formazione (derivato, composto, incrocio). Sono inoltre registrate alcune altre informazioni riguardanti la natura del lemma (appellativo, antropónimo, topónimo, termine di gergo ecc.).

Un punto importante per la concezione del database è che esso vuole rendere conto di tutte le osservazioni etimologiche e relative alla formazione enunciate nella sezione di commento degli articoli: nel VSI, la discussione non riguarda infatti unicamente l'etimo della parola che compare a lemma, ma analizza molto spesso anche sue forme particolari, frutto di incroci o di paretimologie, oppure le modalità della loro derivazione e composizione. Si è imposta così da subito la necessità di distinguere due campi per la cattura dei lessemi dialettali, il campo (provvisoriamente chiamato) del «termine» e quello del «lemma». Due esempi potranno chiarire meglio il senso di questa distinzione.

1. Il lemma del VSI *andat* «accesso, passaggio» presenta una variante *andigh* [ándik] alla quale è dedicato un commento specifico nella discussione:

«[*andat*] ha la stessa origine dell'it. *andito* < ANDITU, per cui vedi REW 410 [...]. Molto probabilmente da un incontro con PORTICU è sorto il tipo *andik* (per cui v. anche crem. *andegh*, *andeghèt* «androne, corritojo» SAMARANI 17; *andek* «andito della stalla» AIS 6.1169 [...])» (VSI, vol.1, p. 170).

La variante richiede allora una scheda supplementare nel database, con la registrazione della forma *andigh* nel campo «termine», simile a ma non coincidente con quello del «lemma», e l'indicazione a etimo del latino PORTICUS per rendere conto del fatto che essa è il risultato di un incrocio.

INDICE DEGLI ETIMI

etimo	origine	termine	lemma	rimando	pag.			
porticus	lat.	ándigh	andat		1.170			
lunghezza tonica	et. respinto	et. sconosciuto	et. incerto	incrocio	antroponom.	deonem.	toponom.	gergo
<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Formazione delle parole:								
derivato	<input type="checkbox"/>	composto	<input type="checkbox"/>	cambiam. pref.	<input type="text"/>	cambiam. suff.	<input checked="" type="checkbox"/>	<input type="text"/>
prefissi	<input type="text"/>		infissi	<input type="text"/>		suffissi	<input type="text" value="jicu"/>	
oss.	<input type="text"/>							

Fig 3. Indice degli etimi: la scheda di *andigh*

2. La forma *aradell* «aratro» è sorta mediante un suffisso a partire dal suo sinonimo *araa* (< latino *ARATUM, latino classico ARATRUM). Nell'indice cartaceo relativo alla formazione delle parole viene solo indicato che un caso del cumulo -ATU + -ELLU si trova sotto il lemma *araa*²; sulla scheda in FileMaker del database, *aradell* figura invece materialmente nel campo «termine», accanto alla segnalazione del «lemma» *araa*² che lo accoglie³.

INDICE DEGLI ETIMI

etimo	origine	termine	lemma	rimando	pag.			
aratum	lat.	aradell	araa 2		1.239			
lunghezza tonica	et. respinto	et. sconosciuto	et. incerto	incrocio	antroponom.	deonom.	toponom.	gergo
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Formazione delle parole:								
derivato	<input checked="" type="checkbox"/>	composto	<input type="checkbox"/>	cambiam. pref.	<input type="checkbox"/>	cambiam. suff.	<input type="checkbox"/>	
prefissi	<input type="checkbox"/>		infixi	<input type="checkbox"/>		suffissi	<input type="checkbox"/> -atu + -ellu	
oss.	<input type="text"/>							

Fig. 4. Indice degli etimi: la scheda di *aradell*

3 La pratica di trattare i derivati e i composti sotto la loro base dialettale, riscontrabile fin dagli inizi del VSI, è stata più rigidamente formalizzata a partire dalla lettera C. Allo stato attuale, un derivato o un composto vengono trattati sotto la relativa base se, per illustrarli, si ritiene sufficiente fornirne la traduzione e la qualifica grammaticale. Quora, invece, i contenuti che veicolano siano qualitativamente pregnanti (per esempio, se essi presentano locuzioni, modi di dire, proverbi, credenze, approfondimenti etnografici ecc.), si decide di farne degli articoli indipendenti.

Sotto quale aspetto si presenterà questo nuovo database online e come avverrà il collegamento con il VSI online già pubblicato? L'interfaccia ricorderà certamente in vari punti quella del *Glossaire des patois de la Suisse romande*, da tempo consultabile in rete. Sarà possibile effettuare ricerche in tutti i campi che figurano nella scheda di FileMaker: si potrà così cercare per etimo, per suffisso o prefisso, cercare tutti i termini composti, quelli frutto di incrocio, l'insieme dei lemmi gergali e così via.

Allo stadio attuale del lavoro, si immagina che i risultati di ogni interrogazione potranno venire rappresentati in una lista a due colonne, corrispondenti rispettivamente al campo «termine» e al campo «lemma». Per fare un esempio di ricerca possibile, poniamo che l'utente sia interessato ai continuatori del latino ANIMA nei dialetti della Svizzera italiana; l'esito della richiesta sarà il seguente:

	<u>termine</u>	<u>lemma</u>
	ánima	ánima
derivati	armá	armá ²
	armass	armass
	armèla	armèla
	armí	armí
	armözz	armözz
	armüsc	armüsc
composti	belèrma	ánima
	bonèrma	ánima
	debilèrma	ánima
incroci	armándola	armándola

Cliccando su un elemento della lista «lemma» potrà accedere al relativo articolo del VSI nella versione online; cliccando invece su uno della lista «termine» verrà rimandato a un'ulteriore scheda, contenente le indicazioni relative alla sua origine (etimo, lingua di appartenenza oppure onomatopea, fonosimbolismo ecc., grado di sicurezza dell'etimo) e alla sua formazione (presenza di prefissi e/o di suffissi), indicazioni che costituiranno nel loro insieme una sorta di «carta di identità» linguistico-etimologica.

Il progetto in corso stimola anche dei desideri. Nel 2004 è stato dato alle stampe il *Lessico dialettale della Svizzera italiana* (LSI), che raccoglie in 5 volumi il tesoro lessicale completo della regione. In futuro, è auspicabile che un LSI online venga ad affiancare il VSI online. Un consultatore potrebbe così entrare, anche con il suo tablet o il suo smartphone, nel LSI online e ricevere le informazioni più immediate e sintetiche che offre; qualora la parola fosse già trattata dal VSI, potrebbe infine accedere alla versione online di quest'ultimo e agli approfondimenti che lo caratterizzano. Mettere in rete il LSI comporta infine che esso sia affiancato dal suo necessario complemento, uscito in versione cartacea nel 2013, il *Repertorio italiano-dialetti* (RID); con questo strumento si creerebbe un accesso ai dati del LSI online e del VSI online a partire dai traducanti italiani dei termini dialettali.

Bibliografia

LSI = *Lessico dialettale della Svizzera italiana*, 5 vol., Bellinzona: Centro di dialettologia e di etnografia, 2004.

Monti, Pietro (1845): *Vocabolario della città e diocesi di Como con esempi e riscontri di lingue antiche e moderne*, Milano: Società Tipografica de' Classici italiani.

RID = *Repertorio italiano-dialetti*, 2 vol., Bellinzona: Centro di dialettologia e di etnografia, 2013.

VSI = *Vocabolario dei dialetti della Svizzera italiana*, Lugano/Bellinzona: Centro di dialettologia e di etnografia, 1952–.

Dicziunari Rumantsch Grischun – Der lange Weg zur Retrodigitalisierung und zur Online-Publikation

Ursin Lutz

Einleitung

Wer den Namen «Dicziunari Rumantsch Grischun» (DRG) zum ersten Mal hört, verbindet damit oft die Vorstellung eines Wörterbuchs, das den Wortschatz der im Jahre 1982 geschaffenen bündnerromanischen Standardsprache Rumantsch Grischun zum Inhalt hat. Wird aber die Tatsache berücksichtigt, dass das Institut dal DRG schon seit mehr als 100 Jahren besteht, erkennt man leicht, dass dies nicht der Fall sein kann. Die Namensgebung des DRG soll zum Ausdruck bringen, dass wir es hier mit einem Wörterbuch zu tun haben, in dem alle Wörter der im Kanton Graubünden schriftlich und mündlich verwendeten bündnerromanischen Idiome und Dialekte erfasst und erklärt werden.

Das Institut dal DRG wurde im Jahr 1904 gegründet, als Trägerverein fungiert die im Jahre 1886 gegründete Societad Retorumantscha. Nach einer grundlegenden Materialsammlung, die über zwei Jahrzehnte andauerte, erschien der erste Faszikel (A – ADEMLAT) im Jahr 1939. Aktuell sind 14 Bände und sechs Doppelfaszikel (Stichwörter von A bis MINDRAMAINA auf fast 11 000 doppel-spaltigen A4-Seiten) publiziert.

Das Institut dal DRG in Chur, wo das Dicziunari Rumantsch Grischun redigiert wird, ist ein wichtiges Dokumentations-, Informations- und Forschungszentrum zur bündnerromanischen Sprache und zur alpinen Kultur. Es verfügt über eine umfangreiche Arbeitsbibliothek mit über 30 000 Titeln, die auch Besuchern zur Verfügung stehen. Das DRG mit seinen verschiedenen Materialsammlungen und seinen Publikationen ist heute eine unverzichtbare Adresse für all jene, die sich mit Fragen und Forschungen zur bündnerromanischen Sprache, Geschichte und Kultur beschäftigen.

Grundlagen

Die ersten Schritte mit dem Computer

Nachdem die Redaktion des DRG während vieler Jahrzehnte ihre Artikel von Hand geschrieben hatte – Copy & Paste wurde noch mit Schere und Klebstreifen bewerkstelligt –, konnten im Jahr 1991 die ersten zwei Macintosh-Computer erworben werden.¹ Den Umstieg auf Computer machte die Tatsache nötig, dass die damalige Druckerei, die Druckerei Winterthur AG, ihre Maschinensatzabteilung, nur noch für den Druck des DRG in Betrieb gehalten, schloss (DRG 9: III). Das eigentliche Redigieren am Computer startete im Jahr 1992, nach Lösung der technischen Probleme um die Sonderzeichen: Ein erster Schritt auf dem Weg zur (Retro-)Digitalisierung des DRG war geschafft (Tomaschett 2004: 10). Der Redaktionsprozess wurde nach und nach auf den Computer verlagert, die Redaktorinnen und Redaktoren erfassten ihre Artikel fortan mit Microsoft Word. Auch die Layoutarbeiten erfolgten von Anfang an im Haus, zunächst mit dem Programm QuarkXPress.

Ebenfalls ab dem Jahr 1992 wurden zusätzlich zahlreiche Karteien und Register, vorher nur auf Zetteln und Listen vorhanden, in FileMaker-Datenbanken erfasst.² Diese Datenbanken wurden seither gepflegt und laufend aktualisiert; der DRG-Redaktion erweisen sie noch heute wertvolle Dienste.

- 1 Cf. AnSR 1992: 234 [Jahresbericht SRR 1991]. Eine erste Prüfung zur Einführung des computergestützten Redigierens erfolgte jedoch bereits 1986: «Die Societad Retorumantscha liess im Sommer 86 eine Detailanalyse und ein Rahmenkonzept für die Einführung der EDV durch einen Spezialisten erstellen» (EDV-Einsatz 1991). Im Jahr 1987 schliesslich tauchte zum ersten Mal das Kapitel «Dicziunari ed ordinatur» [Dicziunari und Computer; Übersetzung UL] im Jahresbericht der Societad Retorumantscha auf (AnSR 1988: 233). Gemäss dem Bericht für das Jahr 1990 waren die Vorarbeiten zur Einführung des Computers so weit fortgeschritten, dass man bereits Offerten für die Anschaffung der Geräte vorliegen hatte. Allerdings fehlte noch eine Lösung für die Eingabe der zahlreichen Sonderzeichen (AnSR 1991: 161).
- 2 Cf. AnSR 1993: 336f. [Jahresbericht SRR 1992]: «Register dals chavazzins e renviaments ... Nossa secretaria ... ha gia registrà ... en tut varga 15000 unitads» [Register der Stichwörter und Verweise. Unsere Sekretärin hat gesamthaft bereits über 15000 Einheiten erfasst; Übersetzung UL]. Ab dem Jahr 1996 ist eine tabellarische Übersicht mit dem Titel «Stadi da la programmaziun da la banca da datas» [Stand der Datenbankprogrammierung; Übersetzung UL] fester Bestandteil der Jahresberichte der Societad Retorumantscha (cf. AnSR 1997: 259).

Herausforderung Sonderzeichen und deren Eingabe

Seit Beginn der Einführung von EDV-Mitteln in den Arbeitsablauf der DRG-Redaktion erwies sich eine einwandfreie und den älteren, mit Bleidruck hergestellten DRG-Bänden entsprechende Abbildung der DRG-Schriften mit all ihren Sonderzeichen als grösste technische Herausforderung.³ Bei den ersten Gehversuchen der Redaktoren am Computer zeigte sich schon, dass diese Komponente über die nächsten Jahrzehnte zum Dauerbrenner unter den technischen Problemen werden sollte.⁴

Für die Digitalisierung der Zeichensätze wurden die bestehenden DRG-Schriften, basierend auf dem Zeichensatz Bauer Bodoni, aus dem 10-Punkt-Bleisatz der Druckerei Winterthur AG hochvergrössert und als Computerschriften gespeichert. Da die Zeichenkapazität des damaligen ASCII-Standards nicht ausreichend war, wurden die einzelnen Zeichen in verschiedenen Schriften abgelegt, insbesondere wurde für die phonetische Schrift neben der regulären kursiven Schrift ein eigener Schriftsatz angelegt. DRG-spezifische Zeichen wurden auf freien Positionen der ASCII-Tabelle platziert. Über die für Macintosh programmierbaren Tastaturbelegungen konnte das Aufrufen der einzelnen Zeichen über die jeweiligen Tasten bzw. über Tastenkombinationen individuell gesteuert werden. Mit diesem Ansatz fand die gesamte Zeichensteuerung auf Ebene Betriebssystem und nicht auf Programmebene statt, folglich konnten die DRG-Schriften schon immer in sämtlichen Programmen eingesetzt werden.

Der Benutzerfreundlichkeit wurde von Anfang an ein sehr grosser Wert beigegeben. So wollte man zur Eingabe der vielen Sonderzeichen eine lange Liste mit komplizierten Tastenkombinationen bewusst vermeiden. Nach vielen Versuchen und eingehendem Tüfteln wurde eine höchst einfache, aber ebenso geniale Lösung gefunden: Die Sonderzeichen, die mit dem jeweiligen Grundbuchstaben kombiniert werden sollten, wurden auf dem Dezimalblock platziert und mit selbst hergestellten Etiketten sichtbar gemacht. So wurde eine für den Redaktor sehr zeitsparende, intuitive und angenehme Lösung gefunden. Von dieser Errungenschaft profitieren wir bis heute!

3 Diese Tatsache bestätigt bereits der Jahresbericht für das Jahr 1990: In diesem wird berichtet, dass für die Einführung von EDV-Mitteln eine Offerte vorliege, jedoch ohne eine technische Lösung für die Sonderzeichen (cf. AnSR 1991: 161).

4 Cf. AnSR 1993: 338 [Jahresbericht SRR 1992]: «Per l'empreu hai sa tractà d'optimar noss sistem: las scrittiras, ils segns spezials ed ils cumonds da la tastatura» [Zunächst ging es darum, unser System zu optimieren: die Schriften, die Sonderzeichen und die Tastaturbefehle; Übersetzung UL].



Abb. 1. Belegung der Sonderzeichen auf dem Dezimalblock und Anpassung der Tasten mit Tastaturklebern

So gut diese Lösung mit der damaligen Konfiguration aus Gerät, Betriebssystem und Software auch funktionierte, so anfällig war sie für Probleme bei jeglichen Veränderungen dieser Ausgangskonfiguration. Diese schmerzliche Erfahrung musste man bereits im Jahre 1992 machen. Im entsprechenden Jahresbericht wird von einem Update des Mac-Betriebssystems von Version 6.07 zur Version 7 geschrieben, das zu massiven Problemen beim Ausdrucken führte. Nach umfangreichen Abklärungen entdeckte man, dass die Fettschrift mit dem neuen Betriebssystem nicht kompatibel war.⁵

Richtig dramatisch wurde es aber im Jahr 2006: Nach Anschaffung und Inbetriebnahme der neuen Mac-Generation, erstmals mit dem Betriebssystem OS X 10 ausgestattet, musste man mit Schrecken feststellen, dass die DRG-Schriften aus den Achtzigerjahren nun definitiv ausgedient hatten. Die Redaktoren konnten zwar wie gewohnt ihre Artikel mit Microsoft Word erfassen, aber beim Import in das Layoutprogramm QuarkXPress wurden die Sonderzeichen derart zerschossen, dass an eine manuelle Überarbeitung nicht zu denken war. So wurde zum Beispiel ein bis anhin freier Platz in der ASCII-Tabelle mit dem Eurozeichen belegt, was zur Folge hatte, dass während des Imports zahlreiche phonetische Zeichen durch Eurozeichen ersetzt wurden. Mit einem eher unkonventionellen Workaround konnte der Publikationsrhythmus des DRG zwar beibehalten werden, aber diese Erfahrung führte zur endgültigen Erkenntnis, dass sich der computergestützte Redaktionsprozess so weit wie möglich von kommerzieller Software würde lossagen müssen, um nicht in Kürze vor dem nächsten Abgrund zu stehen.⁶

5 Cf. AnSR 1993: 339 [Jahresbericht SRR 1992].

6 Cf. AnSR 2007: 300f. [Jahresbericht SRR 2006].

Projekt «Digitales Wörtermuseum»

Dieses massive Problem, das die DRG-Publikation ernsthaft bedroht hatte, war dann auch der Startschuss zu einer umfassenden Analyse der kompletten EDV-Infrastruktur, also der gesamten Hard- und Software. Für die Behebung der Mängel wurde, nach einiger Vorbereitungszeit, ein Gesamterneuerungskonzept mit dem Arbeitstitel «Digitales Wörtermuseum» erstellt. Als Ziel wurden nichts weniger als die Retrodigitalisierung und die Online-Publikation des DRG sowie die Einführung eines datenbankbasierten Redaktionssystems erklärt, zudem sollten so viele Standardkomponenten wie nur möglich eliminiert werden, um in Zukunft erneute Probleme zu vermeiden.⁷

In einem ersten Schritt wurden die alten DRG-Schriften aktualisiert und in zeitgemässe Schriften im True-Type-Format umgewandelt; in diesem Prozess wurden sämtliche Zeichen gemäss dem Unicode-Standard neu belegt. Die DRG-spezifischen Zeichen wurden im Private-Use-Block abgelegt (Hodapp 2011: 3f.) Diese erste Massnahme rettete die DRG-Publikation, zudem sollte sie für die spätere Erstellung eines DRG-Redaktionssystems von zentraler Bedeutung sein.

Die Entwicklung des neuen Redaktionssystems fing im Jahr 2011 an. 2013 wurde es fertiggestellt und in Betrieb genommen. Hingegen wurde entschieden, das DRG-Online in einem eigenen Projekt umzusetzen.⁸

Im Folgenden wird auf das Redaktionssystem und natürlich vor allem auf das DRG-Online näher eingegangen.

7 Das Projekt wurde im Jahresbericht SRR 2009 angekündigt (AnSR 2010: 358).

8 Cf. AnSR 2014: 228 [Jahresbericht SRR 2013]: «Terminaziun da noss nov sistem da redacziun ... Ultra da quai furma il nov sistem da redacziun era la basa per la retrodigitalisaziun e per la publicaziun online dal DRG cumplet. La realisaziun dad in DRG online vegn dentant instradada pli tard en il rom dad in agen project» [Fertigstellung unseres neuen Redaktionssystems. Zusätzlich bildet das neue Redaktionssystem die Grundlage für die Retrodigitalisierung und für die Online-Publikation des DRG. Das DRG-Online wird jedoch später, im Rahmen eines eigenen Projektes, realisiert; Übersetzung UL].

Das neue Redaktionssystem

Das neue Redaktionssystem sollte idealerweise datenbankbasiert sein und unabhängig von kommerzieller Software betrieben werden können. Zudem sollte es langfristig haltbare und archivierbare XML-Daten liefern. In die Konzipierung des neuen Redaktionssystems flossen auch Überlegungen für eine spätere Online-Publikation der damit redigierten Artikel ein.

Ziel war es also, ein neues Redaktionssystem einzuführen, mit dem die erfassten Artikel nicht nur für die gedruckte Buchversion, sondern auch für die Online-Publikation verwertet werden können.

In enger Zusammenarbeit mit der Firma edp-services ag in Kriens entschieden wir uns nicht für einen traditionellen XML-Editor, sondern für eine für Lexikografen sicherlich angenehmere Arbeitsoberfläche. Die Grundidee für die im Webbrowser betriebene Applikation folgt dem Baukastenprinzip: Dem Redaktor stehen sämtliche für die Erstellung eines Artikels erforderlichen Bausteine zur Verfügung, die er mit dem gewünschten Inhalt füllen kann. Je nach Eingabefeld weiss das System, in welcher Schrift das erfasste Textelement zu erscheinen hat, im Hintergrund und für den Redaktor unsichtbar wird eine XML-Datei angelegt und fortlaufend gespeichert. Zur Artikelstrukturierung lassen sich die erfassten Inhaltsbausteine in Ordnern, sogenannten Struktursteinen, zusammenfassen. Zudem kann die Reihenfolge der Bausteine und Struktursteine per Drag-and-Drop verändert werden; so lassen sich die Artikel bequem korrigieren und umgestalten.

Um den Arbeitsfortschritt zu überprüfen, kann der Redaktor jederzeit eine Artikelvorschau aufrufen, die dem Endlayout relativ nahekommt, zudem können jederzeit einzelne oder mehrere Artikel als PDF-Datei oder als XML-Datei exportiert werden. Die Ausgabe als PDF wird vor allem zum Korrekturlesen verwendet, die XML-Daten können einerseits für die Weiterverarbeitung ins Layoutprogramm Adobe InDesign importiert werden, um die Artikelreihen für den Buchdruck vorzubereiten, andererseits für die Online-Publikation.

The screenshot shows the web interface of the Dicziunari Rumantsch Grischun (DRG) in a browser window. The browser address bar shows the URL 'http://www.drgr.ch' and the page title 'Dicziunari Rumantsch Grischun'. The interface is in German and displays the following elements:

- Navigation tabs:** 'Start', 'Meine Artikel', 'Alle Artikel', 'Meine Verweise', 'Alle Verweise'.
- Article Information:** 'Schachtel Nr.: 366b', 'Lemma: MIQUIR'.
- Article Structure:** A text input field containing 'Start | Artikelkopf'.
- Existing Structure Elements:** A list of elements to be added or removed, including:
 - MIQUIR *m. m. n. n. n.*
 - Lehnwörter: Pächter, Mieter; Halbpächter; Alpengasse
 - C 20 *m. n. n.*
 - C 30 *m. n. n.*
 - C 41, 44, 47–58 *m. n. n.*
 - C 62 *m. n. n.*
 - C 75, 83 *m. n. n., m. n. n.*
 - C 81, 88–91, 93 *m. n. n.*
 - S 1–6 *m. n. n.*
 - S 66 *alle m. n. n.*
 - S 7 *m. n. n.*
 - *W. n.*
 - Lexik. III.
 - m. n. n., m. n. n.*
 - Lexikon 91
 - m. n. n., Alpengasse*
 - Lexik.
 - Alpengasse, Alpengasse, der Halb-Pächter*
 - m. n. n.*
- Article Options:** A panel with three buttons: 'Vorschau', 'Speichern', and 'Löschen', each with a radio button.
- New Content Elements:** A vertical list of content building blocks:
 - Strukturstein
 - Notiz
 - Folienmappe
 - Freie Standard
 - Lexikonkürzel Standard
 - Synonymverzeichnis Standard
 - Hauptlemma invariant (1)
 - Sublemma Standard
 - Sublemma invariant
 - Bedeutung
 - GI-Kategorie
 - Phonetische Form
 - Einleitung Wörterbuchbeleg
 - Wörterbuchbeleg
 - Verteilung

At the bottom of the page, there is a copyright notice: '© 2013 by Institut für Dicziunari Rumantsch Grischun, Ringstrasse 34, CH-7000 Chur/Coira info@drgr.ch | http://www.drgr.ch'.

Abb. 2. Das Redaktionssystem des DRG – links die bereits erfassten Bausteine, rechts der «Baukasten» mit den zur Auswahl stehenden Inhaltsbausteinen

The screenshot shows the web interface of the 'Dicziunari Rumantsch Grischun'. The main content area displays the article preview for the word 'MIGIUR'. The article is structured as follows:

MIGIUR m., suedi. (Lehmann) Pflücker, Meier; Halbpflücker; Alghosses: Malinszelbesitzer, Malinszelbesitzer. C 20 *migür*, C 30 *migür*, C 41, 44, 45-48 *migür*, C 62 *migür*, C 74, 83 *migür*, *migür*, C 84, 88-91, 93 *migür* S 1-6 *migür* S 66 *üter migür*, S 7 *migür*. - Wh: Gots. DB: *migür*, Meier; Cassin 91 m. d'uf, Alghosses; Com: il *migür*, il *migür*, der Halbpflücker; Bern, Goms. 123 il *migür*, il *migür*, il *migür*, il *migür*, Vaxx, Vos, m., Pflücker, Halbpflücker; Vos, suedi, BD, m., Pflücker, Meier; Vos, Surs. *migür*, Alghosse, Malinszeller d'ostel Malinszeller, Pflücker, Yersulter; Vos, suedi, *migür*, Pflücker.

1. 'Lehmann, Vassil, Vogt' (hist.) Vgl. algh. → *migür* (H. 527, Abs. D. → *mür* (H. 210, Abs. D.) - Lit.: C: Igl *uocher muer* (Schub), *uocher* il *central* della *possessione agricola* in *Sares*, Igl *fu*, *müer* *igür* *migür* e *müer* *igür* ... *con* *de* *par* *schöin*, der *agrarer* (Schub), Ah, die Zentrale der ländlichen Besitzes im Oberhalbstein, der Ort, an welchem die von Bischof eingesetzten Lehensleute den Zins zu entrichten hatten (Ann. 20, 231). *Clere* *uocher* *inductus* *mür* *in* *die* *igür* *vaxx*, *con* *de* *un* *par* *schöin* *de* *un* *migür* *inductus*, diese Rechte setzen auf ihren alten Höfen wie auch auf den neugegründeten deutschen Lehenshöfen ein (Salen 191, 21). - Lit.: S: *En* *il* *trou* *mür* *de* *la* *possession* *igür*, *par* *passer* *igür* *con* *de* *un* *migür* *honoris* *con* *un* *clere*, *il* *par* *honor* *igür* *de* *un* *igür* *inductus* *migür* *uocher* *et* *con* *müer* *honoris* *dependenti* *de* *un* *Sigur*, im Mittelalter herrschte der Adel, wenige Mächtige lehensleihen ihre Lohndienste dem König. Das König kam damals noch selbst etwas an Besitz erstanden und war so glücklich von seinem Lehensherrn abhängig (Grischa Ann. 1839, 96, 183.2). *Le* *cas* *d'* *hospitium* / *Posses* *de* *Sigur*. (*Le* *cas* *de* *un* *migür*) *Con* *pluris* *posses* (*Ne* *uocher* *migür*, der Herr in Hospital geht dem Herrn, der hant sein Vassal mit absoluter Macht über unter Vgl. (H. 101, Goms 3, 151), *Alle* *trou* *della* *dotigra* *dil* *müer* (de *l'igür*) *stanz* *con* *uocher*, *igür* *uocher* *il* *migür* *uocher*, *de* *un* *il* *müer* *della* *Jappa*, *hant*, nach an der Spitze des Stadtgerichtes stand zuerst, wie schon bemerkt, der herkömmliche Vogt, später der jeweilige Anwesen der Graf (Ann. 53, 100). *Allegia* *de* *un* *con* *de* (*uoch* *igür* *uocher* *de* *Com*) *con* *migür* *e* *un* *uocher* *de* *un* *uocher* *de* *uocher*, in Finn unten hatte er über Bischof von Chur seine Vögte und seine Lehenleute und er selber gehörte dem Adel an (Hörsig, Pap. 2, 39).

2. 'Pflücker, Meier, Gutbesitzer, pachtweise Inhaber, Geschäftsführer, der einen Betrieb bewirtschaftet und verwaltet', belegt für Sars, Mal, Aulder, Pars, Schel, Bains, Bains, Trin, in S. Algh. *uocher*. - Alghosses: *Andre* *igür* *de* *müer*, *Bis*, *con* *de* *migür*, Pflücker, Bains *de* *migür*, *uocher*, - Lit.

Abb. 3. Vorschau eines bereits erfassten Artikels

Das DRG-Online, ein Projekt in drei Etappen

Als im Jahr 2015 die Vorbereitungsarbeiten für das DRG-Online anliefen und erste Skizzen gemacht wurden, mussten wir zuerst feststellen, dass die vorhandenen Daten der mit dem Computer redigierten Bände 9–13 für die Online-Publikation nutzlos waren. Die fertig redigierten und publizierten Artikel lagen einerseits als QuarkXPress-Dateien vor und andererseits als PDF-Dokumente. Da diese Daten nur Informationen zur Darstellung der einzelnen Artikelelemente, nicht aber zu ihren jeweiligen Funktionen enthielten, war es nicht möglich, die Artikel sauber in eine strukturierte Datenbank zu importieren. Demnach musste eine Lösung für die Retrodigitalisierung aller noch nicht im neuen Redaktionssystem erfassten Artikel gefunden werden.⁹

Retrodigitalisierung der DRG-Bände 1–13

Da das Schriftbild des DRG sehr komplex ist, sich durch zahlreiche Schrift- und Stilwechsel innerhalb der Artikel auszeichnet und sich über die vielen Jahrzehnte sogar leicht verändert hat, kam Einscannen und anschließendes Digitalisieren mit OCR-Software nicht in Frage. Zudem waren die Anforderungen, die von Beginn weg an das finale DRG-Online gestellt wurden, sehr hoch und setzten eine strukturierte Datenbank voraus, in der die einzelnen Artikelelemente erkennbar und somit auch auswertbar sind. So wurden wir relativ schnell auf das der deutschen Universität Trier angegliederte Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften aufmerksam. Bei der Prüfung der Trierer Methode stellte sich sehr bald heraus, dass wir nicht nur zu einer hochpräzisen, fast fehlerfreien Volltextdigitalisierung unserer DRG-Bände gelangen würden, sondern dass auch detaillierte, von spezialisierten Computerlinguisten erstellte TEI-konforme XML-Daten hergestellt würden. Als Vorteile der Volltextdigitalisierung nennt das Kompetenzzentrum in Trier die folgenden Punkte:

- Volltextdigitalisate ermöglichen den Zugriff auf jedes einzelne Zeichen eines Textes;

9 Die erste mit dem neuen Redaktionssystem erfasste Artikelreihe umfasst die Wortstrecke MESSADAT – METTENT und wurde von Ursin Lutz redigiert (DRG 14, 339–380, Publikation im Jahr 2016). Ab dem Stichwort MICHEL (DRG 14, 640f., Publikation in den Jahren 2018/2019) wurden alle Artikel mit dem neuen Redaktionssystem erfasst.

- die Menge an zu verwaltenden Daten ist geringer als bei Imagedigitalisaten;
- der Aufbau von Indizes oder Metadaten ist in vielen Fällen automatisch durchführbar;
- nur Textdigitalisate ermöglichen ein barrierefreies Internet.

Für die Texterfassung kooperiert das Kompetenzzentrum mit dem chinesischen Erfassungsbüro TQY DoubleKey in Nanjing, da die chinesischen Datentypistinnen aufgrund der Komplexität und Feingliedrigkeit ihrer eigenen Schrift auch feinste Schrift- und Zeichenunterschiede erfassen und als Nichtmuttersprachlerinnen keine ungewollten «Verbesserungen» der Vorlage vornehmen.

In differenzierten Erfassungsanweisungen, die am Kompetenzzentrum in Trier vor dem Abtippen in China erstellt werden, wird anhand von Beispielen aufgeführt, wie die verwendeten Alphabete und Sonderzeichen sowie die verschiedenen bedeutungstragenden typografischen Besonderheiten und Layoutmerkmale bereits bei der Erfassung zu kennzeichnen sind. Um Informationsverluste zu vermeiden, wird dieses Regelwerk ins Chinesische übertragen.

In zwei unabhängig voneinander arbeitenden Teams fertigen die chinesischen Datentypistinnen jeweils eine vollständige elektronische Abschrift der Texte an. Dabei werden die typografischen Merkmale wie Kursivierung, Sperrung, Hoch- und Tiefstellung und Schriftgrößenwechsel gemäss den Erfassungsanweisungen durch eindeutige Codierungen gekennzeichnet. Ebenso werden Zeilen-, Spalten- und Seitenumbrüche markiert. Dieses sogenannte Character und Page Encoding führt zu einer ausgabendiplomatischen Abschrift der Vorlage. Die Codierung der Sonderzeichen richtet sich so weit wie möglich nach den Tustep-Konventionen.

Nach der Erfassung werden die beiden Eingabeversionen automatisch miteinander verglichen. Wie bei vielen anderen Arbeitsschritten auch kommt dabei die speziell für EDV-philologische Zwecke entwickelte Tübinger Software Tustep zum Einsatz. Mittels eines Tustep-Programmmoduls wird ein Vergleichsproto-

koll generiert, das im Kompetenzzentrum in Trier kontrolliert wird. Ergebnis ist eine Textversion mit einer Genauigkeit von bis zu 99,997 Prozent, das heisst, auf 100 000 Zeichen sind nicht mehr als drei Fehler zu erwarten.¹⁰



Abb. 4. Schematische Darstellung der Arbeitsschritte von der Retrodigitalisierung der gedruckten DRG-Bände zur Website

10 Dieser Abschnitt orientiert sich stark an der Website des Kompetenzzentrums: <https://www.kompetenzzentrum.uni-trier.de/de/schwerpunkte/volltextdigitalisierung/>.

TEI-konforme XML-Auszeichnung der elektronischen Abschrift

In einem zweiten Schritt wurden die in China erfassten Textdaten in Trier analysiert und unter Einhaltung der TEI-Normen in XML-Daten konvertiert. Dabei wurden sämtliche für das Layout relevanten Informationen wie Abstände, Einzüge, Schriften und Schriftgrösse in XML-Tags hinterlegt wie auch sämtliche für die Artikelstruktur inhaltlich relevanten Informationen analysiert und erfasst. Zudem wurden alle Sonderzeichen nach dem Unicode-Standard erfasst. Bei den DRG-spezifischen Sonderzeichen einigte man sich auf eindeutige Codierungen, um Eristere in der Webdarstellung konsistent darstellen zu können.

Dieses Vorgehen erlaubt einerseits eine saubere und konsequente Darstellung der einzelnen DRG-Artikel als Website, andererseits aber auch die Erstellung einer Datenbank für linguistische Abfragen und Auswertungen sowie Zusammenstellungen von spezifischen Listen, die zum Beispiel bei der Erstellung von aktualisierten Abkürzungs- oder Literaturlisten viel Arbeit ersparen können.

Erstellung eines Webportals als Benutzer-Frontend

Im letzten Schritt auf dem Weg zum DRG-Online wurde in Zusammenarbeit mit der edp-services ag ein Webportal als Benutzer-Frontend erstellt. Aus Sicht des DRG sollten folgende Anforderungen erfüllt sein:

- Eine akkurate Darstellung der DRG-Artikel inklusive Abbildungen und deren Legenden, die den gedruckten DRG-Bänden so getreu wie möglich ist. Im Gegensatz zur gedruckten Version soll aber eine Standardschrift verwendet werden, um grösstmögliche Lesbarkeit und Durchsuchbarkeit des Textes sowie Plattformunabhängigkeit zu garantieren;
- keine inhaltlichen Vereinfachungen aufgrund von technischen Einschränkungen;
- einfacher und intuitiver Zugang, auch ohne DRG-spezifische Vorkenntnisse;

- Suche mit deutschen und romanischen Stichwörtern, in allen Idiomen und lokalen Ausprägungen;
- Interaktivierung der internen DRG-Synonymverweise sowie der Inhaltsübersichten bei längeren Artikeln;
- einfache Orientierung in den DRG-Bänden und Zitiermöglichkeit im wissenschaftlichen Sinne;
- Ausgabe der Artikel als PDF;
- Integration von Abkürzungs- und Literaturverzeichnissen.




Herausforderungen

Als wir uns an die Umsetzung dieses Vorhabens machten, merkten wir sehr schnell, dass die Aufgabe nicht leicht sein würde und dass wohl oder übel auch einige Abstriche aus technischen Gründen in Kauf genommen werden müssten.

Eine der grössten Herausforderungen war die Heterogenität der DRG-Artikel, die in einem Zeitraum von 75 Jahren entstanden und gedruckt worden waren (1939–2014). So zeigte sich nicht selten, dass in früheren Zeiten vereinzelt mit der Artikelstruktur sehr frei umgegangen worden war oder aber dass einige Lemmata eine unorthodoxe Darstellung des Materials verlangt hatten. Diese Tatsache hatte bereits bei der XML-Konvertierung in Trier Kompromisse zur Folge gehabt und jetzt, bei der finalen Darstellung, erst recht. So kam es immer wieder vor, dass die aktuelle Importkonfiguration nur einem Teil der DRG-Artikel gerecht wurde, bei anderen wurde die Darstellung wieder verschlechtert.

Auch bei diesem Schritt sollten die Sonderzeichen uns vermehrt Kopfzerbrechen bereiten; die Komplexität der DRG-spezifischen Zeichen zeigte sich einmal mehr, trotz der sauberen Erfassung nach dem Unicode-Standard. Nach bestem Wissen und Gewissen wurden hier so wenige Zugeständnisse wie möglich gemacht.

Einige Beispiele hierfür soll diese Tabelle aufzeigen:

Buchform	DRG-Online	Beispiel
	č	CHARBUN (DRG 3, 364)
	[fund]	MALAGUARDÀ (DRG 12, 193)
	x	CALISCH (DRG 3, 20)

Im Fall [fund] entschieden wir uns für die Auflösung des Zeichens, in erster Linie, weil das ursprünglich verwendete Zeichen für die alte Währung Pfund einem breiten Leserkreis unbekannt ist.

Schliesslich erwies sich die Interaktivierung der internen DRG-Synonymverweise als problematisch, und zwar aus mehreren Gründen. Auf der einen Seite weisen vor allem die Synonymverweise in den älteren DRG-Bänden, die auf Folgeartikel verweisen, eine markante Ungenauigkeit auf, die zu einem späteren Zeitpunkt für das DRG-Online verbessert werden muss. Auf der anderen Seite stellten wir fest, dass die Trierer XML-Auszeichnung bei diesem Detail nicht 100 Prozent eindeutig war und Varianten aufwies, was die Aufgabe zusätzlich erschwerte. Ähnliches galt für die Inhaltsübersichten: Auch in diesem Hinblick stellten sich die XML-Daten für eine automatisierte Interaktivierung als unzulänglich heraus.

Lösungsansätze

Wenn man mit 15 323 DRG-Artikeln bzw. fast 71 Millionen Zeichen konfrontiert ist, ist man sehr gut beraten, so viele Arbeitsschritte wie möglich auf automatisiertem Weg zu lösen. Dies gelang zu einem sehr grossen Teil; wo wir aber keine einfache Lösung für die Darstellung von einigen Details sahen, entschieden wir uns zusammen mit den Programmierern der edp-services ag für den pragmatischen Weg: Wir bekamen die Möglichkeit, selber XML-Daten zu verändern und diese direkt in der Live-Datenbank auf dem Server zu ersetzen. Damit konnten wir auf unkomplizierte Weise, in enger Zusammenarbeit zwischen Lexikografen und Informatikern, einzelne Probleme doch noch auf technischem Weg lösen. Darüber hinaus stand es uns Lexikografen offen, manuelle Änderungen vorzunehmen. So erfassten wir für alle Inhaltsübersichten (cf. zum Beispiel MANTEL, DRG 13, 33) die Zeilenumbrüche und die Einzüge; bei grösseren Verbartikeln (cf. LASCHAR, DRG 10, 498) fügten wir für die Darstellung der Formenübersichten Tabellen in die XML-Daten ein.

Für den einfacheren Zugriff auf die einzelnen Artikel wird in der Suchmaske für jeden Artikel eine deutsche Definition angezeigt; auch diese wird grundsätzlich automatisch generiert. Gerade bei den Artikeln aus den älteren DRG-Bänden funktionierte das überhaupt nicht wunschgemäss; so führten wir kurzerhand einen neuen XML-Tag ein und erfassten für zahlreiche Artikel diese Definition von Hand. Um sicherzustellen, dass die Formen in allen Idiomen und Dorf-mundarten gesucht werden können, hinterlegten wir die Einträge aus unserer über Jahrzehnte gepflegten FileMaker-Datenbank, die alle Subformen zu den verschiedenen Lemmata enthält. Auch für die Erstellung des Abkürzungs- und Literaturverzeichnisses konnten wir uns auf die Dienste unserer angestaubten FileMaker-Datenbank verlassen.

Fertigstellung und Inbetriebnahme der ersten Version

Nachdem wir mit der Darstellung der DRG-Artikel und dem Funktionsumfang der ersten Version des DRG-Online zufrieden waren, konnte die Inbetriebnahme und öffentliche Vorstellung nach über drei Jahren Vorbereitungszeit endlich ins Auge gefasst werden. Dafür wurde von der Societad Retorumantscha und dem Institut dal DRG ein Festakt geplant. Im Rahmen dieses Anlasses wurde das DRG-Online mit seinen Funktionalitäten erstmals vorgestellt.

Festlicher Akt vom 7. Dezember 2018

Um die Online-Publikation des DRG als historischen Moment gebührend zu markieren, organisierten die Societad Retorumantscha (SRR) und das Institut dal DRG am 7. Dezember 2018 im Auditorium der Graubündner Kantonalbank in Chur einen Festakt mit Ansprachen (Dr. Cristian Collenberg, Präsident SRR; Dr. Carli Tomaschett, Chefredaktor DRG; Dr. Ursin Lutz, Redaktor und Projektleiter DRG; Rolf Stegemann, Leiter Entwicklung edp-services ag, Kriens) sowie Grussworten der SAGW (Dr. Manuela Cimeli), des Kantons Graubünden (Barbara Gabrielli, Leiterin des Amtes für Kultur Graubünden) und der Freien Universität Bozen (Dr. Giovanni Mischì) sowie einem humoristischen Beitrag des Vizepräsidenten der SRR, Chasper Pult.

Zudem fand anlässlich der Veranstaltung die Premiere des RTR-Films «Made in China – la digitalisaziun dal Dicziunari Rumantsch Grischun» statt, eines Films von Bertilla Giossi.¹¹

Funktionalität und Funktionsweise der ersten Version

Für unsere erste Version stellten wir vor allem eine saubere Darstellung der Artikel in den Vordergrund und beschränkten uns auf eine einfache Suche, die aber intuitiv sein und ohne grosse Anleitungen und Einführungen auskommen sollte. Gelten die DRG-Bände häufig als Bücher mit sieben Siegeln, soll hier der Zugang so direkt wie möglich erfolgen. Für die gedruckte DRG-Variante wird

11 Der halbstündige Dokumentarfilm, der am 9. Dezember 2018 erstmals in der Filmreihe «Contrasts» auf SRF 1 ausgestrahlt wurde, ist online verfügbar.

The screenshot shows a web browser window with the address bar displaying 'drg-online.ch'. The page title is 'DRG-Online - CUCCALORI'. The header features the logo of the 'Dizionario Rumantsch Grischun' and navigation links: 'Institut | Über das Wörterbuch | Angebot | Publikationen | Shop | Poltava Online | DRG-Online'. A search bar is located in the top right corner.

On the left side, there is a sidebar with a search filter set to 'DRG-ONLINE'. Below this, there are sections for 'Neue Suche', 'Resultate: dummkopf', 'Suchtipps', 'Abkürzungen', 'Vorwörter', and 'Hinweise'. A vertical list of letters from A to M is provided for navigation.

The main content area shows the word 'CUCCALORI' with a '+ zurück' link. Below the word, there are two buttons: 'DRG 04 / 312' and 'Artikel als PDF laden'. A note indicates: 'Markieren Sie einfach grosszügig eine Textstelle, um die Band- und Seitenzahl anzuzeigen'. The main text of the article reads: 'CUCCALORI m., sursch. "Dummkopf, Tölpel, einfältiger Mensch". S. Aschaffner: C 41, 89 *hubschfiri*. – Wb.: seit Cassi. *exccalori*. – Röm. *ex buxg excalori*, ein guter, gutmütiger Kerl. Pigniu *in cuccalori dalla bia'aura*, ein Müsiggänger. Dik: *Fusper (stet) cuccalori! Severat buca z'urchivà ampax maglier?* dummt'er Kerl, könntest du es nicht ein wenig geschickter anstellen? – Lit. S: *Ma nel por viaz fatg; cax ie in! cuccalori marcadet jeu huc*, schert euch weg; mit einem solchen Tölpel marke ich nicht (Ann. 60, 124). *Rasba ... rena perdets il pli grond cuccalori*. Vermögen macht des grössten Dummkopf geschick (Tschespet 12, 107, *Muonc*; zahlr. weitere Belege).

Below the main text, there is a paragraph: 'Unter den Bezeichnungen für "Dummkopf" eignet *cuccalori* *schafiri* fast ausschliesslich dem Surselvischen. *agpavri* dagegen wird, jedoch bedeutend seltener, auch in C und E gebraucht. Ausgangsform für diese Sippe scheint *excavari* zu sein; es ist nicht zu trennen von schw.ä. *Gaggelari* "Tölpel" (Schw. M. 3, 1362). *Gaggelari* (Bd. 3, 364) *bat-rin*, *Gaggelari*, *Gogelari*, *Gogelari* (cf. «Der Gaggelore» von Otto RUTIMUS, Biederstein Verlag 1963, wo der Name einen zwerghaften, zu allerlei Possen und Unflat bereiten Dämon bezeichnet). Es scheint sich also um eine im süddeutschen Raum entstandene Bildung zu handeln. In Romanischbüchern vermischte sie sich mit *cax* in gleicher Bedeutung.

At the bottom of the article, there is a section 'A. Schorta'.

Abb. 5. Der Artikel CUCCALORI im DRG-Online

DRG-Online - MANTIGLIA

online.dgzh

Suche mit Bing

DRG-Online
 Wörterbuch
 Romanisch
 Gröschlun

Institut | Über das Wörterbuch | Angebot | Publikations | Shop | Petrarca Online | DRG-Online

DRG-Online

Neue Suche

Resultate: 1 result

Suchtipps

Abkürzungen

Vorwörter

Hilfsseite

A

B

C

D

E

F

G

H

I

J

K

L

M

Zurück

MANTIGLIA

[DRG 13 / 81] Artikel als PDF laden

Interpretieren Sie diesen geschichtliche Textstelle, um die Band- und Seitenreferenzen zu finden

MANTIGLIA ungd., **MANTIGLIA** ungd., f. 'Mantiglia, Schürzenrock', *Petrarca*: II 10, 13, 15, 20, 22–23, 25, 40, 43, 49, 50, 53–55 *manila*; II 21, 40, 42 *manila*; C 10–11, 62, 67, 81, 84, 92 *manila* (C 83 auch *manila*); C 23–24, 27, 31, 41, 46–47, 92 *manila*; S 11, 25–26, 47, 61 *manila* (S 25 auch *manila*); S 34 *manila*. – Wb.: C 100; *manila*, Frauenmütze; P 111, *maniglia* (Eo.), Frauenmütze; *manila* (Eb.), Mann; Vb.: *manila*, RD, *manila*, Mütze, Schabernack, kleines Frauenmütze; Vb.: *manila*, *Petrarca*, *Petrarca*, *Petrarca* Überwurf des alten Hüftenstetels; Vb.: *manila*, *Petrarca*, *Petrarca*.




Abb. M 193: Mann mit Rückentragkorb, Hirtenkrone mit *Petrarca*, Vale um 1920 (Foto W. Derichswiler)

DRG-Online - MANTIGLIA

Zurückfahren

Abb. 6. Ein Ausschnitt des Artikels MANTIGLIA im DRG-Online

das Idiom des Unterengadins (Vallader) als Leitsprache für die Festsetzung der Lemmata verwendet; hier soll jeder Bündnerromane, vom eigenen Idiom oder sogar von der Dorfmundart ausgehend, die gewünschten Artikel finden. Vor allem aber sollen die DRG-Artikel auch über deutsche Suchbegriffe auffindbar sein. Um zum Beispiel zum Artikel LÜNDESCHDI «Montag» zu gelangen, kann man demnach ins Suchfeld *gliendisdis* (Sursilvan), *glenderschis* (Vaz), *lindasde* (Marmorera) oder deutsch *Montag* eintippen.

Ist ein Artikel geöffnet, kann dieser jederzeit als PDF-Dokument gespeichert werden und mit Hilfe der Suchfunktion des verwendeten Browsers durchsucht werden. Um die Band- und Seitenzahl einer Textstelle zu erfahren, kann einfach etwas Text markiert werden, was eine entsprechende Einblendung zur Folge hat.

Die Suche nach deutschen Wörtern hat sich zur Zusammenstellung und Auffindung von Synonymen bestens bewährt; in Sekundenschnelle findet man zahlreiche Begriffe mit der gleichen Bedeutung. Ein beliebtes Beispiel dafür ist die Suche nach dem Begriff *Dummkopf*, bei der nicht weniger als 23 romanische Ausdrücke angezeigt werden, darunter *cuccalori*, *calöri*, *gnahà*, *lali* und *ma-mau*.

Die aktuell online publizierten DRG-Bände 1–13 umfassen die Wortstrecke von A – MEDGIAR; von daher könnte man meinen, dass keine Artikel zu finden sind, die mit Buchstaben anfangen, die im Alphabet nach dem M kommen. Das gilt auf jeden Fall für die Lemmaform, die nach dem unterengadinischen Idiom angesetzt wird. Nun spielen hier vor allem einige surselvische Lautgesetze und orthografische Regeln eine wichtige Rolle, zudem kennt das Bündnerromanische lautliche Entwicklungen, die zu vielfältigen Resultaten führten. Mithilfe der zahlreichen Verweise findet man auch Einträge zu Formen wie *penda* («Band», → BENDA I), *rida* («Kreide», → CRAIDA), *schambun* («Schinken», → DSCHAMBUN), *tgau*n («Hund», → CHAN), *uaul* («Wald», → GOD), *vegnir* («kommen», → GNIR) und *zafalet* («Taschentuch», → FAZÖL).

Was natürlich nicht fehlen darf, ist eine Volltextsuche, mit der man unkompliziert alle DRG-Artikel nach einem bestimmten Begriff oder etwa einer Redewendung wie *die Flinte ins Korn werfen* suchen kann. Dadurch werden die romanischen Entsprechungen *better l'arma ella fletga* (Dardin, wörtlich «die Waffe ins

Farnkraut werfen», Artikel BETTER und FAISCHEL), *fierer la faultsch el canvau* (Dardin, wörtlich «die Sense in die Mahd werfen», Artikel CHANVÀ und FIERER), *fierer il bandun* (Vrin, wörtlich «den Zapfen des Troges werfen», Artikel BANDUN I), Suts. *fierer igl moni tschandervei* (Sutselva, wörtlich «den Stiel hinwerfen», Artikel MANCH I) mit Leichtigkeit und in Kürze aus den verschiedenen DRG-Bänden zusammengetragen.

Ausblick

Hat man es mit einem monumentalen Werk wie dem DRG zu tun, das fortlaufend publiziert wird und jetzt auch online verfügbar ist und auf einer enormen Datenbank beruht, ist die Arbeit nie fertig. Nach der Lancierung einer ersten Version wurde demnach direkt an der Einpflege der neuen Artikel gearbeitet; zudem liegen bereits erste Ideen für neue Funktionalitäten vor.

Fortlaufende Einpflege der publizierten Artikel

Zunächst ging es darum, für die während der Projektphase des DRG-Online publizierte Wortstrecke (MEDI – MICHA I; DRG 13, 1–640) eine Lösung zu finden. Da diese Artikel, mit Ausnahme der Artikel MESSAGERA – METTEL (DRG 14, 339–390), alle noch nach dem alten Verfahren mit Microsoft Word erfasst wurden, werden sie nach der gleichen Methode wie Band 1–13 in China abgetippt und in Trier zu XML-Daten weiterverarbeitet.

Die neueren Artikel wurden alle mit dem datenbankbasierten Redaktionssystem erfasst; so können diese in Zukunft ohne den Umweg über China direkt in die Datenbank des DRG-Online aufgenommen werden. Hier gilt es, die beiden Datenbanken mit einer neuen Export- bzw. Importfunktion nachzurüsten, die vor allem die Umwandlung der Sonderzeichen vornehmen kann. Damit ist langfristig für eine reibungslose Druckproduktion der DRG-Faszikel und der DRG-Bände sowie für eine unkomplizierte Online-Publikation der neuen DRG-Artikel gesorgt.

Erweiterung der Funktionalitäten

Als erste Erweiterung des DRG-Online sind eine Verbesserung der Erkennungsrate der zahlreichen Synonymverweise und eine Korrektur derselben angedacht, damit diese mit Hyperlinks zu den entsprechenden Artikeln hinterlegt werden können.

Die Darstellung der Suchresultate soll um einen kurzen Ausschnitt der Belegstelle erweitert werden, damit der Benutzer schon vor dem Aufrufen des entsprechenden Artikels prüfen kann, ob das Suchergebnis weiterzuverfolgen ist.

Zudem werden die technischen Möglichkeiten zur Ansteuerung der einzelnen Artikelabschnitte erörtert, im Idealfall mit der Interaktivierung der Inhaltsübersichten. Im gleichen Zuge werden die Optionen für die Erstellung von automatisierten Inhaltsübersichten für alle DRG-Artikel geprüft.

In den nächsten Jahren wird sicher die Konzipierung einer erweiterten Suche von zentraler Bedeutung sein. Nicht zu trennen ist diese von den zahlreichen und sehr detailliert vorhandenen DRG-Indizes, die am Schluss eines jeden DRG-Bandes erscheinen. Mit der Einbindung dieser Daten würde sich eine ganze Palette an Suchmöglichkeiten eröffnen:

- nach Wörtern mit dem gleichen Etymon oder die von der gleichen Sprache abstammen;
- nach Wörtern mit dem gleichen Suffix bzw. Präfix;
- nach Wörtern zu einem bestimmten Themenfeld;
- nach Wörtern, die gleiche lautliche Entwicklungen aufweisen;
- nach syntaktischen Elementen wie verbalen oder nominalen Verbindungen.

Gelingt es, diese Suchkriterien zu kombinieren, werden die Möglichkeiten der erweiterten Suche die Nutzbarkeit und Auffindbarkeit der zahlreichen DRG-Artikel beträchtlich erhöhen.

Literatur

AnSR = *Annalas da la Societad Retorumantscha* (1886-): Cuira.

Vor allem:

Rapport annual 1987 (1988): Societad Retorumantscha, in: AnSR, S. 228–233.

Rapport annual 1990 (1991): Societad Retorumantscha, in: AnSR, S. 157–163.

Rapport annual 1991 (1992): Societad Retorumantscha, in: AnSR, S. 231–238.

Rapport annual 1992 (1993): Societad Retorumantscha, in: AnSR, S. 335–341.

Rapport annual 1996 (1997): Societad Retorumantscha, in: AnSR, S. 253–264.

Rapport annual 2006 (2007): Societad Retorumantscha, in: AnSR, S. 293–308.

Rapport annual 2009 (2010): Societad Retorumantscha, in: AnSR, S. 351–366.

Rapport annual 2013 (2014): Societad Retorumantscha, in: AnSR, S. 221–238.

DRG = *Dicziunari Rumantsch Grischun* (1939-). Societad Retorumantscha, Cuira.

EDV-Einsatz 1991 = *EDV-Einsatz beim Dicziunari Rumantsch Grischun*, 24. Januar 1991 [zweiseitiges Daktyloskript, Institut dal DRG, Cuira].

Hodapp, Theresa (2011): *Dokumentation IDRГ Schriften*.

Tomaschett, Carli (2004): *100 onns Institut dal DRG. Retrospectiva e perspectivas*, in: AnSR 117, S. 1–24.

Made in China – Die Digitalisierung des Dicziunari Rumantsch Grischun, Dokumentarfilm, Giossi, Bertilla (Regisseurin): Switzerland, SRF 1, am 09/12/2018 in der Filmreihe «Cuntrasts» ausgestrahlt [<https://www.rtr.ch/play/tv/cuntrasts/video/made-in-china---mit-deutschen-untertiteln?id=9f6fade9-dcb5-4404-92ef-5b7eb8d752ad>] (konsultiert am 29.01.2019).

Website des Kompetenzzentrums für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften der Universität Trier
<https://www.kompetenzzentrum.uni-trier.de/de/schwerpunkte/volltextdigitalisierung> (konsultiert am 28.01.2019).

Im Übergang zum digitalen Wörterbuch. Zum Stand der Digitalisierung des Schweizerischen Idiotikons: Rückblick und Ausblick

Tobias Roth

Das Schweizerische Idiotikon blickt bis heute auf eine über 150-jährige Entstehungszeit zurück und war damit gezwungenermassen den grössten Teil der Zeit ein nichtdigitales Wörterbuch. Natürlich ist aber die Digitalisierung auch am Idiotikon nicht spurlos vorübergegangen. Seit den ersten Retrodigitalisierungsprojekten vor rund zehn Jahren ist das Idiotikon immer mehr zum Anbieter von digitalen Inhalten und zu einer Institution der Digital Humanities geworden. Aktuell nutzen etwas mehr als 50 000 Besucherinnen und Besucher pro Monat die Online-Angebote des Schweizerischen Idiotikons.

Die Digitalisierung des Schweizerischen Idiotikons ist jedoch noch längst nicht abgeschlossen. Es handelt sich um eine fortdauernde Aufgabe mit dem Ziel eines vollständig digitalen Wörterbuchs, das nicht nur digital publiziert, sondern auch digital bearbeitet wird. Was dazu noch fehlt, wird im letzten Teil dieses Beitrags angesprochen. In den ersten Abschnitten werden die bisherige digitale Geschichte des Schweizerischen Idiotikons sowie allgemein die angewendeten Digitalisierungsstrategien beschrieben.

Rückblick und Status quo

Website und Lemmeregister

Der erste Internetauftritt des Schweizerischen Idiotikons in den frühen Nullerjahren bestand aus einer Handvoll einfacher Informationsseiten auf einer Unterseite des Auftritts der Schweizerischen Akademie für Geistes- und Sozialwissenschaften (SAGW). 2008 ging das Schweizerische Idiotikon dann mit einer eigenständigen Website online. Diese war gegenüber der Vorgängerversion inhaltlich und funktional stark ausgebaut. Mit dem elektronischen und auch elektronisch durchsuchbaren Lemmeregister wurde ein erster Digitalisierungsschritt

des Wörterbuchs an die Öffentlichkeit gebracht. Suchbar sind nicht nur die Lemmata selbst, sondern auch Varianten davon. Da die Schreibung im Schweizerdeutschen nicht normiert ist, und die Formen in den einzelnen Dialekten zum Teil beträchtlich voneinander abweichen, ist auch damit zu rechnen, dass über ganz unterschiedliche Formen nach einem Lemma gesucht wird. Für die Suche wurden deshalb von allen Lemmata automatisierte Varianten erzeugt, die in erster Linie bekannte lautliche Phänomene abdecken. Der Algorithmus zeigt eine gewisse Übergenerierung, was aber nicht so stark ins Gewicht fällt, da die Varianten lediglich als Suchstruktur im Hintergrund genutzt werden. Teils werden Varianten auch manuell ergänzt.

Bis heute basiert die Digitalisierungsstrategie des Schweizerischen Idiotikons auf etappenweisem Vorgehen in kleinen Schritten, die sich sinnvoll ergänzen. Neuerungen können so rasch publiziert werden, und Erfahrungen mit den bereits publizierten Werkzeugen und Inhalten können in die Entwicklungsarbeit zurückfließen.

Auf der neuen Website wurden viele Inhalte rund um das Schweizerische Idiotikon publiziert: So Artikel zur Werkgeschichte, eine Bibliografie zum Schweizerdeutschen und neben dem Lemmaregister auch weitere Werkzeuge, die einem die Wörterbuchbenutzung erleichtern sollen, wie etwa Abkürzungsverzeichnis und ein Verzeichnis der Quellensiglen.

Digitales Faksimile

Als weiterer wichtiger Digitalisierungsschritt wurde das Wörterbuch eingescannt. Das Scannen erfolgte extern. Die Scans wurden dann mit dem bereits veröffentlichten elektronischen Lemmaregister verknüpft. Auf diese Weise führte eine Suche nach einem Stichwort über das Lemmaregister direkt zum Faksimile der entsprechenden Wörterbuchseite. Die Grundfunktionalitäten eines elektronischen Wörterbuchs waren damals – wenn auch mit noch eingeschränkter Funktion – bereits umgesetzt.

Zusätzlich wurde das grammatische Register, das ursprünglich handschriftlich auf Karteikarten erfasst worden war, digitalisiert und online publiziert, ebenfalls verknüpft mit dem bereits publizierten Lemmaregister. Es handelt sich beim

grammatischen Register um eine hierarchisch gegliederte Auflistung grammatischer Phänomene zusammen mit Beispielen von Wörtern, die den entsprechenden Phänomenen zugeordnet werden können. Das grammatische Register ist aber weder auf der Kategorien- noch auf der Zuordnungsseite als vollständig anzusehen, da es über weite Strecken impressionistisch als lexikografisches Arbeitsinstrument entstanden ist und nie den Anspruch der Vollständigkeit erhoben hat.

Automatische Texterkennung und Volltextsuche

Es folgte dann die Entwicklung einer Volltextsuche über den gesamten Wörterbuchtext. Da dieser so nicht vorlag, wurde er aus den Scans per OCR (Optical Character Recognition, also automatische Texterkennung) ausgelesen. Es handelt sich dabei um ein automatisches Verfahren, das nie komplett fehlerlos arbeitet und dessen Resultate sehr stark von Art und Qualität der Vorlagen abhängen.

Die Drucke und die Scans sind von guter Qualität. Aus OCR-Sicht positiv zu vermerken ist auch, dass als Druckschrift von Beginn weg eine Antiqua- und keine Frakturschrift verwendet wurde (automatische Texterkennung in Fraktur ist viel schwieriger und unzuverlässiger). Negativ auf die Erkennungsleistung wirken sich vom Druckbild her vor allem die vielen Hochstellungen für etymologisch vorhandene, in der modernen Sprache aber geschwundene Laute sowie die häufigen Formatwechsel zwischen kursiver, aufrechter und gesperrter Schrift aus. Inhaltlich gesehen erschwerend für die automatische Texterkennung sind die komplexe Textstruktur, die Mischung unterschiedlicher Sprachen bzw. Sprachvarietäten und die hohe Informationsdichte, wobei ein derartiger Text natürlich von einem Wörterbuch zu erwarten ist. Automatische Texterkennung betrachtet nicht nur die einzelnen Zeichen losgelöst vom Kontext, sondern bewertet potenziell erkannte Zeichen auch in ihrem Wort- und Satzkontext. Bei einem Wörterbuch wie dem Schweizerischen Idiotikon mit zum Beispiel Reihungen von Abkürzungen und Einzelformen, im Wechsel mit ganzen Sätzen (im Metatext und in Belegsätzen) können die mehrheitlich auf anderen Textsorten basierenden vortrainierten Kontextmuster nicht gewinnbringend eingesetzt werden. Ähnlich verhält es sich mit dem Nebeneinander von Standardsprache (vorwiegend

Metatext), unterschiedlichen Dialekten (in Belegen) und historischen Sprachstufen (ebenfalls in Belegen): Referenzwortlisten, mittels derer in der Texterkennung ein erkanntes Wort plausibilisiert werden könnte, existieren in der nötigen Qualität nur für die Standardsprache; für mundartliche und historische Textteile gibt es kaum Referenzwortlisten, die helfen würden. Im schlimmsten Fall produziert die standarddeutsche Wortliste sogar Interferenzen, nämlich dann, wenn Dialektwörter selber standardnah sind und deshalb vom Algorithmus falsch gelesen werden.

Die hohe Informationsdichte schliesslich hat zwei Hauptauswirkungen auf die automatische Texterkennung beziehungsweise den Digitalisierungsprozess. Auf der einen Seite führt sie dazu, dass sich die automatische Texterkennung kaum auf Redundanzen stützen kann. Selbst wenn man den Algorithmus explizit mit Teilen des Schweizerischen Idiotikons trainieren würde, wären daraus gewonnene Kontextinformationen aufgrund der fehlenden Redundanzen nur von beschränktem Nutzen für die Texterkennung. Die hohe Informationsdichte führt aber auch zu erhöhten Ansprüchen gegenüber dem Digitalisat. Denn auch für menschliche Leserinnen und Leser bedeutet wenig Redundanz, dass Fehler nicht durch redundante Information im Umfeld kompensiert werden können. Ein Digitalisat muss deshalb möglichst korrekt sein, da sonst rasch Information ganz verloren geht.

Beim Schweizerischen Idiotikon wurde das Problem in dieser frühen Phase der Digitalisierung so gelöst, dass das Resultat der automatischen Texterkennung des Wörterbuchs ohne manuelle Korrekturen als reine Suchstruktur in der Volltextsuche verwendet wurde. In der Volltextsuche wird über den auch teils fehlerhaft erkannten Wörterbuchtext gesucht. Als Treffer angezeigt wird danach das gescannte Faksimile, auf dem für menschliche Augen dann der korrekte Text ersichtlich ist. Vielleicht wird wegen OCR-Fehlern nicht alles gefunden; was gefunden wird, wird auf diese Weise jedoch immer korrekt angezeigt.

Es wurde zu diesem Zeitpunkt bewusst davon abgesehen, einen vollständig korrekten digitalen Volltext anzufertigen (sei es durch manuelle Korrektur einer OCR-Vorlage, sei es durch Abtippen, sogenanntes Double-Keying). Der zu betreibende Aufwand wäre beträchtlich, enthält das Wörterbuch doch bisher rund 120 Millionen Zeichen auf gut 15 500 Seiten. Der korrekte Volltext ist dabei aber nur ein Aspekt eines digitalen beziehungsweise digitalisierten Wörter-

buchs. Mindestens genauso wichtig ist die Auszeichnung des Wörterbuchttexts, also die Angabe, wie die einzelnen Textteile zu verstehen sind, ob es sich zum Beispiel um Formen-, Verbreitungs-, Bedeutungsangaben oder auch um Belege und Quellenangaben handelt. Die Auszeichnung des Wörterbuchttexts ist gerade beim Schweizerischen Idiotikon verglichen mit der Gewinnung des digitalen Volltexts eine ungleich grössere Aufgabe. Die Befürchtung war nun einerseits, dass für die Auszeichnung des Texts zu wenig Ressourcen übrig geblieben wären, wenn man mit der Gewinnung des digitalen Volltexts begonnen hätte, und andererseits, dass der Volltext ohne Auszeichnungen gegenüber dem unkorrigierten OCR-Text keinen erheblichen Mehrwert brächte (beziehungsweise nicht genügend, gemessen am Aufwand).

Semantikregister

Der nächste grössere Digitalisierungsschritt für das Schweizerische Idiotikon war darum folgerichtig eine Teilaufgabe der Auszeichnung des Wörterbuchttexts, zwar wie erläutert noch ohne korrekten Volltext, aber selbstverständlich so ausgestaltet, dass die Resultate bei Vorliegen des korrigierten Volltexts weiterhin sinnvoll verwendet werden können. Dieser Digitalisierungsschritt, bei dem es vor allem um die Erschliessung der Bedeutungen geht und der intern unter dem Titel *Semantikregister* läuft, steht momentan kurz vor dem Abschluss.

Es werden dabei die Bedeutungen der Lemmata extrahiert (aus dem OCR-Text herauskopiert oder abgetippt) und diese dann mit dem Faksimile verbunden (über eine exakte Positionierung von Lemmata und Bedeutungsziffern auf der gedruckten Seite). Weiter wird für die bessere Such- und Verlinkbarkeit die standarddeutsche oder eine pseudostandarddeutsche Form angegeben, dann auch ein Bedeutungskern, möglichst in einem Wort, und eine Zuordnung zu einer taxonomischen Klassifikation (leicht modifiziert nach Hallig & von Wartburg 1963). Darüber hinaus werden grundlegende grammatische Informationen (etwa die Wortart) sowie der Belegzeitraum erhoben.

Mithilfe dieser Daten kann in der Online-Version des Wörterbuchs bereits für jedes Lemma eine kurze Bedeutungszusammenfassung beziehungsweise eine Bedeutungsübersicht in der Art eines kurzen Inhaltsverzeichnisses angegeben werden (cf. Abbildung 1). Dies erleichtert den Zugang zum Wörterbuch stark, da die Bedeutungsübersicht direkt bei den Suchtreffern erscheint und so viel

Schnē(w) 9,1372, Schnöuwli ▲

1. eig., Schnee; vom Stoff an sich wie von dessen einzelner Erscheinungsform
2. von Schneeähnlichem
 - a) von Speisen
 - α) zu festem Schaum geschlagenenes Eiweiss
 - β) 'Aphrogala, eine Gattung Speis von Milch, ein Schnee, Nüdelmilch'
 - γ) Kartoffelbrei
 - b) von Speisen, weisse Masse in der unreifen Haselnuss
 - c) schwammiges Fleisch schlecht geratener Apfel, Kohlrabi, Rettige, Rüben
 - d) 'Schnee und Nebel heissen die weniger durchsichtigen und lauterer Teile des einzelnen Kristalls oder der ganzen Kristallgarbe'
3. Pflanzenn., perblütiges Ruhrkraut, *Gnaphalium marg.*
[gedruckt 1925]

Abb. 1. Bedeutungsübersicht zum Lemma *Schnē(w)*

gezielter auf die entsprechende Stelle im Wörterbuchtext gesprungen werden kann. Ausserdem ist so auch eine Volltextsuche einzig über die Texte in den Bedeutungserläuterungen möglich.

Sehr wichtig sind in diesem Zusammenhang auch die genauen Positionierungen der Lemmata und Bedeutungen auf der gescannten Seite. Von der Digitalisierung des Lemmaregisters her sind für die einzelnen Lemmata lediglich Band und Spalte bekannt, nicht aber, wo auf der entsprechenden Spalte sich das Lemma befindet. Das herauszufinden ist nun nicht in jedem Fall ganz einfach (deswegen die manuelle Markierung). Komposita stehen nach ihren Grundwörtern, allerdings wird typischerweise das Grundwort nicht mehr ausgeschrieben. Oft steht also ein Lemma gar nicht vollständig als solches im Text. Die Reihenfolge ist auch auf Ebene der Komposita nicht normalalphabetisch, sondern folgt dem schmellerschen System (cf. Staub 1876). Auf Spalte 16 von Band 11 zum Beispiel finden wir die Lemmata *Änte(n)stall*, *Einungstall*, *Eselstall*, *Understall* und *Üsstall*. Konkret im Text erscheinen Sie in folgender Reihenfolge und Form:

«Einung:- [...] Under:- [...] Enteⁿ- bzw. Ä:- [...] Ûs(s):- [...] Esel:- [...]». Eine automatische Zuordnung ist unter diesen Voraussetzungen mit der notwendigen niedrigen Fehlerrate kaum möglich.

Neben dem zusätzlich digitalisierten Material aus dem Semantikregister ist das Online-Wörterbuch auch weiter informatisch und computerlinguistisch ausgebaut worden. So werden die Positionierungen dazu verwendet, das gesuchte Stichwort im Faksimile farbig zu hinterlegen. Erste Versuche, den Wörterbuchtext dynamischer zu machen, sind mit einer Aktivierung der internen Verweise unternommen worden. Interne Verweise, bei denen explizite Band- und Spaltenangaben gemacht werden, sind leicht automatisch zu erkennen und weisen auch wenige OCR-Fehler auf. Diese internen Band-Spalten-Verweise sind im Faksimile aktiviert worden, sodass man diese nun anklicken kann und gleich an die entsprechende Stelle geführt wird.

REST-API und Mobilversion

Immer mehr zeigte sich, dass das Wörterbuch auch maschinenlesbar zugänglich sein sollte. Konkreter Anlass war eine externe Anfrage sowie ein internes Nutzungsszenario, das auf Webservices aufbaute. 2016 wurde deshalb für die Maschine-zu-Maschine-Kommunikation eine REST-API eingerichtet. Damit können auch andere das Schweizerische Idiotikon auf einfache Art und Weise von eigenen Applikationen aus absuchen beziehungsweise in eigene Applikationen einbinden.

Etwa ein Jahr später folgte eine eigens für Mobilgeräte optimierte Version des Online-Wörterbuchs. Das klassische Online-Wörterbuch ist auf kleineren Mobilgeräten kaum bedienbar, und deshalb wurde die Nachfrage nach einer Mobilversion immer grösser: Das Schweizerische Idiotikon wird nicht nur von Spezialistinnen und Spezialisten bei ihrer Arbeit im Büro verwendet, sondern eben auch von Laien, um unterwegs rasch etwas nachzuschlagen. Da das Online-Wörterbuch, wie oben dargestellt, nicht ohne die gescannten Faksimiles auskommt, mussten diese für eine Darstellung auf Mobilgeräten konvertiert werden. Der Text wurde dazu von einem zweispaltigen in ein einspaltiges Layout überführt, die Scans also gewissermassen in der Mitte zerschnitten. Eine Spalte kann, hält man das Gerät hochformatig, gut lesbar auf einem Mobiltelefon dargestellt werden (cf. Abbildung 2). Zusätzlich wurden die Scans etwas stärker komprimiert, damit Abfragen auch bei langsamerer Internetverbindung noch möglich sind.

Im Rahmen seiner digitalen Tätigkeiten pflegt das Schweizerische Idiotikon Kooperationen mit ähnlich ausgerichteten Institutionen und Projekten im In- und Ausland. So ist es im Wörterbuchnetz des Kompetenzzentrums für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier¹ integriert und dort mit absuchbar. Das Schweizerische Idiotikon hat an der von 2013 bis 2017 laufenden COST-Aktion zur elektronischen Lexikografie «European Network of e-Lexicography (ENeL)»² teilgenommen. Schliesslich ist das Schweizerische Idiotikon Projektpartner bei histHub³, einem Gemeinschaftsprojekt mit dem Ziel, vernetzte und normierte Daten für die historischen Wissenschaften bereitzustellen.

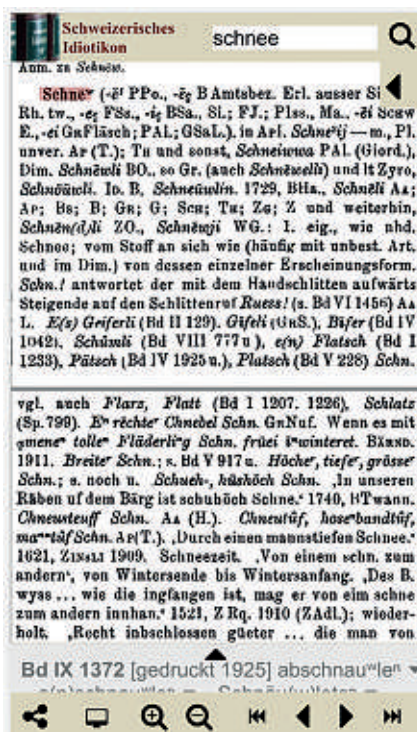


Abb. 2. Mobilversion des Online-Wörterbuchs mit einspaltigem Layout

1 <http://www.woerterbuchnetz.de>.

2 <http://www.elexicography.eu>.

3 <https://histhub.ch>.

Weitere digitale Projekte

Neben dem eigentlichen Wörterbuch sind dem Schweizerischen Idiotikon weitere Teilprojekte mit stark digitalem Fokus aus den Themengebieten Dialektologie, deutsche Sprache sowie Namenforschung angegliedert.

ortsnamen.ch

Das Schweizerische Idiotikon betreibt ortsnamen.ch⁴, das Portal der schweizerischen Ortsnamenforschung (cf. zum Beispiel Bickel & Graf 2015). Auf ortsnamen.ch werden Daten der meist kantonal organisierten Namenbuchprojekte gemeinsam online publiziert und so für ein breiteres Publikum auf neue Weise nutzbar gemacht. Das Portal übernimmt auch eine Archivfunktion für die beteiligten Projekte, erschliesst die Daten neu und macht sie zugänglich für weitere Nutzungsformen, etwa im Rahmen des Kooperationsprojekts histHub (cf. Roth 2016, Weibel & Roth 2018).

Materialien des Sprachatlas der deutschen Schweiz (SDS)

Die Materialien des Sprachatlas der deutschen Schweiz (SDS) befinden sich in der Obhut des Schweizerischen Idiotikons. Es handelt sich dabei vor allem um die Originalaufnahmen, welche die Grundlage für den Sprachatlas bildeten (cf. z.B. Schaller & Schiesser 2017). Es finden sich darin aber auch viele Fotografien und weiteres Zusatzmaterial.

Zwecks sicherer Archivierung und besserer Verfügbarmachung für die Forschungsgemeinschaft wurden die Originalaufnahmen digitalisiert und online publiziert.⁵ Es handelt sich allerdings um einfache Scans, eine Transkription wäre zu aufwendig: Das Material ist einerseits sehr umfangreich, andererseits fast komplett handschriftlich, meist phonetisch transkribierte Formen mit vielen Einsprengseln in Kurzschrift (Steno).

Das fotografische Material soll ebenfalls digitalisiert und, wo sinnvoll, mit dem Schweizerischen Idiotikon verlinkt werden.

4 <https://ortsnamen.ch>.

5 <https://www.sprachatlas.ch>.

Schweizer Textkorpus

Das Schweizer Textkorpus⁶ wurde als Referenzkorpus für das Standarddeutsche des 20. Jahrhunderts in der Schweiz ursprünglich an der Universität Basel entwickelt (cf. Bickel et al. 2009) und später dann ans Schweizerische Idiotikon übertragen. Das Korpus für das 20. Jahrhundert hat einen Umfang von rund 20 Millionen Textwörtern und ist nach Zeit, Werkkategorie (Belletristik, journalistische Texte, Gebrauchstexte, Sachtexte) und Themen beziehungsweise Sachgruppen ausgewogen zusammengesetzt. Vom Konzept her lehnt es sich an das Kernkorpus des DWDS⁷ in Deutschland an, mit dem zusammen (und weiteren Projektpartnern aus Österreich und Südtirol) es das verteilt abfragbare *Korpus C4*⁸ bildet (cf. Dittmann et al. 2012). Das *Korpus C4* ist zwar noch online, hat aber unterdessen eher Beispielcharakter für eine frühe Entwicklung in diesem Bereich.

Seit 2017 wird das Textkorpus nach denselben Prinzipien wie das bestehende Kernkorpus und im selben Umfang (relativ gesehen) um Texte aus den ersten Jahren des 21. Jahrhunderts erweitert. Die Texte dazu wurden bis Ende 2018 gesammelt und sind seit dem ersten Quartal 2019 online. Das Korpus soll fortan kontinuierlich aktualisiert werden und so ein aktuelles Referenzkorpus bleiben.

Seit 2019 wird am Schweizerischen Idiotikon ein schweizerdeutsches Mundartkorpus aufgebaut. Es soll eine wichtige lexikografische Grundlage für das Wörterbuch werden, aber natürlich auch ganz allgemein als linguistisches Forschungsinstrument zur Verfügung stehen.

Aktiv archivierte Projekte

Das Schweizerische Idiotikon beherbergt einige weitere digitale Projekte, an denen aber nicht mehr aktiv weitergearbeitet wird. Sie sind hier in einem aktiven Archivierungszustand, das heisst ihr Online-Auftritt wird sichergestellt (inklusive rein technisch notwendiger Aktualisierungen), und natürlich werden die Daten sicher archiviert. Es sind dies zwei Projekte, die an der Universität Basel

6 <https://www.chtk.ch>.

7 <https://www.dwds.de>.

8 <https://www.korpus-c4.org>.

entstanden sind: Das *Kollokationenwörterbuch*⁹ (cf. Häcki Buhofer et al. 2014, Roth 2014) und das Projekt OLdPhras¹⁰ zur historischen Phraseologie des Deutschen.

Digitalisierungsstrategie

Das Schweizerische Idiotikon hat bisher weitgehend die bei Bickel (2007) formulierte Digitalisierungsstrategie verfolgt und ist dem dort skizzierten etappenweisen Vorgehen in kleinen Schritten auch weiter treu geblieben (cf. auch Landolt & Roth 2019). Ein weiteres wichtiges Merkmal der Digitalisierungsstrategie der letzten Jahre ist das Bestreben des Schweizerischen Idiotikons, als digital aktiv wahrgenommen zu werden.

Etappenweises Vorgehen in kleinen Schritten folgt ganz den Prinzipien agiler Software-Entwicklungs-Methodologien, mit dem von Raymond (1999) verankerten Motto: «Release early. Release often. And listen to your customers.» Die im vorhergehenden Kapitel beschriebenen Digitalisierungsschritte illustrieren dies gut. Ungewohnt für das Schweizerische Idiotikon mit seinen hohen Qualitätsansprüchen war dabei vielleicht die Tatsache, dass auch Unfertiges bereits an die Öffentlichkeit gelangt. Die Rückmeldungen zu und Erfahrungen mit auf diese Weise publizierten Diensten tragen dann aber stärker zur Qualitätsverbesserung bei, als wenn noch nicht publiziert worden wäre. Ausserdem ist sich das Publikum bei digitalen Diensten dieses Vorgehen auch eher gewohnt als in anderen Bereichen.

Der zweite Punkt, die stärker wahrnehmbare digitale Aktivität des Schweizerischen Idiotikons, widerspiegelt sich in den verschiedenen Projekten, die oben vorgestellt wurden, sowie in den umfangreichen, ebenfalls oben beschriebenen Digitalisierungsbemühungen.

9 <https://www.kollokationenwoerterbuch.ch>.

10 <https://www.oldphras.net>.

Übergang zum digitalen Wörterbuch

Was Gegenwart und nähere Zukunft angeht, so befindet sich das Schweizerische Idiotikon im Übergang zum digitalen Wörterbuch. Von seiner Anlage her ist das Schweizerische Idiotikon eindeutig ein traditionelles, gedrucktes Wörterbuch. Durch die Gründung bereits im 19. Jahrhundert ist dies auch gar nicht anders möglich. Die vorangehenden Abschnitte illustrieren die bisherigen Bemühungen hin zu einem digitalen Wörterbuch.

Über weite Strecken ist das Schweizerische Idiotikon bisher noch nicht wirklich ein digitales, sondern mehr ein digital dargestelltes Wörterbuch. Um zu einem wirklich digitalen Wörterbuch zu werden, müssen auch neue Inhalte digital erstellt und nahtlos eingefügt werden können. Es gibt im Wörterbuch vielerorts Ergänzungs- und Korrekturbedarf, was bei der langen Werkgeschichte nicht verwundert. Die bisherigen positiven Erfahrungen mit dem digital konsultierbaren Idiotikon haben gezeigt, dass ein vollständig digitales Idiotikon durchaus erstrebenswert ist. Voraussetzung dazu ist aber natürlich, dass die Tradition des Werks gebührend berücksichtigt wird.

Noch einige Schritte sind notwendig, bis das Schweizerische Idiotikon digital editierbar wird. Eine Grundvoraussetzung ist der korrekt retrodigitalisierte Volltext des gesamten Wörterbuchs: Wenn man Text verändern will, muss man diesen Text zuerst vorliegen haben. Der bisher verfolgte Ansatz mit automatischer Texterkennung ohne Korrektur (s. oben) kommt spätestens hier an seine Grenzen. So ist denn auch für die nächsten Jahre geplant, den korrekten Volltext mit sogenanntem Double-Keying (doppeltes Abschreiben mit Fehlerkorrektur) oder massgeschneidertem und manuell nachkorrigiertem OCR zu gewinnen. Schon früher stand diese Variante als Möglichkeit im Raum, sodass alle bisherigen Digitalisierungsprozesse so gestaltet wurden, dass sie auch mit Vorliegen des korrekten Volltextes noch funktionieren. So werden etwa die Resultate des Semantikregisters auch mit dem korrekten Volltext weiterverwendet werden können, desgleichen vollständig automatisierte Schritte wie die Erkennung interner Verweise.

Ein digitales Wörterbuch verlangt auch nach einer entsprechenden Redaktionsumgebung. Wörterbuchinhalte müssen digital erfasst, informationstechnisch erschlossen und direkt digital publiziert werden können. Die Redaktionsumgebung muss aber auch gewissermassen als Scharnier zwischen alten (retrodi-

gitalisierten) und neu direkt digital geschriebenen Artikeln fungieren. Das Ziel ist ein möglichst harmonisches Miteinander von neuen Artikeln, überarbeiteten und unveränderten retrodigitalisierten Artikeln auf Bearbeitungsebene.

Dieses Miteinander von Alt und Neu wird ebenso auf Darstellungsebene angestrebt. Natürlich bietet ein digitales Wörterbuch vielfältigere Darstellungsmöglichkeiten als ein gedrucktes. Je nach gewünschter Perspektive können die darunterliegenden Daten und der Basistext unterschiedlich visualisiert werden. Trotz diesen Möglichkeiten sollte man aber nicht vergessen, dass die bisherigen für das gedruckte Wörterbuch geschriebenen Artikel ganz explizit für genau dieses Drucklayout geschrieben wurden. Auch ein digitales Wörterbuch sollte immer noch den Zugriff auf diese Druckfassung erlauben, da sonst dem Publikum in der Digitalfassung ein wichtiger Teil des bisherigen Wörterbuchs vorenthalten würde. Ob dabei für das digitale Wörterbuch eine Darstellungsform gewählt wird, die auf dem Drucklayout aufbaut, oder ob das Drucklayout nur als Darstellungsalternative verwendet wird, bedarf weiterer Abklärungen. Da im digitalen Wörterbuch darstellungsunabhängiger produziert werden soll, muss dies aber auch nicht definitiv entschieden werden. Es ist durchaus wahrscheinlich, dass technische Entwicklungen und Trends im Internet dazu führen, dass in einigen Jahren neue Darstellungsformen gefordert sind.

Bereits geplant als Ausbauschritte auf Darstellungsseite sind verschiedene Massnahmen, um den Wörterbuchtext anzureichern, um so beim Verstehen und der Interpretation zu helfen. So sollen die vielen Abkürzungen, seien es geografische Abkürzungen, Quellenkürzel oder allgemeine Abkürzungen, aufgelöst werden. Ein Teil der wörterbuchinternen Verweise ist bereits klickbar aktiviert (s. oben), mittel- bis langfristig sollen sämtliche Verweise, sowohl interne als auch externe, zum Beispiel auf andere Wörterbücher, zu aktiven Links werden. Weiter sollen die Wörterbuchartikel um Zusatzdaten ergänzt werden. Dies können enzyklopädische Zusatzdaten zum Beispiel aus Wikidata¹¹ sein oder auch Bilder, etwa aus dem Fotoarchiv des Sprachatlas der deutschen Schweiz und aus externen Quellen.

11 <https://www.wikidata.org>.

Neben seiner über 150-jährigen allgemeinen Werkgeschichte kann das Schweizerische Idiotikon unterdessen auch schon eine beachtliche digitale Werkgeschichte vorweisen. Es ist ausserdem auch in verwandten Bereichen digital aktiv, wo es um schweizerdeutsche Dialekte und um die deutsche Sprache in der Schweiz geht.

Die Umwandlung des Schweizerischen Idiotikons in ein vollständig digitales Wörterbuch ist das nächste grössere Ziel und gleichzeitig eine konsequente Weiterführung der bisher verfolgten Digitalisierungsstrategie. Der Redaktion wird die komplette Digitalisierung die flexible Bearbeitung und Ergänzung des Wörterbuchs ermöglichen. Über verbesserte Darstellungsformen kann das Wörterbuch näher zu den Leuten rücken und von noch mehr Menschen (und Maschinen) benutzt werden.

Literatur

Bickel, Hans (2007): «Idiotikon digital. Überlegungen zu einer elektronischen Ausgabe des Schweizerdeutschen Wörterbuchs», in: *Schweizerdeutsches Wörterbuch. Schweizerisches Idiotikon. Bericht über das Jahr 2006*, Zürich, S. 13–26.

Bickel, Hans, Markus Gasser, Annelies Häcki Buhofer, Lorenz Hofer & Christoph Schön (2009): «Schweizer Text Korpus – Theoretische Grundlagen, Korpusdesign und Abfragemöglichkeiten», in: Häcki Buhofer Annelies (Hg.): *Fortschritte in Sprach- und Textkorpusdesign und linguistischer Korpusanalyse II*, Linguistik online 39 (3/2009), S. 5–31 [<http://dx.doi.org/10.13092/lo.39.474>].

Bickel, Hans & Martin Hannes Graf (2015): «ortsnamen.ch – Portal der schweizerischen Ortsnamenforschung», in: *Bulletin SAGW* 4, S. 57–58.

Dittmann, Henrik, Matej Ďurčo, Alexander Geyken, Tobias Roth & Kai Zimmer (2012): «Korpus C4 – a distributed corpus of German varieties», in: Schmidt, Thomas & Kai Wörner, (Hg.): *Multilingual Corpora and Multilingual Corpus Analysis*, Amsterdam: Benjamins, S. 339–346.

Häcki Buhofer, Annelies, Marcel Dräger, Stefanie Meier & Tobias Roth (2014): *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen: Francke.

Hallig, Rudolf & Walther von Wartburg (1963): *Begriffssystem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas. 2., neu bearbeitete und erweiterte Auflage*. Berlin: Akademie-Verlag.

Landolt, Christoph & Tobias Roth (im Druck): «Schweizerisches Idiotikon – Wörterbuch der schweizerdeutschen Sprache», in: Stöckle Philipp (Hg.): *Dialektlexikographie im 21. Jahrhundert* (ZDL-Beiheft).

Raymond, Eric S. (1999): *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Cambridge, Mass.: O'Reilly Media.

Roth, Tobias (2016): «Isolation and Mapping of Place-Name Forms in Toponymic Data», in: Stefanie Dipper, Friedrich Neubarth & Heike Zinsmeister (Hg.): *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)* (=Bochumer Linguistische Arbeitsberichte 16), S. 221–225 [https://www.linguistics.rub.de/konvens16/pub/28_konvensproc.pdf].

Roth, Tobias (2014): *Wortverbindungen und Verbindungen von Wörtern. Lexikografische und distributionelle Aspekte kombinatorischer Begriffsbildung zwischen Syntax und Morphologie* (Basler Studien zur deutschen Sprache und Literatur 94). Tübingen: Francke.

Schaller, Pascale & Alexandra Schiesser (2017): *Die Vermessung der Sprache. Zu Geschichte und Bedeutung des Sprachatlas der deutschen Schweiz* (Swiss academies reports, 12, 4), Bern: Schweizerische Akademie der Geistes- und Sozialwissenschaften.

Staub, Friedrich (1876): *Die Reihenfolge in mundartlichen Wörterbüchern und die Revision des Alphabetes. Ein Vorschlag zur Vereinigung; vorgelegt vom Bureau des Schweizerdeutschen Idiotikons* [Zürich].

Weibel, Manuela & Tobias Roth (2018): «On Modelling a Typology of Geographic Places for the Collaborative Open Data Platform histHub», in: Eetu Mäkelä, Mikko Tolonen & Jouni Tuominen (Hg.): *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference* (DHN 2018): Helsinki, 7.–9.3.2018, CEUR-WS.org, S. 170–178 [<http://CEUR-WS.org/Vol-2084/shortplus2.pdf>].

Rétrodigitalisation du Glossaire des patois de la Suisse romande : inauguration du portail web

Alexandre Huber

Historique

Pouvoir consulter une version électronique du Glossaire des patois de la Suisse romande (GPSR) est un rêve qu'ont eu bon nombre de lecteurs, mais aussi de rédacteurs du dictionnaire. Pour les premiers, ce serait la possibilité de s'affranchir de l'ordre alphabétique des articles, qui régit la structure de chaque fascicule imprimé depuis 1924; pour les seconds, ce serait en outre la possibilité de gagner en efficacité de rédaction en bénéficiant de toute la puissance de la recherche plein texte. Celle-ci leur permettrait par exemple de trouver plus facilement des cas analogues à ceux qu'ils doivent résoudre dans leurs articles en cours de rédaction et leur éviterait certainement le risque de consacrer du temps à élaborer des solutions qui existent déjà.

En 1998, Eric Flückiger, rédacteur au GPSR, effectua un scannage complet de plus de 6000 pages du dictionnaire pour tenter un premier essai de rétrodigitalisation, fondé sur le principe de l'océrisation (reconnaissance optique de caractères). Cette démarche donna un résultat très intéressant, mais encore imparfait dans la mesure où les nombreux signes spéciaux utilisés pour la transcription phonétique des patois romands ne pouvaient être valablement reconnus par les logiciels d'océrisation.

En 2014, un scannage professionnel fut confié à une entreprise de la région lausannoise. Malgré des fichiers image d'excellente qualité, le problème persista. Comme en 1998, les logiciels d'océrisation disponibles dans le commerce ne pouvaient fournir une rétrodigitalisation fiable du dictionnaire. Cela était dû aux signes spéciaux qui foisonnent dans le GPSR et qui lui sont propres.

La solution est venue en 2016 grâce à un financement de l'Académie suisse des sciences humaines et sociales, qui a permis de fructueuses collaborations avec plusieurs partenaires, tout d'abord avec le Center for Digital Humanities de l'Université de Trèves. Ce centre de compétence, connu pour avoir réalisé la

version électronique du dictionnaire de Grimm¹, a effectué la rétrodigitalisation du GPSR sur la base du scannage de 2014 et en utilisant le principe de la double saisie. Ce procédé, qui consiste à dactylographier deux fois l'imprimé original (pour faire des corrections en fonction des divergences), assure un taux de fiabilité extrêmement élevé.

En 2017, le fichier XML des huit tomes du GPSR est livré à Neuchâtel, puis transmis à une équipe d'informaticiens dirigée par Fabrice Camus, professeur à la Haute École de gestion ARC (Institut de digitalisation des organisations). En se fondant sur les balises introduites par l'Université de Trèves, Fabrice Camus restitue la très complexe typographie du GPSR avec ses nombreux signes spéciaux et ses fréquents changements de style. Les polices Unicode du GPSR, entièrement modernisées par Sarah Kremer de l'Atelier national de recherche typographique de Nancy, sont converties au format woff (Web Open Font Format) pour qu'elles soient lisibles sur n'importe quel terminal sans qu'elles doivent y être concrètement installées. Peu à peu le portail web prend forme, en bénéficiant notamment des propositions du GRI, c'est-à-dire du Groupe de réflexion informatique du Glossaire (formé de Raphaël Maître, de Christel Nissille et du soussigné).

Aujourd'hui (12 septembre 2018), la rétrodigitalisation des 7197 pages du GPSR est effective et le portail web qui permet d'y accéder peut être officiellement inauguré.

Exemples de recherches

Grâce au portail web, les possibilités de recherches se sont multipliées. Autrefois uniquement consultable par le biais des lemmes classés alphabétiquement, le GPSR est devenu protéiforme. Sa riche matière se consulte dorénavant par des accès variés et se décline en une série de dictionnaires parallèles et complémentaires. Auparavant cantonné à la structure d'un dictionnaire patois-français, le GPSR est devenu, grâce aux nouvelles requêtes rendues possibles par la rétrodigitalisation, un dictionnaire français-patois, une encyclopédie, un répertoire onomastique et, dans un proche avenir, il fonctionnera comme une iconothèque.

1 *Der digitale Grimm*, <http://dwb.uni-trier.de>.

Recherches dans le sens patois-français

Même si le GPSR imprimé comporte un certain nombre de lemmes français (lorsque les formes patoises ont d'exacts correspondants dans cette langue), il a globalement la structure d'un dictionnaire patois-français. C'est donc dans ce sens que s'y font traditionnellement les recherches les plus naturelles. Typiquement, le sens d'un mot patois énigmatique est recherché dans les huit volumes déjà parus, dans lesquels on se repère au moyen de l'ordre alphabétique des articles. Cette façon de procéder n'est pas toujours couronnée de succès, car le mot patois recherché peut être phonétiquement éloigné du lemme sous lequel il se trouve. Rares sont les lecteurs qui auraient spontanément l'idée de rechercher :

- *artans* sous *èrtîns* « héritage » (t. VI, p. 663a);
- *dyèch* sous *gèis* « chèvre » (t. VIII, p. 205a);
- *şuva* sous *fîva* « épicéa » (t. VII, p. 494a);
- *zîga* sous *èga* « jument » (t. VI, p. 150a).

Maintenant qu'existe le GPSR rétrodigitalisé, tout est plus facile. Prenons l'exemple d'une personne qui, chez un bouquiniste, aurait acquis *Ouna Fourdêrà dè-j-èlyudzo* de Tobi di-j-èlyudzo (1906) et qui voudrait comprendre le titre de ce recueil d'historiettes patoises, notamment le sens du dernier mot *èlyudzo*. Si elle dispose d'une connexion internet, il suffit à cette personne d'inscrire dans son navigateur l'adresse portail-gpsr.unine.ch. Une fois cela fait, la page d'accueil du portail apparaît : elle présente une description des principales fonctionnalités du moteur de recherche. Dans le cas précis, il s'agit d'aller à l'onglet « recherche simple » et d'inscrire dans le seul champ disponible *elyudzo*, sans se soucier des accents.

Fig. 1. Recherche simple

La recherche, une fois lancée, va faire apparaître une liste de résultats. Il s'agit des articles où figure le mot *elyudzo*, chaque fois mis en surbrillance avec son contexte immédiat.

DE; (VI, 40)

... l'alpage (V Hérém. LAV. 358). *Tobi dij elyudzo*, T. des éclairs, pseudonyme d'auteur (F...

DÉLUGE (VI, 268)

...trouve dans des mots indigènes tels que *elyudzo*, *ezudzo* «éclairs» (sous *elyudzo*), *lyudzo*...

èkòvā (VI, 231)

...ent le pont d'un bateau (Vd Orm.). *On-n elyudzo d'èrdzin ka ly èkòvè la yè* [cf. *ayèr* I]...

èludzj (VI, 255)

... Roche. *Atm. cathol. SR*, 1949, 77); cf. *elyudzo* 1° comparaisons. 4° Être mouvant, en pa...

èlyudzo (VI, 256)

...). **Loc. et comparaisons.** *Nə vèr tyè lèj elyudzo*, ressentir un éblouissement à la suite...
 ...ue facilit (B Bone.). *Chè lèvā kmin ou-n elyudzo*, se lever subitement (F Gru.). *Durā kom...*
 ...des é. (N Boudry). | *Ouna fourdèrà dè-j-elyudzo*, litt. un tablier plein d'é., titre don...
 ...patois de F Crés. (*Bibl.* 659); *Tobi dij elyudzo*, pseudonyme de l'auteur C. RUFFIEUX. | ...

fòdèrā (VII, 583)

... (V Grim.). Au fig. *Ouna fourdèrà dè-j-elyudzo*, litt. un t. plein d'étincelles, titre ...

Fig. 2. Résultats de la recherche simple

Avec ses quatre occurrences d'*elyudzo*, le cinquième article semble être le plus intéressant. Pour le consulter, il suffit d'effectuer un double-clic sur son lemme. L'article apparaît alors dans son entier. On apprend finalement que le mot signifie « éclair » et que, sous la plume de Cyprien Ruffieux², il prend, par métaphore, le sens d'« historiette », de « bon mot » dans le titre *Fourdèrà dè-j-elyudzo*, littéralement « tablier plein d'éclairs ». La structure de ce titre n'est pas sans rappeler les mots français *anthologie* et *florilège*, littéralement « cueillette/choix de fleurs ».

2 Le véritable patronyme de l'instituteur et poète gruérien qui a écrit sous le pseudonyme de Tobi di-j-elyudzo.

Quant aux mots patois évoqués plus haut (*artans*, *dyèch*, *șuva*, *ziga*), on les repère dans le GPSR rétrodigitalisé avec la même facilité :

ærtîns (VI, 663)

...- B 12, *ærtans*, *â-* N 2, 41, 42, 50, 51, *ærtans* B 22, 23 (-*âns*), 2 STALDER (d'où BR. 47...
...*èchtaman*, *vo vyâs* [voulez] *avaâ é-n bal* *ærtans*, si ce vieillard ne fait pas de testame...

Fig. 3. Résultat pour le mot patois *artans*

gêts (VIII, 205)

...ER, *gêks* 84 corr., *gyé(k)s* 80 a R. DUC, *dyèch* 22 MULLER; *gayis* J 35 ALF, 45 rare, 46,...

Fig. 4. Résultat pour le mot patois *dyèch*

fîva (VI, 494)

...Vd 10 var., *hlyva* F 13, 1 Vuadens Co., *șuva* 1 BOR., SAVOY, CL. GLASSON, *Dict. pat.,...*
...es balais de sapin ou de *fie* (ib.). La *șuva* l é on bou ka va bin pô travayi; l é pr...
...n blanc (F Hte-Glâne. L'HOMME, 158). La *șuva* l é mèya tyè lou vouçnou pòr lè lan, lit...

Fig. 5. Résultat pour le mot patois *șuva*

èga (VI, 150)

...lîga 83 corr. (Gch. *ly-*), *ouîga* 60 var., *ziga* Vd 5 TP. (*Bull. Gl.* II, 38). Anc. *egua*,...

Fig. 6. Résultat pour le mot patois *ziga*

Pour connaître le(s) sens de ces mots, il ne reste plus qu'à afficher la totalité des articles.

Recherches dans le sens français-patois

Ce type de recherche n'est guère aisé dans le GPSR imprimé. Seuls les articles à lemmes français peuvent être d'une certaine utilité. Ainsi l'article *champignon* (t. III, p. 294a) fournit un certain nombre de correspondants patois: on y découvre une forme vaudoise et fribourgeoise *tsanpənyon*, une forme genevoise *şanpinyon*, une forme neuchâteloise *tchanpinyon*, etc., toutes remontant au même étymon (< latin *CAMPANIA* + suffixe *-ŌNE*). Il s'agit majoritairement d'emprunts au français, plus ou moins adaptés. Malheureusement, les désignations autochtones du champignon ne se trouvent pas dans cet article et ne peuvent pas être repérées rapidement dans le GPSR imprimé. En revanche, grâce au portail web, on y accède en une fraction de seconde. Il suffit d'inscrire *champignon* dans le champ prévu à cet effet :

Fig. 7. Recherche simple du mot français *champignon*

Après avoir lancé la recherche, on obtient une liste de 28 résultats, dont voici les onze premiers :

Nombre de résultats trouvés : 28 article(s)

AMADOU (II, 322)

...près nos glossaires cantonaux). Amadou, **champignon** qui, après diverses préparations, serva...

baronta (II, 261)

...*la da pôl-nta*, épi de maïs (Évol.). **3° Champignon** (Grim.). **4°** Par anal. Pl. Testicules (É...

bòkèt (II, 467)

...*ius* et *fomentarius* Fries) et tout autre **champignon** acaulé croissant sur un tronc, une souc...

bòtèt (II, 473)

...- F XVI^e s. — *Tabl. Suppl. 47; ALF 227 (champignon), 1435 (amadou), Suppl. 307.*
... par erreur, Bn, forme *bohla*; Dum.). **1° Champignon**, sans distinction d'espèces (Vd Ollon, ...
Dérivé du lat. *holētus* «espèce de **champignon**»; REW 1193; FEW, I, 426. «*Bouteis* et «b...

bòlèn (II, 475)

|| S. m. **1° Champignon** en général; spécialt espèces de bolets ...

boqasè (II, 573)

...adou, c'est la texture spongieuse de ce **champignon** qui aurait suggéré la valeur de matièr...

boulra (II, 634)

|| S. m. **1° Champignon** en général (Boécourt, Bois, Épauv.). *Lé...*
...amadou. || Lycoperdon (Plagne). || Tout **champignon** vénéneux (Pomm).

champignon (III, 292)

|| S. m. **1° Champignon**, nom générique (Vd, V, F-B). *Néd a plu...*
...que de vin (Vd Flendr.). || Spécialt. **1.** **Champignon** des prés, *Psalliota campestris* L. (B Aj...
...l ne faut pas l'abattre (Vd Penth.). **3.** **Champignon** des caves, *Merulius laerymans* Fries (Vd...

CHAMPIGNON (III, 294)

CHAMPIGNON, *tsanpənyŋon* Vd 16, 31, 50, 62, -pi- V 2...
...ue., *chanpanyon* Vd 3 Forel. **Anc.** patois **champignon** G XVIII^e s. *Cris.* — *ALF 227.*
|| S. m. **1° Champignon**, nom générique (Vd-B). *Alô è tsanpyngon...*
...oître les ch. (Vd Sav.). || Spécialt. **1.** **Champignon** comestible, par oppos. à *bòlèk* (sous *bò...*
...tible, par oppos. à *bòlèk* (sous *bòtèt*) «**champignon** en général» (V St-Luc). **2.**
en général» (V St-Luc). **2.** **Champignon** de couche, *Psalliota campestris* L. (F G...
...yon de *kurti* [jardin], id. (F Attal.). «*Champignon* de Pariss, id. (fr. rég. SR). **3.** *Tsanpa...*
...n, etc., résulte d'un croisement du fr. **champignon** avec *champagnou*. Cf. FEW, II, 152 b.

CHAMPIGNONNIER (III, 294)

Dérivé de **champignon** par le suff. -a r i u adapté comme après p...

CHAPEAU (III, 334)

9° Champignon (Vd V. de Joux, B Bois). *Lé tsapè sè pl...*

Fig. 8. Résultats de la recherche simple du mot français *champignon*

En consultant les articles dans leur intégralité, on accède sans peine aux désignations autochtones du champignon. On retiendra surtout :

- *baroula*, substantif féminin d'étymologie inconnue, recueilli par Jules Gilliéron à Grimentz ;
- *bòkèl*, substantif féminin jurassien (dialectes oïliques), appartenant à la famille de *bouc* et désignant « tout champignon acaule croissant sur un tronc, une souche d'arbre ou du bois » ;
- *bòlài*, substantif masculin, surtout vaudois et valaisan, issu par voie vernaculaire du latin BOLĒTU (les formes patoises empruntées se trouvent sous le lemme français de même étymon *bolet*) ;
- *bòlìn*, variante occasionnelle du précédent ;
- *boulraq*, substantif masculin jurassien, appartenant à la descendance du latin BOLĒTU ;
- *champagnou*, lemme (français régional, forme fournie par E. Gauthey) sous lequel viennent se ranger des substantifs patois de presque toute la Suisse romande (sauf Genève) remontant au latin CAMPANIA + -ÖLU, formation bien attestée en ancien picard et en moyen français (*FEW*, II, 152b) ;
- *tsapé* dans la Vallée de Joux et *tchèpé* aux Bois (tous les deux sous le lemme français *chapeau*) désignent métonymiquement (et métaphoriquement) le champignon, en tant que terme générique ;
- etc.

Recherches encyclopédiques

Le GPSR n'est pas seulement un dictionnaire de langue, il propose également un contenu encyclopédique de manière à renseigner le lecteur qui voudrait s'informer sur certains aspects techniques, culturels, historiques et folkloriques de la Suisse romande. Prenons l'exemple d'une personne qui aurait lu *Plantes et Savoirs des Alpes* de Sabine Brüscheweiler (1999), ouvrage de phytothérapie anniviarde, et qui voudrait en savoir davantage sur les moyens de soigner les entorses. Le GPSR rétrodigitalisé lui fournira non seulement des désignations patoises de l'entorse (*desouè, dètouècha, ékouâsa, eska, étontch, étontchur, étòrsa, étouatchur, fòrsirà*, etc.), mais encore lui permettra de lire online une notice folklorique sur les remèdes soulageant ce type de blessure. Cette notice se trouve à la fin de l'article *entorse*. On y apprend notamment que:

- la consoude pilée et mélangée à de la poix de sapin fraîche fait un excellent emplâtre (Dompierre, canton de Fribourg);
- l'aigremoine et le blé printanier bouillis dans de la lie de vin sont appliqués avec profit sur le membre foulé, pour autant que l'on dise sept fois: « Remets ce pied (ou ce bras) comme les sept sacrements remettent le péché » en faisant le signe de croix à triple traverse (Épauvillers, canton du Jura);
- autre « remède », administré dans la Vallée de Joux: asséner un bon coup de poing sur l'entorse au moment où la personne souffrante ne s'attend à rien!

Les notices encyclopédiques et folkloriques sont nombreuses dans le GPSR. Elles viennent étayer le sémantisme d'un mot et se rencontrent surtout quand les réalités romandes ont quelque chose de particulier qui ne s'observerait pas dans d'autres contrées. À titre d'exemples, voici quelques articles présentant de telles notices:

- abbaye, absinthe, âme, armalyj... (lettre A-, t. I);
- baptême, bénichon, bisse, boucherie... (lettre B-, t. II);
- carnaval, charivari, charrue, consortage... (lettre C-, t. III-IV);
- dent, désalpe, deuil, diable, doigt... (lettre D-, t. V);

- enterrement, envie, escargot, éternuer... (lettre *E-*, t. VI);
- fée, fondue, four, fourneau, fromage... (lettre *F-*, t. VII);
- gage, gautschage, génépi, Gourze... (lettre *G-*, t. VIII).

Certaines notices sont placées là où on ne les attend guère. Qui penserait trouver dans les huit tomes publiés du GPSR (lettres *A-* à *G-*) une notice consacrée à la lessive, aux narcisses, à la pie, au rouet, ou encore à la taupe ? Les possibilités de recherche dans le GPSR rétrodigitalisé font apparaître que ces sujets sont traités dans des articles ayant pour lemme un terme romand :

- *agasə*, pour un commentaire sur la mauvaise réputation de la pie;
- *bourgo*, pour un commentaire sur le rouet (types, parties, importance et croyances);
- *buya*, pour un commentaire sur la grande lessive aux cendres;
- *dèrbon*, pour un commentaire sur la taupe (croyances et médecine populaire);
- *goitreux*, pour un commentaire sur la Fêtes des Narcisses de Montreux.

Les renseignements encyclopédiques sont parfois plus diffus : certaines problématiques s'égrènent au fil des articles et sont signalées par un système de sous-titres, volontiers imprimés en gras pour que le lecteur puisse les repérer lorsqu'il parcourt le GPSR. Ces sous-titres, qui obéissent à une certaine standardisation, constituent d'excellents critères de recherche pour celui qui veut reconstituer l'ensemble d'une problématique :

- « encycl » pour les notices encyclopédiques;
- « folkl », « chanson* », « croyance* », « formulette* », « dicton* », « prov » pour le folklore (l'astérisque permet d'inclure le pluriel);
- « bot », « zool » pour la flore et la faune, sans oublier les noms savants latins pour rechercher une plante ou un animal précis;
- etc.

Recherches en onomastique

Le traitement des noms de lieux et de familles contribue grandement à la spécificité du GPSR... et à son image de marque. En tirant parti de la grande enquête toponymique et anthroponymique effectuée dès 1902 (et pendant une trentaine d'années) par le Professeur Ernest Muret, les articles du GPSR incluent de l'onomastique quand elle peut contribuer utilement à la description de la langue vernaculaire. L'onomastique est introduite par des sous-titres en gras qui permettent de la repérer facilement. Ceux-ci peuvent constituer des critères de recherche intéressants dans le GPSR rétrodigitalisé. Il s'agit d'utiliser les séquences « nom* de lieu* », « nom* de fam », ou encore « nom* de pers ». Suivant ce que l'on désire trouver, on choisit un de ces trois syntagmes, puis on l'introduit dans l'onglet « recherche simple » en prenant garde d'utiliser l'option « recherche personnalisée »; cela permet que le tri se fasse en fonction de la suite exacte des éléments. Si l'on s'intéresse à une commune particulière, par exemple à Puidoux dans le canton de Vaud, on peut tenter de restreindre la requête en utilisant l'opérateur logique PROX (qui limite la zone de recherche à une succession d'un nombre déterminé de mots). Cela donne le filtre suivant :

The screenshot shows a search interface titled "Recherche personnalisée". At the top right is a close button (X). Below the title is a search bar containing the text "(nom* de lieu*) PROX Puidoux" with a help icon (?) to its right. Below the search bar are two main sections. The left section contains a checkbox labeled "Rechercher dans les lemmes uniquement:" which is currently unchecked, and a dropdown menu for "Nombre de résultats par page:" set to "20". The right section contains a dropdown menu for "Proximité des mots (en nombre de mots):" set to "10", with a note below it: "Utilisé uniquement si vous faites une recherche de proximité avec le mot-clé PROX." At the bottom right, there are two buttons: "Réinitialiser" and "Rechercher" (with a magnifying glass icon).

Fig. 9. Recherche personnalisée

Après avoir lancé la recherche, on obtient une liste de résultats, dont voici le début :

ansétole (I, 456)

...*anselo*, -a adj, 14, *anselo* Vd 61, *éselo* 3 **Puidoux** n. de l. **Fr.** *anselle*; *anselle* N **PIER.** ...
...d Sent.). Tas de bardeaux (F Gruyères). **Nom de lieu.** *Les Ancelles* Vd
: *Les Ancelles* Vd **Puidoux**; *éz éslo*, **2°** Adjectivt *inselo*, -a, (bo...

1. ARCHER (I, 585)

...*chiers* V Ayent 1279 (GREMAUD, II, 281). **Noms de lieux:** *in Pra Archer* Vd
: *in Pra Archer* Vd **Puidoux** 1215 (MDR, I 2, 148, copie); *En Champ L...*

badou (II, 187)

... Vd, *Badoud* fréquent en F; *badou* Vd, F, **Noms de lieux:** *La Badouda* F Misery; *a la badôda*; *La B...*

BARBETTE (II, 285)

...*tainis* ordres religieux (V Lens J.). **4° Noms de lieux:** *en Barbettaz* Vd
: *en Barbettaz* Vd **Puidoux**; *es Barbettes* Perroy (S&P, 436); *le môc...*

CORDONNIER (IV, 321)

...url.). **2.** Médecin (argot mil. Roux). **3° Nom de lieu.** *Les Cordonniers* Vd
: *Les Cordonniers* Vd **Puidoux**; *ei kordányi*, près, champs, bois, Cf. e...

COURROIE (IV, 444)

11° Noms de lieux. *La Corraye, les Esserts de la Corraye* ...
...*a Corraye, les Esserts de la Corraye* Vd **Puidoux** (cf. S&P, 454); *a la kôrâ, éiz ésê d la...*

Crochette (IV, 580)

3° Nom de lieu. *La Crochette* Vd
: *La Crochette* Vd **Puidoux** (CN 1243); *a la krôtséta*, vigne.

Dézaleyre. (V, 638)

...**ézaleyre.** | S. f. attesté seult dans le **nom de lieu** *la Dézaleyre* Vd
la Dézaleyre Vd **Puidoux**, *en la Desallegre* 1694; *ei [aux] dézalâ...*
Probabl't var. fém. de *Dézaley*, **nom de lieu** à Vd
à Vd **Puidoux**; le suff. est -a r i a.

Mûl.

DIX-SEPT (V, 786)

... il va sur ses dix-sept ans (V Isér.). **Nom de lieu.** *Les Dix-Sept* Vd
: *Les Dix-Sept* Vd **Puidoux**; *lê dytzasa*, vignes. **2° Composés.** **1.** *Dy...*

FOURMI (VII, 833)

4° Noms de lieux. *La Fourmi* Vd
: *La Fourmi* Vd **Puidoux** (CN 1244); *a la froumi*, bâtiment, près, ...

Fig. 10. Résultats de la recherche personnalisée : les noms de lieux de la commune de Puidoux (VD)

Opérateurs logiques

PROX n'est pas le seul opérateur logique disponible. En voici la liste complète :

- ET tous les éléments doivent être pris en compte ;
- OU au moins un des éléments doit être pris en compte ;
- NON l'élément qui suit est exclu ;
- PROX les éléments doivent figurer dans un périmètre limité (paramétrable) ;
- * pour effectuer des troncations (de 0 à n caractères) ;
- () pour personnaliser les priorités lorsque plusieurs opérateurs logiques sont utilisés ;
- « » pour rechercher une suite exacte d'éléments.

Projet d'une base de données iconographiques

Le GPSR possède un riche fonds iconographique, constitué non seulement par les croquis esquissés sur les fiches des correspondants, mais encore par les quelque 2000 dessins originaux du peintre Paul Bœsch effectués lors d'une grande enquête ethnographique menée entre 1943 et 1947 par le Professeur Wilhelm Eglhoff. À cela s'ajoute une vaste collection de photographies prises à la même époque. Une petite partie de cette documentation a été publiée au fil des articles du GPSR, généralement pour illustrer certaines acceptions techniques. Il est prévu qu'en 2020 ces illustrations soient ajoutées au GPSR rétrodigitalisé afin qu'il reflète l'entier de ce qu'on trouve dans la version imprimée. Une fonctionnalité nouvelle sera proposée : les illustrations seront non seulement affichables à partir des articles concernés, mais, après avoir été assorties d'un certain nombre d'informations, elles pourront constituer un domaine de recherche à part entière. En y ajoutant la documentation non publiée du fonds iconographique du GPSR, on aura à disposition une grande iconothèque informatisée. Des images pourront être sélectionnées en fonction de critères de tri variés et s'afficher sous forme de galeries :

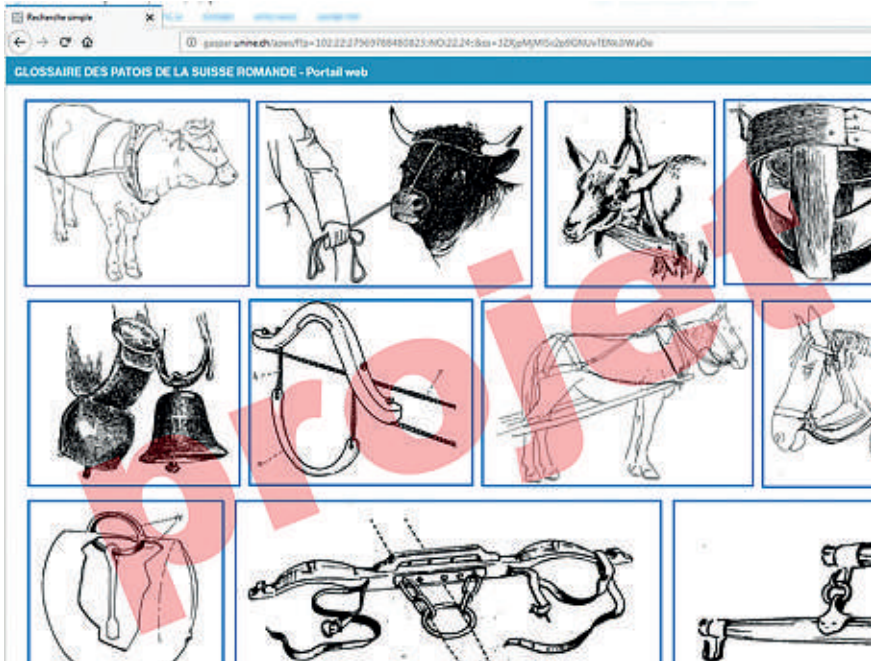


Fig. 11. Projet d'une galerie d'images informatisée

Afficher à l'écran des collections d'images peut être un but en soi. Cela peut aussi être un moyen d'accéder à des informations du GPSR difficiles à atteindre. Prenons l'exemple d'une personne qui voudrait connaître le nom d'un objet trouvé dans une étable et savoir à quoi il sert :

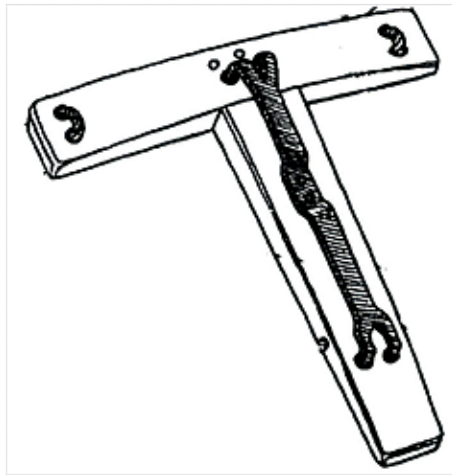


Fig. 12. Vue de face d'un mystérieux objet

En sélectionnant dans l'icône thèque une catégorie d'images comme les bovidés, le harnachement, l'élevage, etc., il aurait sous les yeux une galerie d'images, dans laquelle il pourrait repérer ce qui l'intéresse, en l'occurrence l'illustration figurant au milieu de l'écran :

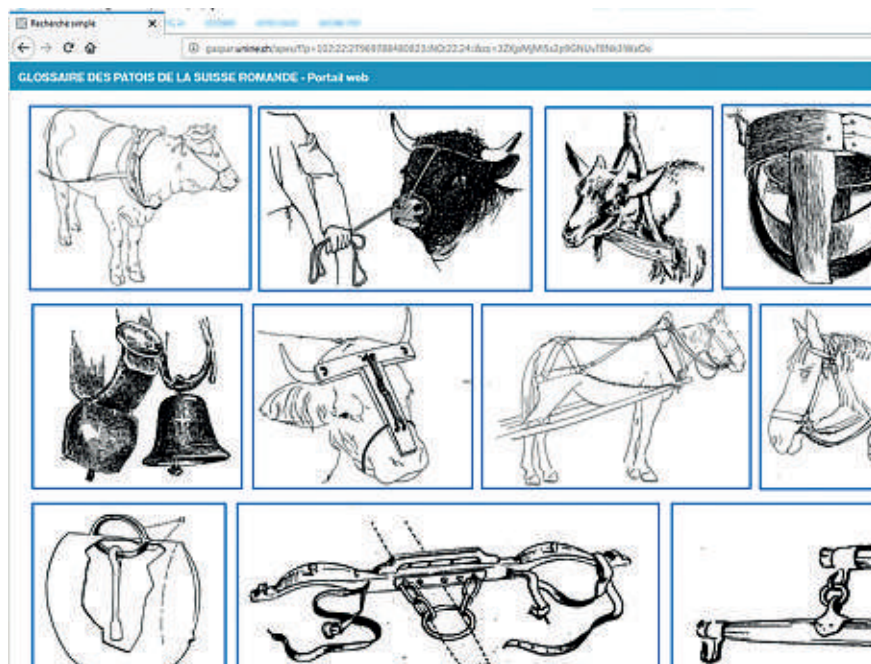


Fig. 13. Prototype d'une page de résultats de la galerie d'images informatisée

Grâce à un lien informatisé, cette illustration pourrait renvoyer directement à l'emplacement où elle se trouve dans le GPSR. Il s'agit de l'article *côte* (sens 3° 13). On y apprend que l'objet s'appelle en patois *kōta*, qu'il est attesté dans les Alpes vaudoises, et qu'attaché aux cornes et au museau des vaches, il sert à les empêcher de briser les haies. En effet, au contact d'un obstacle, un ressort plat en métal se courbe et applique deux paires de pointes sur la tête de l'animal, l'une sur le front et l'autre sur le chanfrein !

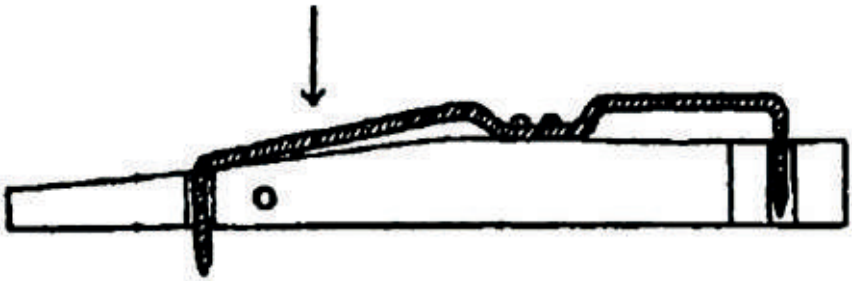


Fig. 14. Vue de côté de l'objet

Grâce à l'informatique, l'accessibilité du GPSR s'est grandement diversifiée. Si l'on ajoute aux différents types de recherches présentées ci-dessus celles – plus linguistiques – exposées dans un autre article (Huber & Greub, à paraître), on peut désormais considérer que l'internaute dispose d'un très large éventail de requêtes différentes pour obtenir les informations qu'il souhaite. Cette foison d'outils de recherche sophistiqués voit son efficacité augmentée par un autre avantage de l'informatique: alors que le lecteur traditionnel du GPSR avait un rayon de sa bibliothèque entièrement occupé par les lourds volumes du dictionnaire, l'internaute moderne peut les emporter pour ainsi dire dans sa poche, partout où il va. Cette gageure est rendue possible grâce à son smartphone et à la conception ergonomique du portail web, qui autorise la consultation sur des écrans de petites dimensions. Le GPSR rétrodigitalisé a ainsi gagné une forte maniabilité, lui permettant non seulement d'être le fidèle assistant du dialectologue, de l'ethnologue ou du toponymiste lors de leurs enquêtes de terrain, mais encore le compagnon du promeneur, de l'alpiniste, du garde-chasse, du guide de montagne, etc. Imaginons qu'un randonneur, féru de trekking dans le Val d'Hérens, quitte le hameau de Villa en direction du Mayen Blanc. Son chemin passera devant un chalet dont la façade porte l'inscription *L'oûra dóou byènyo*. Grâce à son smartphone et à une consultation in situ du GPSR rétrodigitalisé, il

aura la grande satisfaction d'obtenir rapidement la signification de ces quelques mots patois. En guise d'incitation à consulter le portail web du GPSR, nous laissons à chacun le plaisir d'y effectuer cette recherche...³

3 Une interrogation par le biais de *oura*, puis de *byenyo* fournira vite la solution: l'inscription signifie « Le vent du glacier ».

Bibliographie

Brüschweiler, Sabine (1999): *Plantes et Savoirs des Alpes*, Sierre: Éditions Monographic SA.

FEW = Wartburg, Walther von (1922–2002): *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes*, 25 vols., Bonn et al. : Klopp et al.

Huber, Alexandre & Yan Greub (à paraître): « Rétrodigitalisation du Glossaire des patois de la Suisse romande: le rêve devient réalité », in: Aquino-Weber, Dorothee, Federica Diémoz & Maguelone Sauzet, *Quelle place pour les patois en Suisse romande aujourd'hui ?*, 21-22 sept. 2017, Neuchâtel: Éditions Alphil.

Tobi di-j-èlyudzo (1906): *Ouna Fourdèrà dè-j-èlyudzo*, Bulle: Imprimerie commerciale.

SAGW

Die Schweizerische Akademie der Geistes- und Sozialwissenschaften (SAGW) koordiniert, fördert und vertritt die geistes- und sozialwissenschaftliche Forschung in der Schweiz. Ihr gehören 61 Fachgesellschaften und mehr als 20 Kommissionen an. Zudem leitet sie mehrere grosse Forschungsunternehmen. Die SAGW versteht sich als Mittlerin zwischen Forschenden, politischen Entscheidungsträgerinnen und Entscheidungsträgern, Behörden und der Öffentlichkeit. Die SAGW verfügt über ein Budget von rund 16 Millionen Franken. Sie wird von einem Vorstand mit 19 Mitgliedern aus dem universitären Umfeld geleitet. Im Generalsekretariat arbeiten 14 Mitarbeiterinnen und Mitarbeiter.

ASSH

L'Académie suisse des sciences humaines et sociales (ASSH) coordonne, encourage et représente la recherche en sciences humaines et sociales en Suisse. En tant qu'organisation faitière, elle regroupe 61 sociétés savantes et plus de 20 commissions scientifiques. Elle dirige également plusieurs entreprises de recherche de taille importante. L'ASSH fonctionne comme intermédiaire entre les chercheurs et chercheuses, les responsables politiques, les autorités et le grand public. Disposant d'un budget annuel de quelque 16 millions de francs, elle est dirigée par un Comité de dix-neuf membres issus du milieu universitaire. Le Secrétariat général compte quatorze collaboratrices et collaborateurs.

