Projet ANR 2009 CORD 023

TRACE

TRADUCTION ROBUSTE PAR ANALYSE ET CORRECTION D'ERREURS

DES CORPUS D'ERREURS POUR TRACE

Septembre 2012

François Yvon & Natalia Segal

Des corpus d'erreurs pour TRACE

François Yvon et Natalia Segal
LIMSI/CNRS et Softissimo/Reverso

Juillet 2012

Résumé

Ce rapport technique documente les différents corpus d'erreurs orthographiques qui ont été collectés et annotés durant le projet ANR/CONTINT-2009/TRACE dans le cadre du sous-lot 1.2.

On décrit notamment la procédure suivie pour sélectionner le corpus, avant d'analyser plus en détail les principales anomalies détectées et corrigées.

Introduction

Le projet TRACE prévoyait l'acquisition et l'annotation de données représentatives des textes qui sont soumis au moteur de traduction en ligne de Softissimo, et ce pour les deux langues anglais et français. Dans la mesure où les travaux sur l'anglais ont été finalement abandonnés, seuls des échantillons de données françaises ont été collectés.

Ce rapport documente les différentes étapes de la constitution de ce corpus, et inclut également une brève analyse des données collectées, en les comparant en particulier au corpus WiCoPaCo [Max and Wisniewski, 2010] ou plutôt au sous-ensemble correspondant à des révisions visant le plus probablement à corriger des fautes [Wisniewski et al., 2010].

1 Les sources

Les données rassemblées dans le corpus proviennent de quatre sources principales :

— **OBG :** des blogs collectés sur la plate-forme de blogs over-blog [1] ;
— **COR :** des fragments collectés sur l'interface de correction d'orthographe du français de Reverso-Softissimo [2] ;
— **REV :** des fragments collectés sur l'interface de traduction automatique de Reverso-Softissimo [3] ;

1. `www.over-blog.com`
2. `www.reverso.net/orthographe/correcteur-francais/`
3. `www.reverso.net/orthographe/text_translation.aspx?lang=FR`

peut être que je sortirais avec des amies samedi soir donc si vous voulez venir il n'y a pas de problème
Bonjour Lili depuis une semaine nous ne savon pas ce qui vas ce passer avec nos résevation avion car tout nos vol étais avec Mexicana .Comme tu est surement au courant que Mexicana est en difficulté financière ou faillite
Avez vous eu le temps de jeter un oeil sur le ticket ITSM relatif au message d'erreur ci dessous (nous nous sommes entretenu hier)
c'est rare d'avoir une connection avec une personne... j'ai cette connection avec toi..
Je suis Maria Noel Cruz, je travail au bureau de la Com à Total Austral. Je vous contacte parce qu'il y a une fille qui ne trouve pas le link personnel pour accéder à l'enquête Total Survey. Serait-il possible de lui envoyer le link une autre fois ? Elle s'appelle Daniel Franza, son mail est daniela.franza@total

TABLE 1 – Exemples de fragments inclus dans le corpus de correction

	# segments	# phrases	# mots
COR	1 217	2 679	47 971
INT	451	674	14 643
OBG	1 001	3 531	71 332
REV	420	1 124	19 468
all	3 089	8 008	153 414

TABLE 2 – Analyse statistique des corpus d'erreurs

— **INT :** des fragments collectés sur un intranet d'entreprise.

Sur ces sources principales, un certain nombre de critères de sélection ont été appliqués afin que chaque segment :

— comprenne au moins 100 caractères ;
— soit écrit principalement en français ;
— se termine par un point ;
— contienne au moins une faute détectée par le correcteur en ligne de Reverso-Softissimo (une version dédiée du correcteur développée par la société Cordial [4]).

Chaque segment ainsi collecté correspond à une ou à plusieurs phrases. Des exemples de données collectées sont reproduits dans le tableau 1.

Comme on peut le voir sur les quelques exemples de la table 1, les fragments sélectionnés correspondent aussi bien à des phrases contenant soit des erreurs de natures variées (*résevation* ; *comme tu est au courant* ; etc.), soit des passages incorrectement normalisés (absence de majuscule initiale, absence de point final, etc.).

En utilisant des outils standard de segmentation en phrases et en tokens, on obtient les statistiques figurant dans la table 2 qui contient, pour chaque partie du corpus, les nombres respectifs de fragments, de phrases, et de mots.

Au total, le nombre de mots recueillis est conforme avec les objectifs du projet (qui visait un corpus de plus de 100 K mots) ; on notera que près de

4. www.cordial.fr

la moitié des phrases (et des mots) correspond à des données collectées sur la plate-forme OverBlog.

2 Normalisation et production des références

Les données ainsi recueillies ont été corrigées manuellement par un unique annotateur en deux étapes : d'abord un premier échantillon de 1 000 fragments, qui a donné lieu à un échange avec le correcteur et a permis de finaliser un guide recensant les consignes à suivre pour les corrections (donné en annexe). Dans un second temps, les 2 000 fragments restants ont été corrigés.

Une dernière révision finale effectuée en interne à Softissimo a enfin été effectuée pour produire les fichiers de référence.

Environ un quart des fragments recueillis sont accompagnés de commentaires du correcteur, qui détaillent ou justifient les corrections effectuées.

3 Les erreurs de WiCoPaCo

Dans cette section, nous rappelons les principales caractéristiques d'un corpus d'erreurs extrait de WiCoPaCo [Max and Wisniewski, 2010] qui est utilisé pour servir de point de comparaison avec les corpus recueillis et annotés pour le projet. On se reportera à [Wisniewski et al., 2010] pour une présentation plus complète de ce corpus.

Le corpus WiCoPaCo *(Wikipedia Correction and Paraphrase Corpus)* est un corpus de modifications locales extrait des révisions des articles de Wikipédia. Sa construction repose sur l'observation que la plupart des révisions « mineures » d'un article (celles qui ne portent que sur quelques mots) sont des corrections d'erreurs (orthographiques, grammaticales, typographiques, ...) ou des améliorations du style. La construction de ce corpus se compose de deux étapes. Dans une première étape, un ensemble de modifications locales est extrait [5]. Pour cela, nous calculons l'ensemble des différences textuelles entre deux versions d'une même page à l'aide d'un algorithme de recherche de plus grandes sous-séquences communes [6]. L'objectif étant d'extraire des modifications locales, seules les modifications portant sur sept mots au plus sont prises en compte.

Cette première étape permet d'extraire un très grand nombre de modifications locales. Un ensemble de filtres est donc appliqué afin de ne sélectionner que les plus intéressantes. En particulier, les modifications qui ne conservent pas un minimum de mots et qui ne concernent que des signes de ponctuation ou des changements de casse sont exclues. Le premier filtre permet de rejeter (de manière grossière) des corrections « sémantiques » ne conservant pas le sens ; le second permet de limiter la taille du corpus. Les modifications extraites sont ensuite normalisées (notamment en supprimant toutes les informations de mise

5. Les modifications effectuées par les « robots de correction » de Wikipédia sont ignorées (voir `fr.wikipedia.org/wiki/Wikipedia:Bot`).

6. Nous avons utilisé une implémentation identique à celle du programme `diff` standard.

en page), segmentées et sauvegardées dans un format XML. Lors de l'extraction, les informations permettant de faire le lien entre la modification et la page Wikipédia sont conservées, et le contexte (le paragraphe dans lequel la modification est effectuée) est également extrait.

Le corpus WiCoPaCo permet de construire facilement un corpus de corrections orthographiques : il suffit pour cela de distinguer, parmi toutes les entrées du corpus, les corrections des reformulations et du vandalisme. Il est également possible d'extraire la correction de cette erreur, en faisant une hypothèse forte : il faut supposer qu'aucune « petite » modification avec la dernière version d'un article (au moment du téléchargement) n'introduit d'erreur et que le contenu après la modification peut donc être considéré comme une référence correcte. Une étude rapide du corpus montre que cette hypothèse est valable dans la grande majorité des cas.

Le corpus d'erreurs est donc construit en sélectionnant les modifications ne comportant ni signe de ponctuation, ni chiffre, ni nombre écrit en toutes lettres (sauf « un » et « une »), ni plus d'une lettre en majuscule. Ces critères permettent d'écarter des modifications ne portant pas sur l'orthographe du mot et notamment certaines corrections de nature sémantique. Les modifications portant sur plus d'un mot sont également rejetées.

Comme l'on dispose, pour chaque modification, du mot corrigé (avant et après correction), la distinction des erreurs peut être faite quasi automatiquement en appliquant les règles suivantes :

— on conserve les **erreurs lexicales**, correspondant à l'édition d'un mot inconnu en un mot connu, la correction donnant lieu à strictement moins que 6 éditions ;
— on conserve également les **erreurs grammaticales** qui correspondent aux entrées dans lesquelles un mot connu est remplacé par un autre mot connu suffisamment proche (la distance d'édition doit être strictement plus petite que 4) ;
— tous les autres cas sont rejetés.

L'application de ces règles permet d'extraire un corpus de 74 100 erreurs grammaticales et 72 493 erreurs lexicales en contexte.

4 Analyse du corpus

4.1 Décompte grossier des erreurs

Un premier traitement consiste à produire des versions parallèles des segments avant et après correction. À ce stade, chaque segment correspond à une ou à plusieurs « phrases ».

Pour obtenir des premiers décomptes d'erreurs, on applique l'outil sclite [7] qui est usuellement utilisé pour comparer les sorties de systèmes de traitement automatique de la parole avec des références manuelles. Les résultats, venti-

7. www.itl.nist.gov/iad/mig/tools/

Corpus	# Seg	# Wrd	Cor	Sub	Del	Ins	Err	S.Err
WIKI	138 270	8 136 168	97,6	1,7	0,7	0,4	2,8	100,0
COR	1 217	47 974	80,9	11,2	7,9	1,5	20,6	99,6
INT	451	14 647	89,9	7,0	3,2	1,4	11,6	92,7
OBG	1 000	71 209	91,8	5,0	3,2	1,3	9,5	91,5
REV	420	19 483	88,5	7,2	4,4	1,2	12,8	97,4

TABLE 3 – Décomptes d'erreurs

lés par sous-corpus, sont donnés dans le tableau 3, qui inclut également, pour comparaison, les chiffres obtenus avec WICOPACO.

On notera que les corpus sont inégalement bruités, puisque les données recueillies sur l'interface de correction de Reverso contiennent plus de deux fois plus d'erreurs que celles recueillies sur la plate-forme de blogs. On note également qu'une proportion non négligeable de phrases ne contient en fait aucune erreur (8.5% pour le corpus de blogs) : compte tenu du mode de sélection retenu, ces chiffres donnent une idée du taux de fausses alarmes du correcteur automatique utilisé lors de la sélection des fragments de textes. Par comparaison, les données de WICOPACO sont considérablement moins bruitées, avec un taux d'erreur inférieur à 3% ; en revanche, toutes les phrases de corpus présentent une erreur.

4.2 Analyse des erreurs

Une tentative d'analyser plus finement ces différents corpus a également été entreprise. À cet effet, ces corpus ont été traités de manière à (i) produire un alignement entre erreurs et référence au niveau des mots (ii) informer linguistiquement le site de l'erreur en identifiant le POS [8] du mot correct, son lemme, son contexte lexical et syntaxique, ainsi qu'à (iii) produire un alignement au niveau graphique entre les mots en erreur et leur correction, permettant également d'identifier le lieu de l'erreur. Un exemple d'annotation est reproduit dans le tableau 4.

4.2.1 Calcul des annotations

Les annotations sont produites de manière automatique en alignant le texte d'origine, le texte corrigé, ainsi qu'une analyse morpho-syntaxique du texte corrigé. Ceci permet de calculer un certain nombre d'annotations sur les fichiers d'erreur.

POS et POS simplifié : le POS est une des catégories produites par analyse avec le TreeTagger [9]. Dans sa version simplifiée, toutes les étiquettes verbales

8. Calculé avec l'outil TreeTagger accessible depuis `www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

9. Une liste est complète est disponible à l'adresse `www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.

forme erronée	verifié
forme corrigée	vérifié
POS corrigé	VER :pper (Verbe participe passé)
lemme corrigé	vérifier
contexte lexical	... 18/08/10 uniquement. J'ai vérifié ma fiche de paie et ...
contexte syntaxique	PRO :PER VER :pres VER :pper DET :POS NOM
contexte simplifié	VER VER DET
transformation(s) graphique(s)	e→é, ve→vé, er→ér, ver→vér
erreur initiale	0 (NON)
erreur finale	0 (NON)
erreur de capitalisation	0 (NON)
erreur d'accentuation	1 (OUI)
distance d'édition	1
erreur grammaticale	0 (NON)

TABLE 4 – Exemples d'annotations du corpus d'erreurs

sont projetées sur la seule étique VER, et le même traitement est appliqué aux DET et PRO, de manière à limiter un peu le nombre de catégories. Dans certains cas, on fournit également un POS plus détaillé, respectant le format de catégories étendu du projet Multext.

Contexte syntaxique : séquence comprenant en plus de la catégorie de l'erreur les catégories des deux mots à gauche et à droite, lorsqu'ils existent. Dans sa version simplifiée, le contexte est tronqué à un seul mot, et on utilise les POS simplifiés.

Erreurs de capitalisation : qualifie les situations dans lesquelles les chaînes originales et corrigées sont identiques lorsque l'on transforme la majuscule initiale de l'une en minuscule.

Erreurs d'accent : qualifie les situations dans lesquelles les chaînes originales et corrigées sont identiques lorsque tous les accents sont supprimés (par ex. un 'é' devient 'e', etc.).

Erreurs d'accord : qualifie les situations dans lesquelles l'erreur porte sur le suffixe et appartient à un ensemble restreint de patrons prédéfis (ajouter supprimer un 's' ou un 'e', confondre 't' et 's', etc.) Cette liste donne une première approximation des fautes d'accord. On distingue non-faute (0), faute d'accord (1), autre erreur en suffixe (2), autre erreur 3.

Erreurs grammaticales : qualifie les situations dans lesquelles le mot originel comme le mot corrigé sont deux mots présents dans la liste des formes du français recensée par Boula De Mareuil et al. [2000]. On distingue quatre situations : les plus communes sont quand le mot erronné est inconnu (respectivement connu) et sa correction connue (code 1, respectivement 3). Il arrive également que les deux soient inconnus (code 0), et plus exceptionnellement qu'un mot connu soit corrigé par un mot inconnu (code 2).

	COR	INT	OVB	REV	Total	WIC
nb. phrases	2478	668	3325	1105	7576	310617
nb. tokens	45609	14700	70333	19250	149982	8018550
longueur moyenne	18,40	22,00	21,15	17,42	19,79	25,81
nb. d'erreurs	9966	1924	7302	2921	22113	150431
par phrase	4,02	2,88	2,19	2,64	2,91	0,48
% d'erreurs	21,86	13,08	10,39	15,17	14,80	2,87
% délétions	1,36	1,32	1,07	1,29	1,21	0,06
% substitutions	13,54	8,76	6,55	9,88	9,31	1,70
% insertions	6,96	3,01	2,75	3,99	4,21	0,11

TABLE 5 – Distribution des erreurs par type et par corpus

Localisation : un drapeau est levé lorsque l'erreur porte sur la première (respectivement dernière) lettre du mot.

Distance : la distance d'édition entre la forme originale et corrigée.

Participes passés : un ensemble de descripteurs concerne finalement les formes étiquetées au participe passé pour lequel on calcul l'auxiliaire le plus proche (dans le passé), la distance à cet auxiliaire, ainsi qu'un drapeau qui repère les formes pronominales.

À partir de ces annotations, il est possible de calculer des statistiques plus précises concernant les erreurs contenues dans ces différents corpus et d'en contraster ainsi le contenu.

4.2.2 Analyse des fichiers annotés

Après segmentation en phrases, une analyse statistique des différents corpus donne les valeurs [10] dans le tableau 5.

Les statistiques reportées dans le tableau 5 confortent l'analyse faite au vu du tableau 3, à savoir en particulier que les données capturées sur l'interface de correction de Reverso sont de loin les plus bruitées. Par comparaison, les données soumises au système de traduction sont un peu moins souvent erronées, même si le taux d'erreur reste très important. Par contraste, il apparait clairement que la procédure de constitution de WICOPACO a conduit à un corpus aux caractéristiques très différentes.

4.2.3 Localisation et typage des erreurs

Dans ce paragraphe, nous étudions les erreurs du point de vue graphique en analysant les principaux types de substitution, ainsi que leur localisation. Les

10. Le nombre de tokens est entendu **après alignement au niveau des tokens** : on prend toujours la longueur *maximum de la phrase originale*.

	COR	INT	OVB	REV	Total	WIC
initiale	45,44	44,91	50,54	52,31	51,98	16,08
finale	53,55	45,14	52,15	49,63	51,78	41,21
initiale ou finale	83,63	77,54	87,97	86,59	84,90	52,86
initiale et finale	15,37	12,51	14,72	15,35	14,88	4,43
capitalisation	20,97	21,45	25,97	28,65	23,71	0
accentuation	14,23	17,63	9,87	14.45	13,14	21,78
grammaticale	54,76	54,39	58,68	62,98	57,14	42,61
longueur moyenne	5,19	5,95	5,35	5,10	5,30	6,63
distance moyenne	1,70	1,67	1,64	1,48	1,69	1,24

TABLE 6 – Distribution des erreurs par type et par corpus

principaux résultats globaux et par sous-corpus portant sur ces quelque 13 970 instances sont résumés dans le tableau 6 et, de nouveau, contrastés avec les 136 104 erreurs de WICOPACO.

On note en particulier la très forte proportion d'erreurs se situant en début ou en fin de mots, la première étant partiellement expliquée par la proportion importante d'erreurs relatives à la capitalisation (ou nom) de l'initiale du mot. Ceci explique peut-être la forte proportion d'erreurs qualifiées de grammaticales, qui correspondent à plus de 55% des erreurs. Dans de nombreux cas, il ne s'agit en fait que d'un problème de capitalisation, et l'erreur ne correspond donc pas réellement à un problème de syntaxe. Si l'on omet ces configurations, le nombre d'erreurs de syntaxe tombe à 38,34%, se rapprochant du taux observé sur WICOPACO.

Autre fait marquant, la longueur moyenne des mots erronés est très sensiblement supérieure à la longueur moyenne des mots calculée sur l'ensemble du corpus (5,3 caractères contre 4,3) : les erreurs concernent en moyenne des mots longs. Cette tendance est encore plus marquée sur les données de WICOPACO ($6,63$ caractères en moyenne pour les mots erronés contre $4,63$ pour l'ensemble des mots). Notons finalement qu'en dépit de fluctuations entre corpus, le nombre d'édition moyen est relativement élévé : plus près de deux que de un en moyenne. Une conséquence est qu'en ne considérant que les voisins à distance 1, on ne corrige, suivant les corpus, qu'entre 70% et 78% des erreurs ; prendre en compte les voisins à distance 2 ne permet de corriger au mieux que 90 % des erreurs et il faut considérer jusqu'à 5 erreurs par mot pour traiter environ 95% des erreurs. Le contraste est de nouveau très marqué avec WICOPACO pour lequel plus de 95 % des erreurs sont à distance inférieure à 2.

4.2.4 Analyse des transformations graphiques

Dans cette section, nous étudions sommairement les principales transformations orthographiques qui sont les erreurs. On en dénombre dans notre corpus un total de 2 887, parmi lesquelles seules 621 ont une fréquence supérieure à 1 et 380 une fréquence supérieure à 2 : leur distribution est donc très inégale. La

liste des 64 transformations les plus fréquentes est dans le tableau 7. On notera que ces transformations ne couvrent qu'un peu plus de 60% des transformations attestées.

_→s	e→é	a→à	j→J	s→_	_→e	_→-	m→M
7,1	10,9	14,3	17,7	20,7	23,1	24,7	26,2
e→è	e→_	c→C	b→B	_→’	l→L	e→E	er→é_
27,7	29,2	30,6	31,8	33,1	34,3	35,4	36,5
...→_	p→P	s→S	a→A	t→s	t→T	i→I	s→t
37,6	38,6	39,6	40,6	41,5	42,3	43,2	44,0
e→ê	t→_	é_→er	_→t	é→è	P→p	d→D	e→’
44,7	45,5	46,2	46,8	47,5	48,2	48,8	49,3
n→N	o→ô	f→F	C→c	S→s	c→ç	,→.	M→m
49,9	50,4	50,9	51,4	51,9	52,3	52,7	53,2
s→ç	A→a	,→...	_→r	u→ù	v→V	L→l	__→nt
53,6	54,0	54,4	54,8	55,2	55,6	56,0	56,3
D→d	e→a	c→s	J→j	’→e	%→_	__→le	o→O
56,7	57,1	57,4	57,8	58,1	58,4	58,7	59,0
.→?	s→c	E→e	a→e	R→r	a→â	_→x	u→U
59,2	59,5	59,8	60,1	60,3	60,6	60,9	61,1

TABLE 7 – Les transformations graphiques les plus fréquentes. On liste la fréquence cumulée des transformations, de la gauche vers la droite puis de haut en bas.

Comme attendu, les transformations les plus fréquentes portent sur des changements simples et linguistiquement transparents : oubli d'un 's', oubli d'un accent, confusion entre 'à' et 'a', etc. On note la forte représentation des transformations qui correspondent à la restauration d'une majuscule (la transformation de 'j' en 'J' est la quatrième plus fréquente dans ce corpus), ainsi qu'un nombre important de changements qui concernent la ponctuation. Pour comparaison, la distribution des erreurs pour WICOPACO (voir le tableau 8) est nettement plus resserrée, même si, à l'exception des capitalisations, les principales erreurs sont sensiblement les mêmes (ajout ou omission d'un 's', ajout ou omission d'un 'e', erreur d'accent, etc.).

_->s	_->e	s->_	e->é	E->É	e->_	oe->œ	_->n
9,42	15,43	21,05	24,99	28,47	31,81	34,22	36,05
é->è	_->t	_->r	i->î	t->_	a->e	é->e	a->à
37,83	39,52	41,20	42,56	43,89	45,17	46,41	47,634162
_->i	e->a	a->â	l->_	n->_	__->nt	_->m	r->_
48,84	49,98	51,09	52,15	53,08	54,00	54,91	55,76

TABLE 8 – Les transformations graphiques les plus fréquentes de WICOPACO.

POS	COR	INT	OVB	REV
ABR	0,17	0,19	0,21	0,15
ADJ	4,07	4,33	4,83	4,27
ADV	5,85	4,62	5,20	6,55
DET	9,04	10,68	10,10	9,14
INT	0,17	0,04	0,10	0,15
KON	4,48	4,42	4,16	4,97
NAM	3,42	3,79	4,16	2,96
NOM	15,98	19,47	18,54	15,84
NUM	1,76	1,74	1,75	1,25
PRO	14,25	10,08	9,92	14,57
PRP	11,56	15,94	14,38	11,85
PUN	7,43	4,15	7,68	5,76
SENT	5,29	4,49	4,68	5,53
SYM	0,029	0,06	0,09	0,04
VER	16,48	15,99	14,20	16,98

TABLE 9 – Distribution des étiquettes morpho-syntaxiques

4.2.5 Analyse des erreurs par parties du discours

Dans cette section, nous nous intéressons à décrire les erreurs du point de vue des parties du discours qui sont affectées. Le tableau 9 donne la distribution des étiquettes morpho-syntaxiques au sein de chaque sous-corpus.

Comme attendu, les catégories les plus fréquentes sont, dans l'ordre, : les verbes, les noms, les pronoms, prépositions et déterminants. À l'autre extrêmité du spectre, les abréviations et les interjections ne sont que marginalement représentées.

Les résultats du tableau 10 montrent une relative variabilité dans la distribution des erreurs par POS. On note en particulier que les erreurs se distribuent majoritairement sur les noms et sur les verbes, et, moins fortement, sur les adjectifs (qui sont toutefois moins bien représentés dans le corpus). Inversement, déterminants et prépositions sont beaucoup moins affectés par les erreurs. Au sein de la catégorie verbe, on note également de fortes variations entre les formes à l'infinitif, qui sont correctes dans plus de 92 % des cas, et les conjugaisons moins courantes (conditionnel, subjonctif, passé simple) pour lesquelles près de 3 formes sur 4 sont erronées. Autre résultat marquant de cette analyse : l'importance (relative) des erreurs portant sur les signes de ponctuation et corrolairement, sur les frontières de phrases (catégorie 'SENT'). Ceci signale à nouveau que ces données rendent compte des principales erreurs orthographiques, mais également de difficultés liés à la normalisation des énoncés.

Une analyse plus fine des contextes d'erreur est exposée dans le tableau 11, qui liste tous les contextes apparaissant au moins 100 fois, sur un total de plus de 1 000 contextes différents. En dépit des relativement faibles effectifs de ce tableau, il est possible d'y retrouver certaines des régularités mentionnées ci-dessus.

	COR			INT			OVB			REV			ALL		
	ins	sub	cop	ins	sub	cop	ins	sub	cop	ins	sub	cop	ins	sub	cop
ABR	26,58	31,65	41,77	10,71	14,29	75,00	4,70	18,12	77,18	10,71	7,14	82,14	11,97	20,42	67,61
ADJ	0,98	22,01	77,01	0,32	15,92	83,76	0,65	10,27	89,08	0,25	13,69	86,07	0,66	14,46	84,87
ADV	4,10	12,20	83,70	1,34	8,06	90,60	1,49	6,43	92,07	2,09	7,31	90,60	2,41	8,56	89,03
DET	1,70	5,01	93,29	1,16	4,52	94,32	0,70	2,35	96,96	0,40	4,38	95,22	0,99	3,58	95,43
INT	6,41	51,28	42,31	0,00	33,33	66,67	4,48	26,87	68,66	3,45	44,83	51,72	5,00	40,56	54,44
KON	1,44	5,84	92,72	0,62	5,15	94,23	0,48	3,90	95,61	0,53	4,56	94,92	0,80	4,72	94,48
NAM	2,01	38,18	59,81	0,73	13,82	85,45	0,93	14,68	84,39	1,60	36,12	62,28	1,28	23,29	75,43
NOM	0,90	18,63	80,47	0,78	13,21	86,01	0,47	8,71	90,82	0,63	12,26	87,11	0,64	12,36	87,00
NUM	3,79	14,29	81,92	2,37	13,04	84,58	1,32	8,47	90,21	2,53	8,02	89,45	2,32	10,73	86,94
PRO	2,53	13,86	83,61	1,50	9,71	88,78	1,27	8,21	90,51	1,12	11,05	87,83	1,73	10,85	87,42
PRP	2,37	8,16	89,48	1,17	3,80	95,03	0,91	2,95	96,14	0,84	6,62	92,54	1,32	4,84	93,85
PUN	42,25	5,29	52,45	18,60	2,16	79,24	14,41	2,83	82,77	33,15	3,84	63,01	25,59	3,69	70,72
SENT	40,91	4,75	54,35	29,95	5,99	64,06	19,05	3,23	77,72	23,31	3,90	72,79	27,72	4,06	68,22
SYM	46,15	0,00	53,85	66,67	0,00	33,33	60,94	3,12	35,94	42,86	0,00	57,14	58,06	2,15	39,78
VER	1,59	19,17	79,24	0,56	11,21	88,23	0,80	9,46	89,74	0,93	13,55	85,52	1,05	13,36	85,58

TABLE 10 – Distribution des erreurs par POS et corpus (ins=insertion ; sub=substitution, cop=copie).

Corpus Trace				
SENT PRO VER 3,61	DET NOM PRP 2,57	VER VER PRP 1,81	PRO VER PRP 1,69	PRP NOM PRP 1,57
PRO VER ADV 1,47	PRO VER VER 1,35	DET NOM SENT 1,34	PRO VER DET 1,28	SENT DET NOM 1,27
DET NOM PUN 1,25	PRP NOM SENT 1,14	PRP NOM PUN 1,02	PUN PRO VER 0,99	SENT PRO PRO 0,99
ADV VER ADV 0,96	PRP NAM PUN 0,89	PRO PRO VER 0,88	SENT NOM PUN 0,88	VER VER DET 0,85
Corpus WiCoPaCo				
DET NOM PRP 5,69	VER VER PRP 3,51	PRP NOM PRP 3,35	DET NOM ADJ 2,45	PRO VER PRP 1,98
NOM ADJ PRP 1,87	NOM VER PRP 1,78	DET NOM VER 1,52	PRP NOM PUN 1,51	PRP NOM ADJ 1,50
NOM ADJ PUN 1,48	DET NOM PUN 1,43	PRO VER DET 1,26	ADV VER PRP 1,02	VER VER DET 0,98
NOM ADJ SENT 0,94	PRP DET NOM 0,93	PUN VER PRP 0,92	PRO VER ADV 0,88	PRP NOM VER 0,88

Table 11 – Les 20 contextes syntaxiques les plus « erreurogènes »

Ainsi, la principale cause d'erreur porte sur les pronoms en position initiale de phrase, on suppose du fait d'un problème de capitalisation. Les principaux contextes d'occurrences des verbes ('VER VER PREP' 'PRO VER PRP', etc.) et noms sont logiquement bien représentés ('DET NOM PRP', 'DET NOM PUN', 'DET NOM VER') sans qu'il soit possible de tirer des conclusions plus précises.

Une dernière illustration de l'analyse des erreurs par catégorie est donnée dans la table 12. Lorsque les erreurs portent sur des mots connus, il est parfois possible d'inférer une étiquette plus fine pour le mot corrigé, et, lorsque l'information n'est pas ambiguë, également pour le mot erroné. On peut en déduire une liste des confusions les plus fréquentes au niveau des catégories principales. Cette information a pu être recalculée pour environ 7 000 erreurs.

Si la situation la plus fréquente est celle où la catégorie est inchangée (ce qui recouvre en particulier les erreurs de capitalisation), on retrouve dans cette table les principales confusions morphologiques (erreur de nombre, de personne, ou de genre).

5 Distribution et formats

Il est prévu de distribuer l'ensemble des corpus collectés et annotés lors du projet, dans une version brute ainsi que dans une version enrichie. L'annexe B contient un descriptif (en anglais) du format de distribution retenu et de la structure de l'archive.

Changement	Fréq.	Commentaire
==	2747	POS inchangé (par ex. capitalisation)
Ncmp→Sp	478	essentiellement a → à
Ncfs→Ncfp	176	singulier/pluriel (noms féminins)
Ncms→Ncmp	171	singulier/pluriel (noms masculins)
Vmip3s-→Vmip1s-	134	3ème/1ère personne (présent)
Vmn—-→Vmps-sm	134	infinitif/part. passé
Vmps-sm→Vmps-sf	101	masculin/féminin (part. pass)
Ncmp→Ncms	90	pluriel/singulier (noms masculins)
Afpfs→Afpfp	89	singulier/pluriel (adjectifs féminins)
Vmps-sm→Vmn—-	88	part passé/infinitif
Vmps-sm→Vmps-pm	65	singulier/pluriel (part. pass)
Afpms→Afpfs	64	masculin/féminin (adjectif sing)
Ncfp→Ncfs	64	singulier/pluriel (noms féminins)
Ds3fss-→Pd-ms–	61	
Vmps-sf→Vmps-sm	52	féminin/masculin (part. pass)
Cc→Pr-fp–	51	
Ncms→Pp1msn-	51	
Vmip1s-→Vmps-sm	51	fini/participe
Afpms→Afpmp	49	singulier/pluriel (adjectifs)
Vmip1s-→Vmip2s-	46	1ère/2ème personne (présent)
Vmip3s-→Vmip3p-	46	singulier/pluriel (present)
Vaip3s-→Vaip2s-	43	3eme/2eme pers (auxiliaire)
Afpfs→Afpms	38	masculin/féminin (adjectifs)
Ncmp→Rgp	37	
Pp3mpn-→Pp3mpd-	37	
Ncms→Vmip1s-	36	
Vmif1s-→Vmcp1s-	36	cond/futur (1ere personnne)
Vmip1s-→Vmip3s-	34	1ère/3ème personne (pass. simple)
Vmps-pm→Vmps-sm	33	pluriel/singulier (part. pass)
Ncms→Da-ms-d	30	

TABLE 12 – Principales confusions "grammaticales"

Références

Philippe Boula De Mareuil, François Yvon, Christophe D'Alessandro, Véronique Aubergé, Jacqueline Vaissière, and Angélique Amelot. A french phonetic lexicon with variants for speech and language processing. In *Second International Conference on Language Resources and Evaluation (LREC)*, pages 273–276. Athens (Greece), 2000.

Aurélien Max and Guillaume Wisniewski. Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the Language Resources and Evaluation Conference*, Marrakech, Morocco, 2010.

Guillaume Wisniewski, Aurélien Max, and François Yvon. Recueil et analyse d'un corpus écologique de corrections orthographiques extrait des révisions de wikipédia. In *Actes de la 10eme conférence sur le Traitement Automatique des Langues Naturelles (TALN'2010)*, Montréal, Canada, 2010.

A Consignes pour la correction

Tous les segments comportant un nombre d'erreurs trop important (de type SMS par exemple) ou n'étant pas en langue française ne sont pas à traiter. Il faudra néanmoins signaler leur mise à l'écart par un commentaire.

Toute correction différente d'une correction orthographique standard ou autre qu'une correction de ponctuation standard devra être rapportée. Le rapport des modifications est également à inscrire dans les commentaires.

Sigles : Les sigles doivent être écrits en capitales et sans point (SNCF, FMI) à l'exception des sigles lexicalisés (sida, laser, Otan) qui peuvent être à fois en minuscules et en majuscules.

Ponctuation et typographie :

— corriger les espaces manquants ou superflus. De cette façon, «septembre1998» doit être séparé afin d'obtenir «septembre 1998».

— apparier les parenthèses et les guillemets. Pour ces derniers, ne pas corriger leur style (« vs ")

— ajouts de virgules lorsqu'elles sont nécessaires à la compréhension de la phrase.

— ajout des traits d'unions. Il faut également ajouter les traits d'unions lorsqu'ils sont nécessaires à la bonne construction d'une question. D'où ajout du point d'interrogation en cas de question.

— ajout de l'apostrophe lorsqu'il est absent mais pas de correction s'il est droit et non courbe.

En ce qui concerne les accents, ils doivent TOUS être ajoutés sur les lettres en minuscules. En ce qui concerne les lettres capitales, les accents sur ce type de lettres sont facultatifs. Ainsi, « DEMENAGEMENT » ne peut pas être considéré comme une faute.

Négations : dans le but d'avoir une aussi bonne correction que possible, les négations sont à corriger. Ainsi, «je suis pas d'accord», doit être remplacé par «je ne suis pas d'accord».

Abréviations : elles ne sont pas à remplacer si elles sont suffisamment claires et pourraient constituer des entrées du dictionnaire. «stp» ne doit pas être corrigé alors que «tjrs» doit l'être.

Anglicismes : les anglicismes doivent être tolérés et ne pas être corrigés à condition qu'ils soient passés dans le langage courant et considérés comme des emprunts linguistiques. Ainsi, on peut conserver «e-mail» alors que «des réponses non-judgmental» doit être corrigé par «des réponses sans porter de jugement».

Noms propres : des corrections doivent être apportées aux noms propres s'il s'agit de noms ou de lieux géographiques susceptibles d'être inclus dans un dictionnaire. De cette manière «Sarkosi» doit être corrigé, tout comme «le Rein» (fleuve). L'emploi de leur traduction dans des segments en français n'est pas toléré. De cette manière, «je suis partie en Uzbekistan» est à corriger par «je suis partie en Ouzbekistan». Toutefois, si la traduction est dans la langue d'origine du mot, «Beijing» pour «Pékin», l'emploi est toléré.

Numéraux-ordinaux : ne pas modifier les numéraux-cardinaux. 1 est aussi

valable que 1er et que premier. Les chiffres romains sont également acceptés. Cependant il faut corriger toutes les fautes d'orthographe et de typographie lorsqu'elles se présentent.

Paronymes : les paronymes sont à corriger. De cette manière, «une lettre en bon uniforme» doit être corrigée par «une lettre en bonne et due forme». Reformuler des phrases : Il arrive que l'erreur provienne d'une mauvaise formulation de la phrase ou d'une mauvaise utilisation d'un mot. Ces erreurs sont à corriger. Il convient néanmoins de différencier entraînant une mauvaise compréhension des erreurs qui ne l'altèrent pas. Ainsi «Vous neccesairez : 4 œufs, une tasse de sucre» est à corriger par «vous aurez besoin de ...» alors que « est-ce que tu vas là bas ? » n'est pas à corriger par « vas-tu là-bas ? »

Ligatures : Les « ae » et « oe » n'ont pas à être corrigés. « Soeur » est aussi acceptable que « Sœur ».

Lexique : les mots grossiers, injurieux, familiers, argotiques, populaires n'ont pas à être modifiés ou supprimés. Lorsqu'un mot présente une erreur mais que même le contexte n'aide pas à la bonne compréhension de ce mot, choisir une alternative. Cependant, il faut indiquer dans les commentaires que la correction entraîne peut être une modification du lexique. Ex : «ça exige une connaissance du thème et je ne connais rien de cette tempe.» pour « de cette trempe ».

Doublons : Si un mot est répété plusieurs fois (sauf pronom réfléchi «nous nous»), le mot en trop doit être supprimé.

Majuscules : Si un nom propre est présent, il doit commencer par une majuscule. Dans le cas où tout un segment serait en majuscule, il ne faut pas le modifier. Cependant, si l'on rencontre des cas tels que « étudiant à l'IUT de NAntes », il faut corriger par « étudiant à l'IUT de Nantes ». De plus, si une lettre capitale est utilisée pour un nom commun ou adjectif, la modification n'est pas à faire ex : «Etude sur la Voiture Verte ».

Pronoms personnels : dans le cas d'apparition de pronoms personnels tronqués comme «j'» ou «t'», il faut leur redonner leur forme entière. Ainsi, « j'vais à la piscine » est corrigé par «je vais à la piscine».

Nouvelle orthographe : les nouvelles règles orthographiques sont acceptées.

Marques déposées : les majuscules et les symboles des marques déposées ne sont pas à rajouter.

B Distribution

```
LHCYR 1.0 (c) 1998 V.V.Zhytnikov (vvzhy@td.lpi.ac.ru)
The LaTeX 2e style for Russian typesetting in bilingual environment
Available at CTAN:/tex-archive/macros/latex/contrib/supported/lhcyr
Main CTAN sites:   ctan.tug.org   ftp.dante.de   ftp.tex.ac.uk

This package is free. You may modify and use it for whatever
```

purpose you want. But you are not allowed to redistribute modified version under the same name.

CONTENTS

1. GENERAL FEATURES

LaTeX 2e style lhcyralt is intended for typesetting Russian or bilingual English-Russian documents. Other latin-alphabet languages will probably work as well but none where tested.

The style is based on the so called lh-fonts in the Alternative (AKA codepage 866) encoding which is standard for Russian in the MS-DOS. The style works by replacing each standard TeX computer modern text font by the analogous lh-font (e.g. cmr10 -> lhr10 etc replacing all 'cm' by 'lh'). Upper part of an lh-font (character codes 0-127) exactly coincides with the corresponding cm-font and its lower part (codes 128-255) contains Russian alphabet in the Alternative encoding together with some extra symbols such as Russian number sign and angular double quotes.

The installation described below is emTeX specific but provides rather good idea how the package can be installed with other TeX distributives. I assume that you have Metafont and it can automatically generate fonts upon request.

The lhcyralt package consists of the following files:

```
README             - this file

lhcyralt.sty       - LaTeX 2e style file

ot1lhdh.fd         - LaTeX 2e font driver files
ot1lhfib.fd
ot1lhfr.fd
ot1lhr.fd
ot1lhss.fd
ot1lhtt.fd
ot1lhvtt.fd
```

```
hyphen.cfg          - LaTeX 2e hyphenation configuration file
rhyphen.tex         - Russian hyphenation patterns by A.Slepukhin
                      in the Alternative encoding

dvidrv.mfj          - mfjob configuration files for automatic
lhjob.mfj             font generation

karabas.tex         - sample and test files
kniga.tex
otchet.tex
pismo.tex
rusfonts.tex
statya.tex
```

2. INSTALLING FONTS

First of all make sure that you have metafont sources for standard cm (Knuth's computer modern) fonts. They are used for the upper part of the lh-fonts. If not then get them at

```
CTAN: /tex-archive/fonts/cm
```

Next get latest lh-fonts distributive. It available at

```
CTAN: /tex-archive/fonts/cyrillic/lh
ftp:  ftp.vsu.ru/pub/tex
```

Create some temporary directory and unpack .zip or .tgz lh-font archive with the directory structure (use -d switch with pkunzip.exe). Go into the subdirectory \tex and create in it directory \tex\wrk (if absent). Return to \tex and run plain TeX on the 01cm-lh.tex

```
tex 01cm-lh.tex
```

This will create several files in the \tex\wkr directory. Copy all *.mf files from the directory \mf (about 750K) and from the directory \tex\wrk into the place where your Metafont looks for *.mf source files.

Now we have to tell mfjob how to generate lh-fonts replace your dvidrv.mfj file with the dvidrv.mfj file provided with the lhcyralt package. If you have some specific dvidrv.mfj containing your local nonstandard definitions you must extract definition for lh-fonts generation from lhcyralt's dvidrv.mfj and manually insert it into your local dvidrv.mfj file.

To create *.tfm files for lh-fonts run mfjob on lhjob.mfj file

```
mfjob lhjob.mfj
```

This may take from several minutes up to an hours depending on

the computer. Copy created *.tfm files into the directory where your TeX looks for the *.tfm.

If you have testfont.tex you may test lh-fonts by running
  tex testfont
when prompted for font name type
  lhr10
  \table\end
and print or preview the resulting tesfont.dvi.

3. GENERATING FORMAT FILE

To enable correct Russian hyphenation you have to create a new LaTeX 2e format file. With the lhcyralt style both English and Russian languages are represented as one combined language (from the TeX's point of view). With such setup you can freely mix English and Russian words with correct hyphenation for both languages without any language switching commands. This approach is orthogonal to the method used by the LaTeX's babel package. With babel each language is stored separately (you may have more than two languages but must use explicit language switching commands). If the babel package is installed on your system you must disable it (at least temporary during new LaTeX format generation).

Rename any hyphen.cfg file to something else (This file may resides in the babel's directory). Copy the
  hyphen.cfg and rhyphen.tex
files from the lhcyralt distributive into directory where TeX can find them. Make sure that you have the file hyphen.tex and this file is the original Knuth's US-English hyphenation patterns file. Sometimes with babel installed this patterns are renamed to ushyph1.tex or something else. If so or in the case then you want to use some other English hyphenation tables (say UK-patterns) edit hyphen.cfg and replace {hyphen.tex} by {ushyph1.tex} or any other pattern file name.

Create LaTeX format file by running

```
tex386.exe /i /o /8 /mt15000 latex.ltx
```

You'll get new latex format file latex.ftm. Use it with the same /mt switch as during format generation:

```
tex386.exe /mt15000 &latex
```

4. USING lhcyralt STYLE

Now copy lhcyralt.sty and all *.fd files into the directory where your TeX can find them and include the line

```
\usepackage{lhcyralt}
```

into the LaTeX document header. With this style you can freely type English and Russian in the Alternative encoding. Besides you get three extra commands

```
\No  \<  \>
```

for Russian number sign and Russian angular double quotes.

With the additional [russtyle] option enabled

```
\usepackage[russtyle]{lhcyralt}
```

all English texts such as "Chapter", "Appendix" are replaced by the corresponding Russian translations "Glava","Prilozhenie" etc.

You may test lhcyralt style by latexing test files

- karabas.tex
- rusfonts.tex

and sample files

- kniga.tex
- otchet.tex
- pismo.tex
- statya.tex