Computer-Aided Error Annotation

A New Tool for Annotating Arabic Error

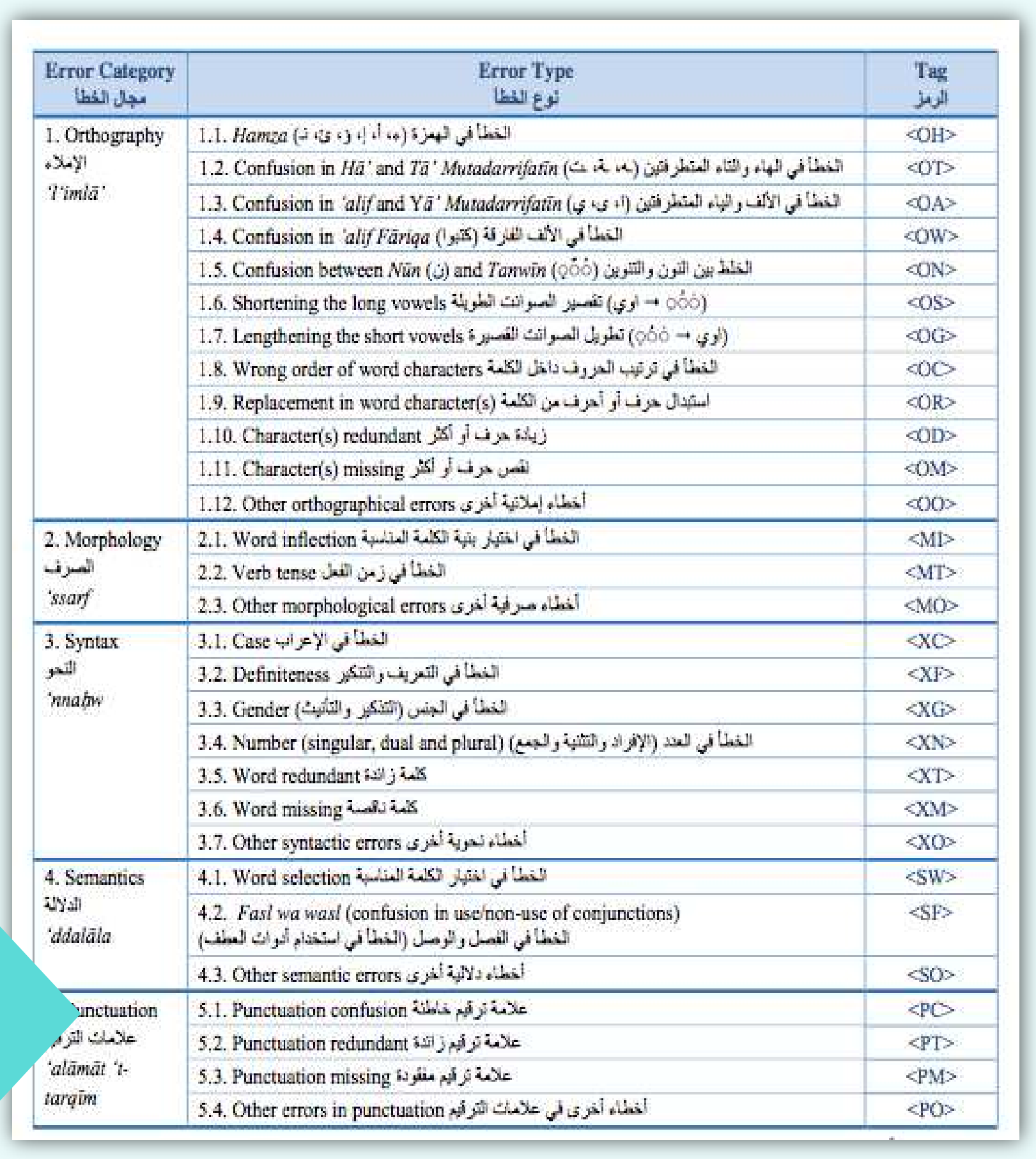
Abdullah Alfaifi and Eric Atwell
University of Leeds {scayga, e.s.atwell}@leeds.ac.uk

UNIVERSITY OF LEEDS

WHY THIS TOOL?

Existing tools for annotating errors in learner corpora are developed for languages other than Arabic.
Thus, this poster introduces a new tool for computer-aided error annotation in Arabic learner corpora.

The new tool includes the Arabic Error Taxonomy (Alfaifi et al., 2013) which was designed specifically for Arabic



Error Category مجال الخطأ	Error Type نوع الخطأ	Tag الرمز
1. Orthography الإملاء 'l'imlā'	1.1. *Hamza* (ء، أ، إ، ؤ، ئ، ئـ) الخطأ في الهمزة	<OH>
	1.2. Confusion in *Hā'* and *Tā' Mutadarrifatīn* (ـه، ـة، ـت) الخطأ في الهاء والتاء المتطرفتين	<OT>
	1.3. Confusion in *'alif* and *Yā' Mutadarrifatīn* (ـا، ـى، ـي) الخطأ في الألف والياء المتطرفتين	<OA>
	1.4. Confusion in *'alif Fāriqa* (كتبوا) الخطأ في الألف الفارقة	<OW>
	1.5. Confusion between *Nūn* (ن) and *Tanwīn* (ـً ـٍ ـٌ) الخلط بين النون والتنوين	<ON>
	1.6. Shortening the long vowels تقصير الصوائت الطويلة (اوي → ـَ ـُ ـِ)	<OS>
	1.7. Lengthening the short vowels تطويل الصوائت القصيرة (ـَ ـُ ـِ → اوي)	<OG>
	1.8. Wrong order of word characters الخطأ في ترتيب الحروف داخل الكلمة	<OC>
	1.9. Replacement in word character(s) استبدال حرف أو أحرف من الكلمة	<OR>
	1.10. Character(s) redundant زيادة حرف أو أكثر	<OD>
	1.11. Character(s) missing نقص حرف أو أكثر	<OM>
	1.12. Other orthographical errors أخطاء إملائية أخرى	<OO>
2. Morphology الصرف 'ssarf	2.1. Word inflection الخطأ في اختيار بنية الكلمة المناسبة	<MI>
	2.2. Verb tense الخطأ في زمن الفعل	<MT>
	2.3. Other morphological errors أخطاء صرفية أخرى	<MO>
3. Syntax النحو 'nnaḥw	3.1. Case الخطأ في الإعراب	<XC>
	3.2. Definiteness الخطأ في التعريف والتنكير	<XF>
	3.3. Gender الخطأ في الجنس (التذكير والتأنيث)	<XG>
	3.4. Number (singular, dual and plural) الخطأ في العدد (الإفراد والتثنية والجمع)	<XN>
	3.5. Word redundant كلمة زائدة	<XT>
	3.6. Word missing كلمة ناقصة	<XM>
	3.7. Other syntactic errors أخطاء نحوية أخرى	<XO>
4. Semantics الدلالة 'ddalāla	4.1. Word selection الخطأ في اختيار الكلمة المناسبة	<SW>
	4.2. *Fasl wa wasl* (confusion in use/non-use of conjunctions) الخطأ في الفصل والوصل (الخطأ في استخدام أدوات العطف)	<SF>
	4.3. Other semantic errors أخطاء دلالية أخرى	<SO>
Punctuation علامات الترقيم 'alāmāt 't-tarqīm	5.1. Punctuation confusion علامة ترقيم خاطئة	<PC>
	5.2. Punctuation redundant علامة ترقيم زائدة	<PT>
	5.3. Punctuation missing علامة ترقيم مفقودة	<PM>
	5.4. Other errors in punctuation أخطاء أخرى في علامات الترقيم	<PO>

Arabic Error Taxonomy

ARABIC LEARNER CORPUS

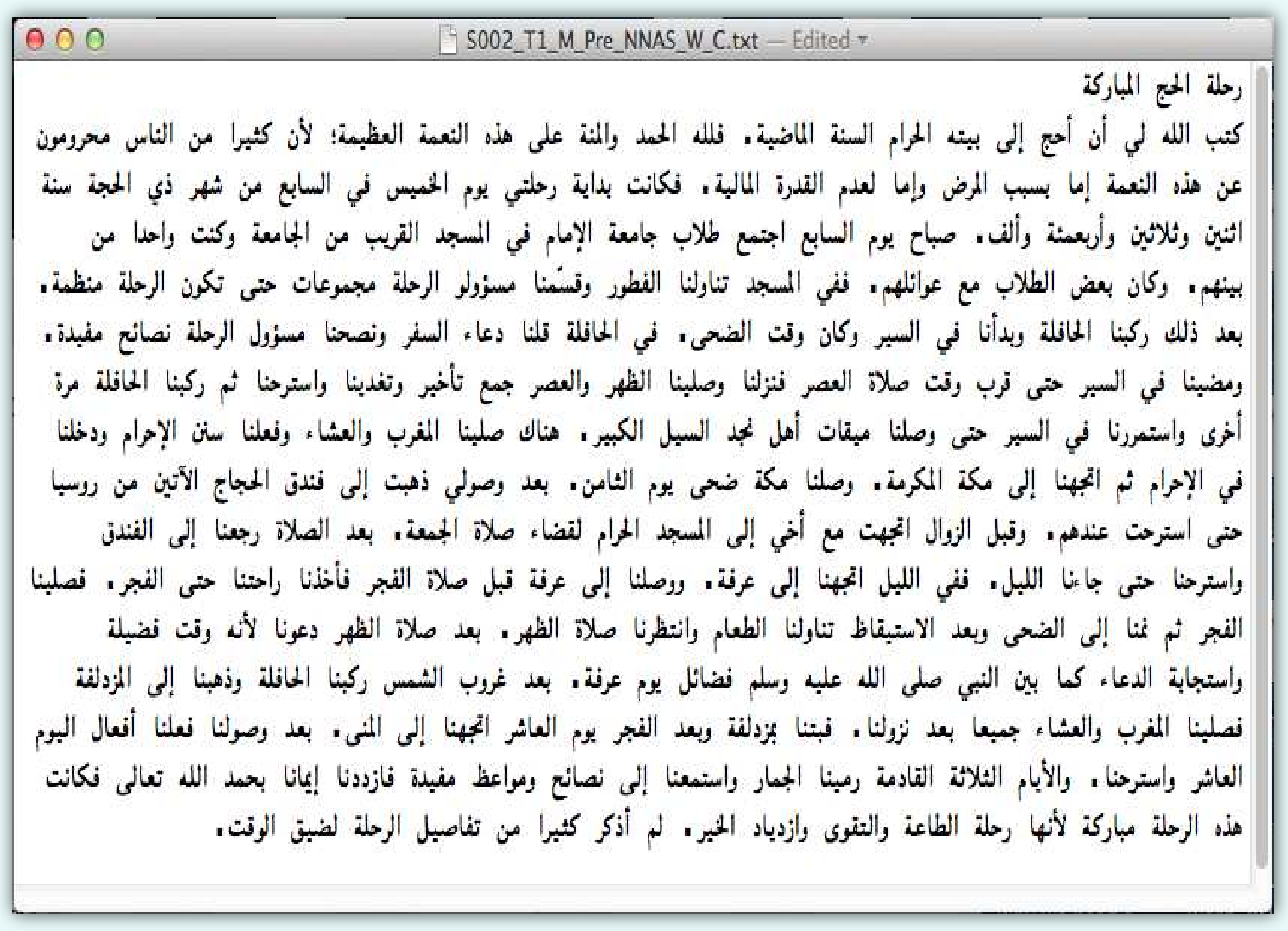
The error annotation will be applied to the Arabic Learner Corpus (ALC), it adds more value to the corpus data.

The corpus includes 282,732 words, produced by 942 learners of Arabic.

ALC ARABIC LEARNER CORPUS
المدونة اللغوية للمتعلمين العربية

CURRENT DATA

The Arabic Learner Corpus currently includes 1585 texts in a raw format.



S002_T1_M_Pre_NNAS_W_C.txt — Edited

رحلة الحج المباركة
كتب الله لي أن أحج إلى بيته الحرام السنة الماضية. فلله الحمد والمنة على هذه النعمة العظيمة؛ لأن كثيرا من الناس محرومون
عن هذه النعمة إما بسبب المرض وإما لعدم القدرة المالية. فكانت بداية رحلتي يوم الخميس في السابع من شهر ذي الحجة سنة
اثنين وثلاثين وأربعمئة وألف. صباح يوم السابع اجتمع طلاب جامعة الإمام في المسجد القريب من الجامعة وكنت واحدا من
بينهم. وكان بعض الطلاب مع عوائلهم. ففي المسجد تناولنا الفطور وقسّمنا مسؤولو الرحلة مجموعات حتى تكون الرحلة منظمة.
بعد ذلك ركبنا الحافلة وبدأنا في السير وكان وقت الضحى. في الحافلة قلنا دعاء السفر ونصحنا مسؤول الرحلة نصائح مفيدة.
ومضينا في السير حتى قرب وقت صلاة العصر فنزلنا وصلينا الظهر والعصر جمع تأخير وتغدينا واسترحنا ثم ركبنا الحافلة مرة
أخرى واستمررنا في السير حتى وصلنا ميقات أهل نجد السيل الكبير. هناك صلينا المغرب والعشاء وفعلنا سنن الإحرام ودخلنا
في الإحرام ثم اتجهنا إلى مكة المكرمة. وصلنا مكة ضحى يوم الثامن. بعد وصولي ذهبت إلى فندق الحجاج الآتين من روسيا
حتى استرحت عندهم. وقبل الزوال اتجهت مع أخي إلى المسجد الحرام لقضاء صلاة الجمعة. بعد الصلاة رجعنا إلى الفندق
واسترحنا حتى جاءنا الليل. ففي الليل اتجهنا إلى عرفة. ووصلنا إلى عرفة قبل صلاة الفجر فأخذنا راحتنا حتى الفجر. فصلينا
الفجر ثم نمنا إلى الضحى وبعد الاستيقاظ تناولنا الطعام وانتظرنا صلاة الظهر. بعد صلاة الظهر دعونا لأنه وقت فضيلة
واستجابة الدعاء كما بين النبي صلى الله عليه وسلم فضائل يوم عرفة. بعد غروب الشمس ركبنا الحافلة وذهبنا إلى المزدلفة
فصلينا المغرب والعشاء جميعا بعد نزولنا. فبتنا بمزدلفة وبعد الفجر يوم العاشر اتجهنا إلى المنى. بعد وصولنا فعلنا أفعال اليوم
العاشر واسترحنا. والأيام الثلاثة القادمة رمينا الجمار واستمعنا إلى نصائح ومواعظ مفيدة فازددنا إيمانا بحمد الله تعالى فكانت
هذه الرحلة مباركة لأنها رحلة الطاعة والتقوى وازدياد الخير. لم أذكر كثيرا من تفاصيل الرحلة لضيق الوقت.

Example of raw text

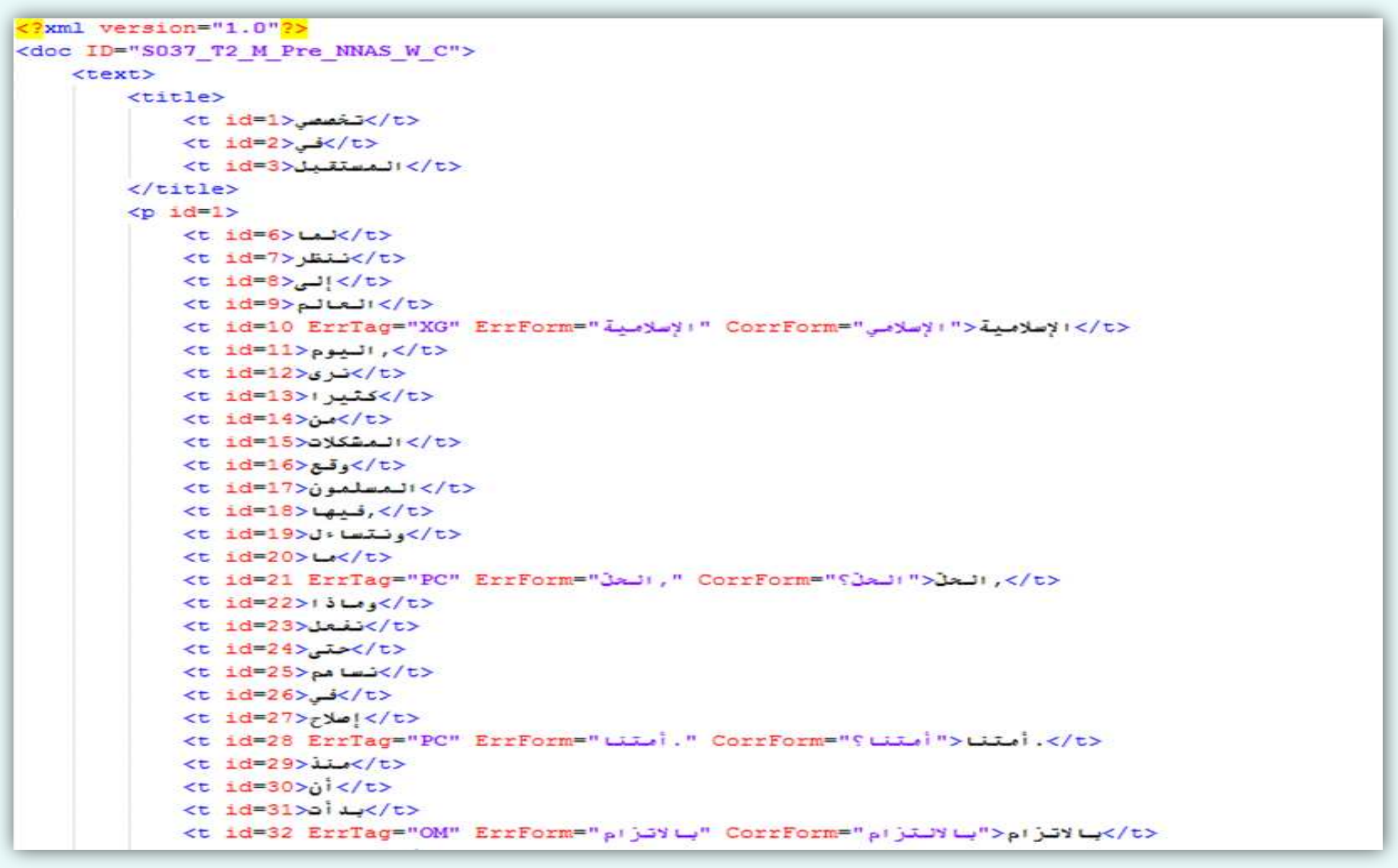
Annotation

TARGET DATA

The ALC will include further information describing the language errors

Error Type | **Error Form** | **Correct Form**

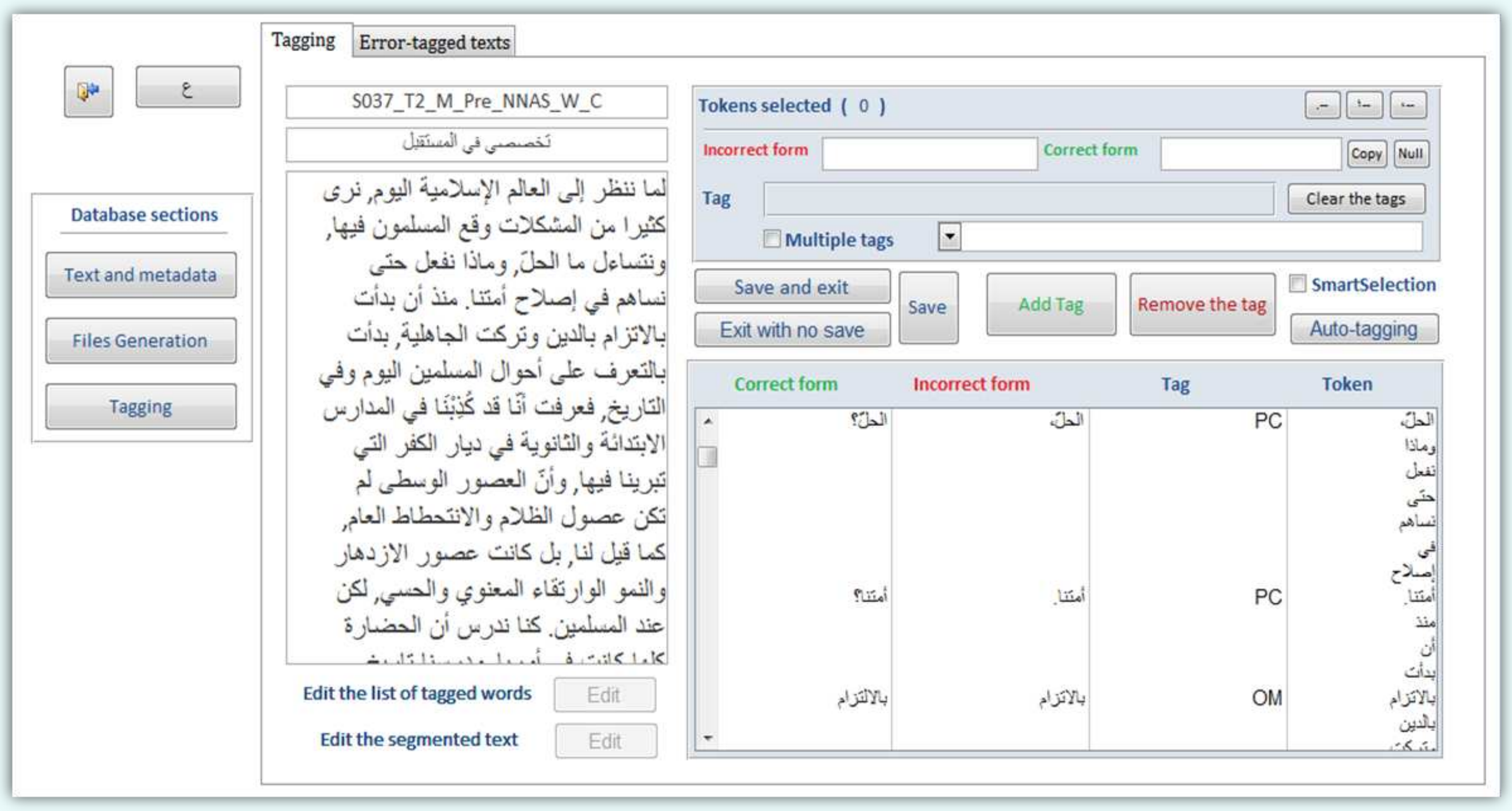
<t id=10 ErrTag="XG" ErrForm="الإسلامية" CorrForm="الإسلامي">الإسلامية</t>



```
<?xml version="1.0"?>
<doc ID="S037_T2_M_Pre_NNAS_W_C">
    <text>
        <title>
            <t id=1>تخصصي</t>
            <t id=2>في</t>
            <t id=3>المستقبل</t>
        </title>
        <p id=1>
            <t id=6>لما</t>
            <t id=7>ننظر</t>
            <t id=8>إلى</t>
            <t id=9>العالم</t>
            <t id=10 ErrTag="XG" ErrForm="الإسلامية" CorrForm="الإسلامي">الإسلامية</t>
            <t id=11>اليوم,</t>
            <t id=12>نرى</t>
            <t id=13>كثيرا</t>
            <t id=14>من</t>
            <t id=15>المشكلات</t>
            <t id=16>وقع</t>
            <t id=17>المسلمون</t>
            <t id=18>فيها,</t>
            <t id=19>ونتساءل</t>
            <t id=20>ما</t>
            <t id=21 ErrTag="PC" ErrForm="الحل," CorrForm="الحل؟">الحل,</t>
            <t id=22>وماذا</t>
            <t id=23>نفعل</t>
            <t id=24>حتى</t>
            <t id=25>نساهم</t>
            <t id=26>في</t>
            <t id=27>إصلاح</t>
            <t id=28 ErrTag="PC" ErrForm="أمتنا." CorrForm="أمتنا؟">أمتنا.</t>
            <t id=29>منذ</t>
            <t id=30>أن</t>
            <t id=31>بدأت</t>
            <t id=32 ErrTag="OM" ErrForm="بالاتزام" CorrForm="بالالتزام">بالاتزام</t>
```

Example of annotated text

THE ANNOTATION TOOL



S037_T2_M_Pre_NNAS_W_C

تخصصي في المستقبل

لما ننظر إلى العالم الإسلامية اليوم, نرى
كثيرا من المشكلات وقع المسلمون فيها,
ونتساءل ما الحلّ, وماذا نفعل حتى
نساهم في إصلاح أمتنا. منذ أن بدأت
بالاتزام بالدين وتركت الجاهلية, بدأت
بالتعرف على أحوال المسلمين اليوم وفي
التاريخ, فعرفت أنّا قد كُنْبَنا في المدارس
الابتدائة والثانوية في ديار الكفر التي
تبرينا فيها, وأنّ العصور الوسطى لم
تكن عصول الظلام والانتحطاط العام,
كما قيل لنا, بل كانت عصور الازدهار
والنمو والارتقاء المعنوي والحسي, لكن
عند المسلمين. كنا ندرس أن الحضارة

Correct form	Incorrect form	Tag	Token
الحلّ؟	الحلّ,	PC	الحلّ,
			وماذا
			نفعل
			حتى
			نساهم
			في
			إصلاح
أمتنا؟	أمتنا.	PC	أمتنا.
			منذ
			أن
			بدأت
بالالتزام	بالاتزام	OM	بالاتزام
			بالدين

Main functions

1. Text Tokenisation

The *Text Tokenisation* function helps in segmenting the text into separate word-based tokens, this enables the annotator to attach tags to those tokens include errors.

2. Smart-Selection

When an error exists more than once, the *Smart-Selection* function enables the annotator to find them all and tag them in a single step with no need to repeat the annotation process with each error.

3. Auto-Tagging

The *Auto-Tagging* feature, which is similar to translation memories, recognises the tokens that have been manually annotated and stores them in a database. So, tokens with the same error can be detected and annotated automatically next time.

For further information about ALC please visit: www.arabiclearnercorpus.com