A Use Case of Automatically Generated Lexicographic Datasets and Their Manual Curation

Dorielle Lonke[1], Raya Abu Ahmad[1], Volodymyr Dzhuranyuk[1], Maayan Or Ner[1], Ilan Kernerman[1]

[1] K Dictionaries, Tel Aviv

E-mail: dorielle@kdictionaries.com, raya@kdictionaries.com, vova@kdictionaries.com, maayan@kdictionaries.com, ilan@kdictionaries.com

Abstract

This paper provides an overview of a multi-layer project combining machine and manual processes in linking multilingual lexicographic resources and leading to the generation of over 200 new language pairs and the update of over 50 existing ones. In the first phase, we create multilingual glossaries by reversing entries from the Password English multilingual dataset of K Dictionaries, reformulating the L1 translations into headwords, aligning them to the original English entries that become their translations, and adding the other language translations of those English entries. The reversal is supplemented by rule-based algorithms to reduce noise; merge, duplicate and separate entries; and check duplicate senses for similar or identical definitions and examples of usage. This is followed by manual detection and amendment of erroneous grammatical categories and faulty meanings, and editing the translation links. The next phase concerns cross-linking each semi-automatically generated multilingual glossary from the first phase with another full lexicographic resource of that L1 from the Global Multilingual Data Series, including its own bilingual versions whenever available. We present the main tasks involved in this project, featuring the automated operations combined with post-editing, the outcomes, our conclusions and further plans.

Keywords: auto-generated data; automatic post-editing; semi-automated processes; manual curation; resource cross-linking

1. Introduction

The creation of up-to-date lexical resources is increasingly facilitated and enhanced by the myriad of methodologies and technologies available for natural language understanding, generation and processing. Traditional requirements and techniques associated with manual compilation of dictionary entries are, on the one hand, empowered by a wide array of automated processes while, on the other hand, supplemented by emerging challenges that stem from these very same processes and others that open new capacities and options for merging different resources with each other.

This paper describes a pipeline of resource convergence and production facets that combine automated processes with manual curation. We begin with crosslingual datasets created by reversing the Password English multilingual dictionary into L1-

English word-to-sense glossaries – by reformulating the L1 translations into headwords, linking them to the original English headwords that become their translations, and adding the other language translations of those English entries – and merge each new L1 resource with another resource of that L1 – some of which are monolingual, bilingual or multilingual – in creating numerous new L1 pairs. The merging process can be outlined as follows:

(1) Use the Password English multilingual dictionary resource (R1).

(2) Reverse R1 – transforming the translations into headwords and the English headwords into their translations – thus producing an L1 to English dataset (R2).

(3) Add the other language translations from R1 onto R2 – using the new English translations as pivots – thus generating an L1 multilingual dataset (R3).

(4) Use another resource of each L1, which may be monolingual, bilingual or multilingual, from the Global Multilingual Data Series of K Dictionaries (R4).

(5) Merge R3 and R4, thus generating a new L1 multilingual resource (R5).

(6) Divide R5 into bilingual sets, thus producing a series of language pairs (R6).

The entire project comprises 19 source languages and 15 target languages (of which 10 are also source languages), so the total number of R6 is 275 language pairs (10x14 + 9x15), involving 25 different languages altogether. Approximately one fifth of these (a little over 50 pairs) were already available in R4, so their corresponding R6 pairs have been updated in the process, whereas all the other language pairs are new. The source and target languages are listed in Table 1.

The pipeline relies on various behind-the-scenes automatic software operations of diverse complexities, with manual editing taking place particularly in curating R2 by means of the specially designated K Index Editorial Tool (KIET), which is used by the editors to review and revise the L1 headword candidates, validate their auto-attributed parts of speech (POS), link to the English equivalents and determine their sense hierarchy, thus detecting and amending erroneous grammatical categories and faulty meanings. The automated processes include rule-based algorithms that reduce noise and merge duplicate entries and senses and check for similar or identical definitions and examples. The rules that serve in this process are devised in accordance with the structure of each target language, taking into consideration semantic variances between English senses and their corresponding translations. Missing POS categories are further provided by matching parallel headwords from a different resource, and more information is introduced from R1, which is later expanded onto matching non-identical but similar POS categories and annotating the glossary to distinguish single lemmas and multiword expressions (MWEs) based on automatic detection. The editor's manual

intervention is minimised by integrating simple rules deduced from repeated evidence of the same error, avoiding redundancies and repetitive amendments of erroneous patterns. Some of the challenges in the post-editing tasks include the detection of such repetitive rules and validating the resulting algorithm, a process which is still mostly done through manual revision and proofing.

Source Languages	**Target Languages**
Arabic	
Chinese Simplified	Chinese Simplified
Czech	Czech
Danish	
Dutch	
	English
	French
	German
Greek	
Hebrew	
Hindi	
Italian	Italian
Japanese	Japanese
Korean	Korean
Norwegian	
Polish	Polish
Portuguese Brazil	Portuguese Brazil
Portuguese Portugal	Portuguese Portugal
Russian	Russian
	Spanish
Swedish	
Thai	
Turkish	
	Ukrainian
	Vietnamese

Table 1: The source and target languages

Section 2 of this paper presents R1, the automatic reversal process and KIET. The actual post-editing of R2 is described in Section 3, along with corresponding automated tasks to produce R3 and combine data components from R4, and the final convergence of R5 is described in Section 4. Section 5 summarises the outcomes of the project and forecasts next steps.

2. The K Index Editorial Tool, its Background and By-products

This section describes the automatic reversal of the English multilingual dictionary (R1), the generation of bilingual (R2) and then multilingual glossaries (R3), and post-editing R2 with the K Index Editorial Tool (KIET).

2.1 The Password English Multilingual Dictionary Resource

The Password English multilingual dictionary (R1) consists of English entries with translation equivalents in nearly fifty languages. The headwords are supplemented with phonetic transcription (IPA) and alternative scripts, POS, grammatical number and sub-categorisation. Each sense of the entry includes a definition and example(s) of usage, and MWEs appear as sub-entries. The translations offer a brief equivalent of each sense and MWE. Figure 1 presents a sample monosemous entry.

jabber ['dʒæbə] *verb*
to talk idly, rapidly and indistinctly:
The students are always jabbering with one another.

AF babble	KO 빨리 지껄이다
AR يَتَكَلَّم بِسُرْعَه	LT plepėti, taukšti
AZ qırıldamaq	LV pļāpāt
BG дърдоря	ML celoteh
BR tagarelar	NL brabbelen
CA balbucejar	NO skravle, plapre løs
CS brebentit	PL paplać
DE schwatzen	PT tagarelar
DK plapre	PRS ور ور کردن
EL φλυαρώ ανόητα και ακατάληπτα	PS ژرژر ږغيدل، بې سنجشه ږغيدل: بې سنجشه وينا، ژرژر خبرې
ES farfullar	PT tagarelar
ET vadistama	RO a bolborosi
FA ور ور کردن	RU тараторить
FI jaaritella	SK trkotať
FR bredouiller	SL klepetati
FY brabbelje	SR brbljati
HE לְבַרְבֵּר	SV padre, babbler, tjattra
HI बकबक करना	TH พูดอย่างรวดเร็วและไม่ชัดเจน; พูดรัว
HR brbljati	TR analcime şekilde konuşmak
HU fecseg	TW 說話急促且含糊不清，閒聊
ID mengoceh	UK плескати язиком; торохтіти
IS masa, blaðra	UR بکواس کرنا
IT ciarlare	VI nói huyên thuyên
JA ペちゃくちゃ言う	ZH 急促而不清楚地说，闲聊

Figure 1: The entry *jabber* in the Password English multilingual resource

2.2 The Reversal Process

The L1-English data are compiled in the process of reversing R1, followed by post-editing R2 as regards the new headwords and POS categories, their links to the English translations and reordering the corresponding senses, including additions or omissions for the auto-generated raw dataset. The L1 entry is created by deriving all identical translations of English entries in R1. The translations are grouped by their POS category and presented to the editor with the original English headword and definition. The editor then determines a new sense order, relying on the English definition as a basic sense indication. This process occurs within the KIET editorial interface. The compilation program follows the algorithm below:

(1) The program runs through all the R1 entries and their corresponding senses. For each sense, it retrieves the translation to L1.

(2) The program creates a new entry in L1 with the same POS as the English headword from which it originated.

(3) If the translation text includes parentheses, commas or semicolons, the text within is divided into separate headwords.

(4) Each L1 headword will include all senses from which it was extracted, including their English definition. This is displayed in the editorial interface, in which the editor can now reorder or remove senses as may be appropriate.

Figure 2 shows an example of the generation of Italian entries from the English entry *thing* in R1. The translation of the second sense as 'a person, especially a person one likes' to Italian is '*persona*, *creatura*'. These translations were thus divided into two separate headwords, *persona* and *creatura*, in R2.

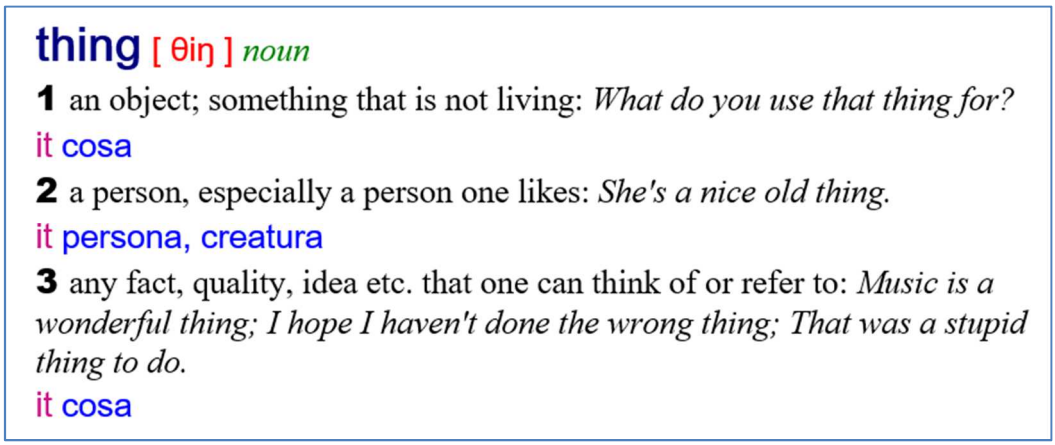

Figure 2: The English entry *thing* with translations to Italian in R1

The English entries *person* and *soul* also contain '*persona*' as a translation of one of their senses, as shown, respectively, in Figures 3 and 4.

As a result, the entry *persona* in the Italian R2 includes all the occurrences of this word as a translation to Italian in R1. All its senses thus comprise these original English meanings, as shown in Figure 5.

person [ˈpəːsn] (*plural* **people** [ˈpiːpl] ˈ**persons**) *noun*

1 a human being: *There's a person outside who wants to speak to you; Some people are never satisfied.*

it persona

2 a person's body: *He never carried money on his person (= with him; in his pockets etc.).*

it sé, persona

Figure 3: The English entry *person* with translations to Italian in R1

soul [səul] *noun*

1 the spirit; the non-physical part of a person, which is often thought to continue in existence after he or she dies: *People often discuss whether animals and plants have souls.*

it anima, spirito

2 a person: *She's a wonderful old soul.*

it anima, persona

3 (of an enterprise etc.) the organizer or leader: *He is the soul of the whole movement.*

it anima

4 soul music: *Her music mixes elements of soul and jazz.*

it soul

Figure 4: The English entry *soul* with translations to Italian in R1

persona *noun*

1. person *noun*

a human being

◊ *There's a person outside who wants to speak to you* ▫ *Some people are never satisfied.*

2. soul *noun*

a person

◊ *She's a wonderful old soul.*

3. thing *noun*

a person, especially a person one likes

◊ *She's a nice old thing.*

Figure 5: the Italian entry *persona* in R2 with the sense division based on the English entries in R1

2.3 K Index Editorial Tool

This post-editing process is done with KIET, which is a bundle consisting of two programs – the admin tool and the editing tool. The admin tool has a graphic user interface (GUI) that enables the project manager to control the backend processes by which data is generated. In these processes, databases on which the editors perform the initial revision are generated from R1, and at a later step, XML files are created from the edited datasets (R2 and R3).

The current version of KIET is based on a revision of the original version developed in 2014 (cf. Egorova, 2015; Kernerman, 2015). The current generation of R2 data was prefaced with a thorough review of the 2014 version, which resulted in several improvement points. The first point of action was adding an admin interface, as the initial KIET version did not include one. The review process raised the need for a GUI on which project managers could control the process of the initial creation of R2 datasets. With the admin tool, project managers can add more languages to the datasets and create new ones by simply entering the required languages into the admin tool, without depending on a software developer to handle the creation. Second, new design features were added to the new (2020) version. It was decided to improve the design and performance in terms of user experience, a point that was previously ranked lower in priority. As more and more languages were added to the R2 project, it became evident that the user experience of the editors was crucial for smooth operation. Through productive cooperation between the software and the content teams, the KIET UX\UI was improved incrementally, with the content team providing input on whether an added feature was intuitive and easy to understand. Third, new features related to the linguistic aspect of the compilation were added in an evolving process that occurred concurrently with the R2 post-editing (described in detail in Section 3) and were added incrementally to the KIET. For example, the POS value list was updated to correspond to ongoing work on the R2 data, and new *GrammaticalNumber* and *Subcategory* fields were added to reflect newfound grammatical information. Further, automatic checks were introduced to reduce duplications (which were also handled in the post-editing stage), as well as a feature alerting the editor about missing information such as POS category. These additions were born from a trial-and-error process pertaining to the revision of the first R2 files in the 2020 project, which contained substantially more duplicates and missing categories than the consecutive versions.

Figures 6 and 7 display screenshots of the original and new KIET main interface, respectively.

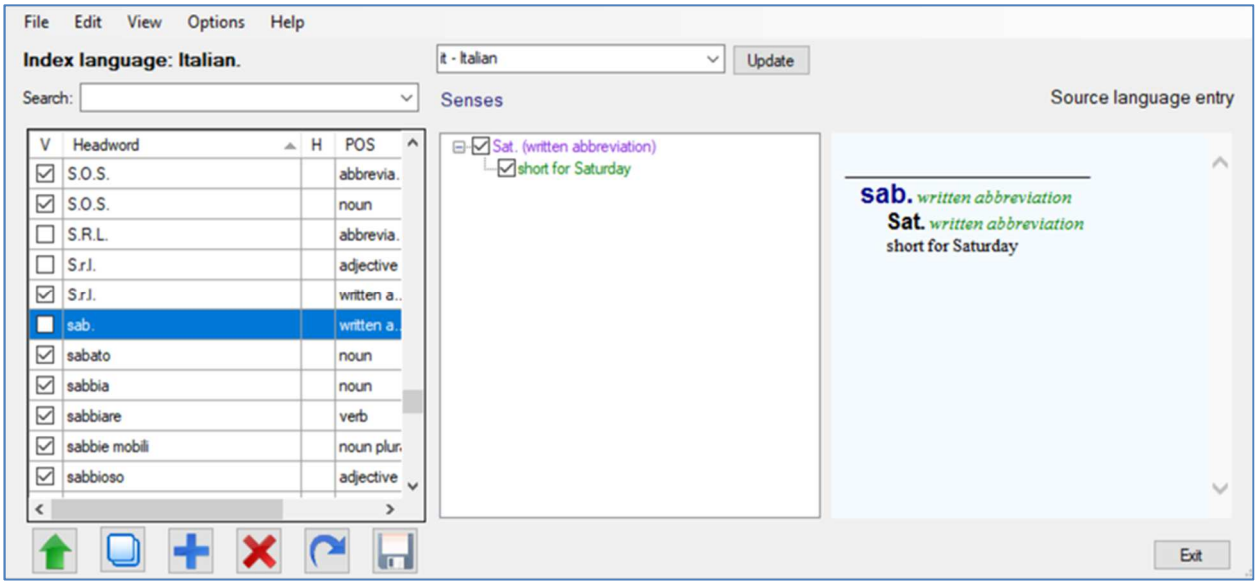

Figure 6: Screenshot of the main interface of the original KIET

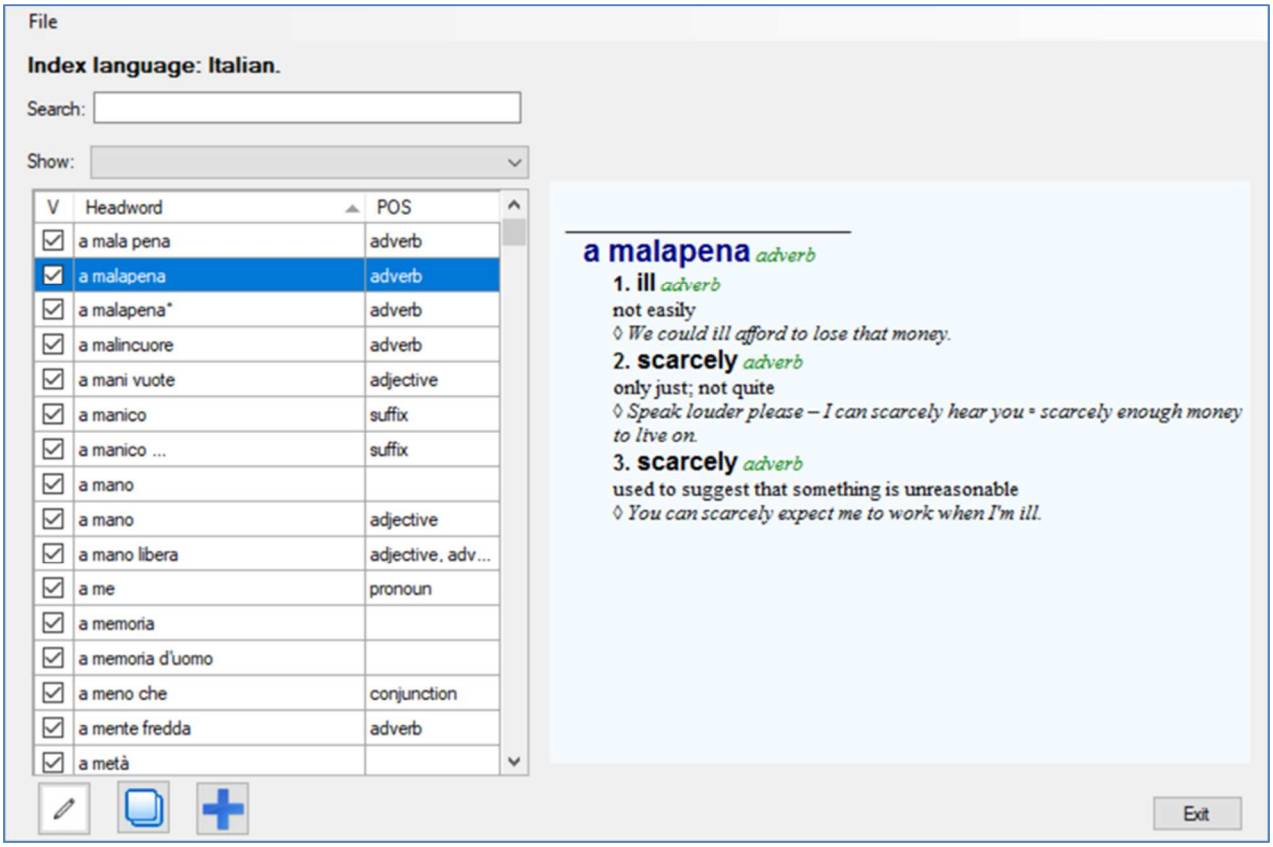

Figure 7: Screenshot of the main interface of the new KIET

2.4 KIET Editorial Interface

The editorial interface of KIET is where the R2 entries are reviewed and revised, enabling the editor to create, remove or duplicate headwords and manage the sense relations and order. Figures 8 and 9 show screenshots of the editorial interface from the initial version (2014) and the current version (2020), respectively.

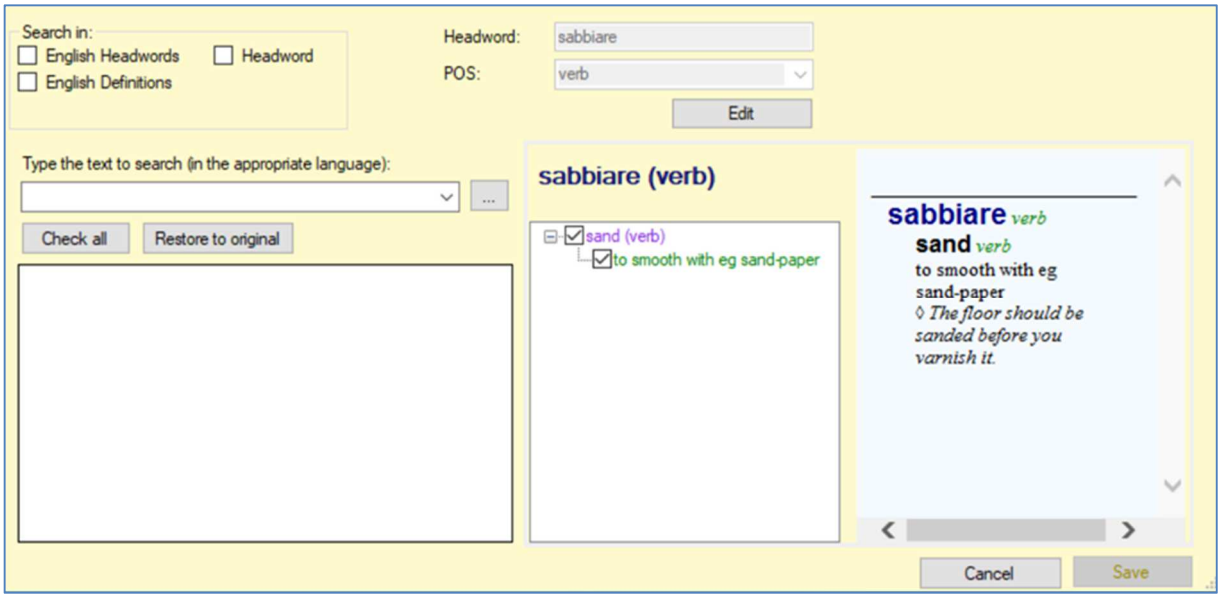

Figure 8: Screenshot of the editorial interface of the original KIET

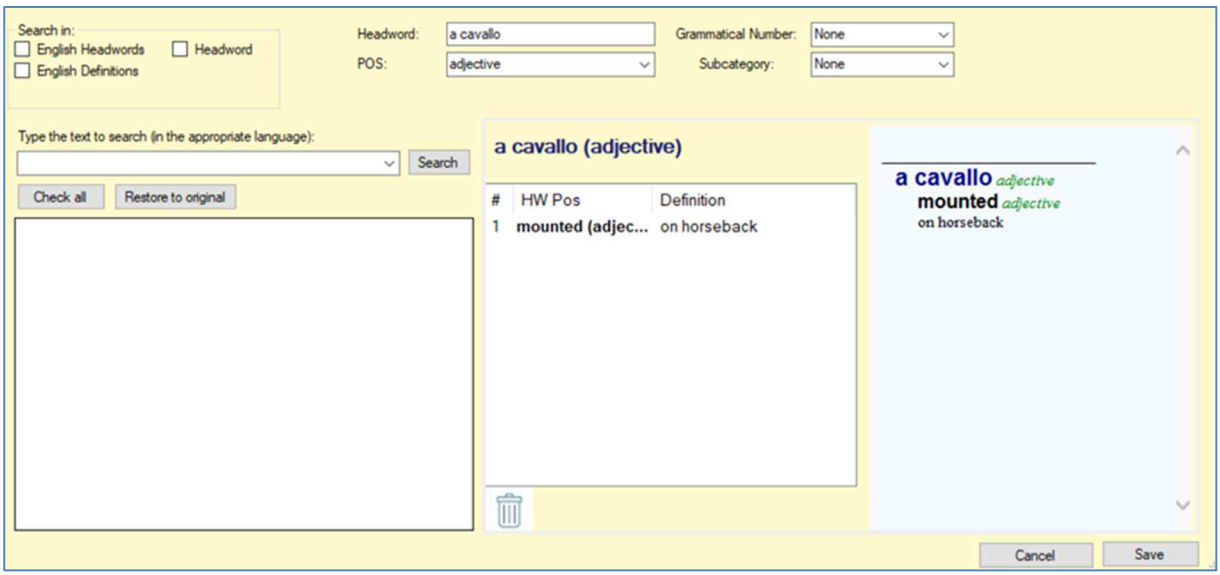

Figure 9: Screenshot of the editorial interface of the new KIET

The main changes in the two versions include a feature that disables the appearance of duplicate entries, that is, entries consisting of the same headword and POS. The first versions of R2 were generated prior to these enhancements and contained many duplications that were handled in the post-editing stage. Post-editing also produced

insights with respect to the implementation of new features, such as the rearrangement of sense order and removal of irrelevant senses. The previous version had design errors that caused the preservation of senses that did not correlate to the senses in the target language, or whose prevalence in that language was much lower than in English. Simultaneously, a newly added menu displays the valid entries as well as those either removed or edited.

The search functionality was enhanced and made more flexible. First, unaccented search was enabled in both (main and editorial) interfaces, removing diacritics and disregarding case. While the first KIET version only allowed searching for a particular entry by the exact headword text, the new version lets the editor search for specific senses of a word by entering either the original English headword or keywords from the definition.

To ensure that a certain structure is maintained, post-editing is only allowed at the entry level. That is, only information pertaining to the spelling of the headword or its grammatical information can be changed. The editor cannot edit the existing sense definitions or add new senses, which is arguably the main shortcoming of R2. The reason for this lies in the R2 structure: while the entry information is generated from a combination of the information pertaining to the original English entry and the equivalent translations in R1, the sense division is based on the English information only. Generally, the R2 senses consist of the English headword and definition, and include the English POS (as further sense indication) and examples of usage. Obliged to remain agnostic to R1, only the sense division and order can be modified in KIET.

3. Post-Editing with Corresponding Automated Tasks

The KIET described in Section 2 was used for the manual editing of the raw (automatically generated) L1-English glossary R2. This post-editing process combined further automated tasks, and the main ones are described in this section. Once the R2 editing was complete, the R1 translations in other languages were added automatically in creating the English pivot-based multilingual glossary R3.

3.1 The Reversed Glossary XML Structure

The multilingual glossary (R3) data is comprised of simple XML documents with a straightforward XML schema. The initial structure consisted of a *DictionaryEntry* element containing two main components. First, the *HeadwordCtn* includes information on the lemma or phrase; initially, it comprised only the headword and POS category, but it was expanded to include more grammatical details such as number or gender, as well as inflected forms. These changes are described below and are part of the post-editing process, which combines automated methods with manual revision and editing.

The second main component of the entry is the *SenseBlock*, including a division into

different meanings (represented by separate *SenseCtn*s), their definitions and examples of usage in English and translation equivalents. The sense division is manifested by the original English information: the headword and POS are wrapped in their own component nested inside the sense, to allow retracing to the original sense in R1. The definition functions as the main sense indicator. Figure 10 presents a sample monosemous entry in the French R2, demonstrating the headword and sense structure.

```
<DictionaryEntry identifier="EN00003471">
  <HeadwordCtn>
    <Headword>charisme</Headword>
    <PartOfSpeech>noun</PartOfSpeech>
  </HeadwordCtn>
  <SenseBlock>
    <SenseCtn id="SE00006070">
      <EnCtn>
        <EnHeadword>charisma</EnHeadword>
        <EnPOS>noun</EnPOS>
        <DefinitionCtn>
          <Definition>a strong personal quality that makes someone
          attract, influence, and inspire other people</Definition>
        </DefinitionCtn>
        <ExampleGrp>
          <Example>He lacked the charisma required to
          become an effective political leader.</Example>
        </ExampleGrp>
      </EnCtn>
    </SenseCtn>
  </SenseBlock>
</DictionaryEntry>
```

Figure 10: XML data of the French monosemous entry *charisme* in R2

3.2 Headwords and Part of Speech Categories

Following the initial automated generation of the R2 sets, it was necessary to introduce editorial amendments reflecting a refinement of the headword forms and the grammatical categories to fit the newfound source languages. The post-editing phase started with revising the headword text and adjusting the POS categories. These modifications were performed manually by the editor of each language and were facilitated by automated processes, including revising the headword text to reflect a more common variant in that language; fixing typos; stripping characters such as slashes, commas, parentheses or brackets; and handling gender inflection. Such cases were either eliminated, inserted into a corresponding tag or divided into independent entries. The primary aim of this initial revision was to verify that all the headword text was cleaned and normalised in order to become fit for automated processing and machine readability.

Alongside the headword revision, the POS category was modified as well. When given

the opportunity to redesign R2 from scratch, the leading heuristic was to simplify the dataset as much as possible, placing the relevant information in designated tags and adhering to a closed list of POS values. As part of the post-editing process, entries missing a POS element were singled out and fixed; existing categories were normalised and stripped from additional information to adhere to a predefined schema of particular POS values; and any additional information that was relevant to the grammar of the word was retained and transferred to corresponding elements, namely the *GrammaticalNumber* and *Subcategory* tags, to reduce noise and facilitate searching the data for relevant information. The POS irregularities can be attributed to two main causes:

(a) The original output did not include a POS category, or the existing category generated by KIET was removed by the editor in the initial editing phase and was not replaced with another value accidentally.

In these cases, an automated process matched the headword text with a corresponding entry in the Global Multilingual Data Series (R4) and inserted the corresponding POS category into the R2 dataset. Since the POS category does not pertain to a particular meaning, it was not necessary to perform any sense alignment prior to the matching. If there were multiple entries with different categories in R4, the information was transferred to an editor to determine the correct category.

(b) The original POS category, which was generated from the English POS category in R1, included additional information, such as grammatical number or subcategory.

In these cases, an automated process located all instances of a POS tag including additional information and separated the POS category from the grammatical information, placing the new information in a corresponding tag.

Figure 11 is a demonstration of an R2 French entry containing the newfound *GrammaticalNumber* tag whose information on plurality is evident from the original English part of speech (*EnPOS*).

As the automated process for generating R2 included attributing the POS of the original English entry in R1 to the new L1 headword in R2, the editors also received a list of headwords whose POS had to be determined or validated. In some cases, no equivalent was available in any parallel resource, so the editors supplemented the information based on their own linguistic knowledge. In other cases, multiple equivalents were found in R4 and were all given to the editor, thus facilitating the decision. In addition, a list of uncertain POS categories was curated, consisting of headwords with POS values that did not belong to a predefined closed list of values – including narrower categories such as ‘proper noun’ instead of ‘noun’ and unconventional or abbreviated text such as ‘adj’, standing for ‘adjective’ – and the editor was asked to select an appropriate POS category from a list of values. As a final

step, the editors were asked to review all headwords tagged as 'plural' or 'abbreviation' (for each element respectively) and to verify whether this tagging was correct. This demonstrates how automatic retrieval of information, albeit not precise or exact, can help the manual work and speed the post-editing process.

```
<DictionaryEntry identifier="EN00000338">
  <HeadwordCtn>
    <Headword>accents</Headword>
    <PartOfSpeech>noun</PartOfSpeech>
    <GrammaticalNumber>plural</GrammaticalNumber>
  </HeadwordCtn>
  <SenseBlock>
    <SenseCtn id="SE00000519" num="">
      <EnCtn>
        <EnHeadword>overtones</EnHeadword>
        <EnPOS>noun plural</EnPOS>
        <DefinitionCtn>
          <Definition>suggestions; hints</Definition>
        </DefinitionCtn>
        <ExampleGrp>
          <Example>There were overtones of discontent in his speech.</Example>
        </ExampleGrp>
      </EnCtn>
    </SenseCtn>
    <SenseCtn id="SE00000520" num="">
      <EnCtn>
        <EnHeadword>strain</EnHeadword>
        <EnPOS>noun</EnPOS>
        <DefinitionCtn>
          <Definition>(often in plural) the sound of a tune</Definition>
        </DefinitionCtn>
        <ExampleGrp>
          <Example>I heard the strains of a hymn coming from the church.</Example>
        </ExampleGrp>
      </EnCtn>
    </SenseCtn>
  </SenseBlock>
</DictionaryEntry>
```

Figure 11: XML data of the French polysemous entry *accents* in R2

3.3 Eliminating Duplicate Entries and Senses

As mentioned in Section 3.2, as part of the automated POS attribution process, missing categories were supplemented from the Global Series, and variants of existing categories were cleaned and normalised. This process in turn resulted in another data issue, which was also handled and solved automatically as part of the post-editing pipeline. Amending the headword text and POS categories resulted in many cases in which the same headword text and POS appeared for two separate *DictionaryEntry* elements in the data, that is, two separate entries that originally included the same headword text, but different POS categories were now duplicate cases of the same entry. However, just removing one of the entries would not suffice, since the senses were in most cases different for each entry. The purpose of the automated task was to eliminate duplicate entries in the data while retaining all information from the sense level. This was divided into two steps. The first step, handling the duplicate entries, was designed according

to the following algorithm and combined an automated process with extra human validation:

(1) For each entry, check if there is another entry that shares the same headword text and POS category.

(2) If one entry includes additional grammatical information (such as number or subcategory), the revision is delegated to the editor to manually verify that the entrees are indeed separate entries and make the proper modifications to distinguish them.

(3) If there is no additional information, take all senses from the second occurring entry and append them to the *SenseBlock* of the first occurring entry, then remove the second entry from the dataset.

This process is general enough to catch many cases, but at the same time remove the risk of accidentally concatenating two entries that are not in fact identical; involving the editors in the automated post-editing process allowed the flexibility and speed of an entirely automated pipeline while still retaining the benefits of humanly curated data that is checked and validated after every step. The second step, which included the revision of duplicate senses following the grouping together of senses from two separate entries, was done separately, so as to break down the deletion process into smaller, manageable steps that could be verified upon execution, thus reducing the error margin to a minimum.

To preface the sense elimination step, it is important to reiterate the compilation process of the R2 dataset: as presented in Section 2, this data is constructed by retrieving translations from English entries in R1. Translations from different entries are grouped together by POS categories, and the editor is requested to rank the sense order by importance or prevalence, relying on the English definition as an indication for the sense (since no additional information is given for the entry in L1). Then, the L1 entry is created for that R2, including all senses belonging to the corresponding English entries sharing the POS category. Upon revision of the resulting R2, and after amending headword text and POS categories as previously described, we generate separate entries encompassing the same lemma, for which multiple and different senses belong. When concatenating together the amended entries, it is now necessary to check that no duplications occur within the collection of the different senses. This phase is slightly more complicated than the previous one of eliminating duplicate entries, as it must take into account the meaning variations and carefully consider whether two senses reflect the same meaning. This process, like the previous one, combined an automated process with manual post-editing. Relying on the four types of information that currently exist within a *SenseCtn* for an individual sense, which is the English headword, the English POS, the English definition text and examples of usage, an algorithm was constructed according to the following guidelines:

(1) Comparing each following sense to the first one as an anchor, an automated process checked whether the sense pair included the identical English headword and POS information. If so, and there was no additional information, the second sense was removed from the dataset.

(2) If additional information existed, the process then compared the definition text: if the definition text was identical, then the process merged the two senses by deleting the second sense and taking any examples it contained and appending them to the *ExampleCtn* of the first sense; if no examples existed, no action was required.

(3) If the definition text was not identical, the senses were transferred to the editor for manual editing.

The editor then had to determine whether the two definitions encompassed the same meaning, or if they were distant enough to count as separate senses. Figure 12 presents a sample of a merged entry in which the original English headword and POS information are identical, but the definitions reflect separate meanings:

Figure 12: Italian polysemous entry *aberto* in R2

Here, the manual check was able to determine that these are all separate meanings of the English word *open*, thus leaving the initial sense division as is and retaining all relevant example phrases and sentences. Some senses containing three or more examples are the result of an automated process comparing two senses which had the same definition text and grouping together their separate examples to one sense, demonstrating uniform usage for a singular meaning.

The numbers of problematic entries varied between languages. Some, such as French or Italian, initially included a small number of suspicious duplicates, and others, such as Chinese, had much higher numbers of duplicate or erroneous headwords to be examined and modified, ranging between 100 and 5,200 entries per language. The automated process managed to reduce manual work by more than half, resulting in a significantly lower number of cases for editorial review and revision. In the case of Chinese, the initial process of eliminating duplicates covered as many as 5,000 cases, leaving approximately 200 entries only for manual post-editing and curation. This process could be further automated by relying on additional tools and resources that enable the definitions to be compared, checked for their closeness, or rated for their similarity by a particular metric (Kaltenböck and Kernerman, 2017). The current process relied on straightforward string comparison and applied human judgement to determine sense division, due to time constraints and the uncertainty of such similarity tests. However, it would be interesting to incorporate such tests in more elaborate automatic post-editing pipelines.

3.4 Further Revision and Evaluation

Nearly every step explained above required the editor to verify and validate the automatic outcome, as well as to point out additional problems with the data that might need further (automatic) tackling. The design of the pipeline itself allowed for the minimal amount of material to be manually reviewed, by taking care of tasks that can be handled entirely automatically first and delivering anomalous tasks to editors second. A list of unconventional duplicate entries and senses was also reviewed manually, bearing in mind to amend any automatically integrated information that was incorrect, while keeping all relevant information by concatenating it from the duplicate entries, thus creating one full final entry. Similarly, a list of headwords with slashes, brackets and other abnormal characters was reviewed, stating the correct text to be amended and whether another entry was to be added. For example, the original Swedish headword '[allt]sedan' was separated to two new headwords 'allt sedan' and 'sedan'.

The process of identifying and separating variants from headwords containing slashes revealed a sub-category of cases in which the text after the slash was not an individual word but rather a suffix for the feminine form of the headword for languages with gender inflection. These were identified by a dash preceding the suffix, indicating the need to replace the masculine suffix of the original word. For example, the French R2

included the headwords with text 'acteur/-trice', 'alarmant/-ante' and 'champion/-onne', which surfaced when searching for headwords with peculiar characters such as slashes. These cases were handled almost entirely automatically, by devising a rule for generating the full feminine form based on the root and the masculine form, verifying the results automatically, and then manually checking them to obtain even more security. The process is described below:

(1) Generating the feminine form was carried out according to the following rule, based on French grammar:

   a. If the suffix begins with a vowel V, the root form is taken as all characters up to the same vowel in the ultimate position of the word, and the suffix, i.e., the text after /-, is then appended to the root, e.g., champion/-onne → champi + onne → championne; alarmant/-ante → alarm + ante → alarmante.

   b. If the suffix begins with a consonant C, the root form is taken as all characters up to the same consonant in the ultimate position of the word, and the suffix, i.e., the text after /-, is then appended to the root, e.g., acteur/-trice → ac + trice → actrice. It should be noted that the V/C distinction is based on the existing orthography and not on French morphological rules.

The resulting forms ('acteur' and 'actrice', 'champion' and 'championne', 'alarmant' and 'alarmante') were then looked up in existing French resources or morphological lists and marked as safe if said forms existed in any such resource. If not, they underwent an automatic translation process, relying on machine translation tools to translate both forms back to English and check whether they match. A match indicates that the automatic generation succeeded in high likelihood. For example, 'champion' and 'championne' both translate to the English 'champion' and were thus marked as a success. The pair 'acteur' and 'actrice', in turn, were located in R4 and marked as a success too.

(2) Following suit, the editor reviewed the automatically generated forms and their success mark and amended the results if necessary.

The benefits of having an existing suggestion for a form as well as a metric to evaluate the success for the automatic generation is twofold: it saves time by eliminating the need to manually enter a value, and it greatly reduces the chances for typos or spelling mistakes. However, relying solely on written characters and their placement relative to each other to devise an automatic rule carries its own risks. The inclusion of manual editorial work in this case also proved to be of high importance: the editor was able to amend errors caused by the algorithm, as well as identify cases that were not marked as a success and identify whether or not they encompass a gender inflection, or a typo.

(3) The reviewed masculine and feminine forms were then incorporated in the data by keeping the masculine form in the headword and introducing an *InflectionCtn* component in which the feminine form was inserted. Grammatical information pertaining to gender was also added to *GrammaticalGender* tags. Figure 13 presents an example of the instantiation of this modelling for the entry 'acteur'.

```
<DictionaryEntry identifier="EN00000447">
  <HeadwordCtn>
    <Headword>acteur</Headword>
    <PartOfSpeech>noun</PartOfSpeech>
      <GrammaticalGender>masculine</GrammaticalGender>
    <InflectionCtn>
      <Inflection>actrice</Inflection>
      <GrammaticalGender>feminine</GrammaticalGender>
    </InflectionCtn>
  </HeadwordCtn>
  <SenseBlock>
    <SenseCtn id="SE00000726" num="">
      <EnCtn>
        <EnHeadword>actor</EnHeadword>
        <EnPOS>noun</EnPOS>
        <DefinitionCtn>
          <Definition>a performer in a play.</Definition>
        </DefinitionCtn>
        <ExampleGrp>
          <Example>a film/movie actor.</Example>
        </ExampleGrp>
      </EnCtn>
    </SenseCtn>
  </SenseBlock>
</DictionaryEntry>
```

Figure 13: XML data of the French polysemous entry *acteur* in R2

Naturally, this process of further revising the headword texts for any R2 dataset may result in newfound duplicate entries. The previously described process of identifying duplicate entries, concatenating them and eliminating their duplicate senses was performed incrementally after each revision of the headword text and could be performed again and again until the revision was finalised.

To find possible misspellings among the resulting headwords, a spell-checking pipeline was defined and implemented for each language. First, all textual data was checked automatically using existing or custom spell-checkers, and then the results were reviewed by the editor, who corrected true misspellings. At the end of the process, the amended text was merged back to the dataset. Obviously, spell-checking in a multilingual environment is a rather challenging task. For some languages, existing tools or simple pipelines yield satisfying results, with a small number of false positives and high recall, that is, most of the misspellings were detected by the system. However, for other languages, mostly morphologically-rich or low-resource ones, the task requires

more tuning and specific implementation. A high number of false positives is counterproductive, as it generates additional editorial work, which is expensive and impractical. Possible solutions may involve morphological analysis as a pre-processing step, mining additional "known words" vocabularies from corpora and utilisation of other available resources.

4. The Full Resource Conversion and the Final Outcomes

In the second phase of the project, the R3 resources were merged with the Global Multilingual Data Series (R4), consisting of a collection of extensive lexicographic cores for different languages. Each language core includes a wide lexical base featuring rich semantic and grammatical information arranged in well-structured datasets, within the framework of a single comprehensive macrostructure and all adhering to the same entry microstructure, with most of these language cores having bilingual and multilingual versions in varied numbers.

The main entry components of the R4 sets include phonetic transcription (in IPA) and alternative scripts, POS, irregular forms, grammatical subcategorisation, gender and number, as well as sense division based on frequency with definitions, examples of usage, related MWEs and other attributes such as synonyms, antonyms and subject domain.

To converge R3 with R4, it was necessary to develop a meticulous algorithm, first to match the headwords in each resource and then to link senses correctly for polysemous entries in either or both resources.

MWEs and nested entries were also taken into consideration so as to expand the database of entries for which the merging is performed and raise the chances of a match. The matching algorithm then searched for the headwords within the expanded collection and matched them with corresponding entries from R3. The algorithm was constructed as follows:

(1) A dataset was created for the R4 entries, including POS categories, synonyms and inflections.

(2) The matching program ran through this dataset, and for each headword or inflection, and their corresponding POS, it checked whether the pair exists in R3 as a headword and POS pair, disregarding the POS component for MWEs.

(3) If the headword and POS pair was identified within R3, it was added to a set of all matching R3 entries.

The result is a set of matching pairs – R4 entries and their corresponding entries from R3 that were found as headwords or MWEs or as an inflection of an entry in R4. The following stage, which consisted of a sense alignment of sorts, was comprised of two steps. The first relied on translations to perform the initial sense linking. The second

relied on synonyms, if existing, to further expand the possibility of matching the R4 sense with a corresponding R3 sense. The sense matching algorithm is as follows:

(1) The algorithm loops through all of the senses of the matching entry, focusing on the available translations of the senses.

(2) The algorithm then loops through all the senses of the R3 entry; for each sense, if any of the translations of the R4 sense matches any of its translations, the sense is registered as a matching sense pair, and the number of matching translations is counted.

(3) If any of the R4 senses have synonyms, they are searched for within the R3 resource. If an R3 entry identical to the synonym is located, then the program runs through all its senses, comparing it to the sense of the R4 entry in which the synonym was originally found; the same process of translation comparison is performed for the matching synonym entry.

After reviewing all the R3 senses that were singled out as possible matches, the most fitting one is selected. The parameter in this case is the highest number of matching translations. The guiding principle in the process of sense linking was that each R3 sense can match no more than one R4 sense for the same entry. The percentage threshold for the matching varied for each language, mainly due to a discrepancy in the number of target languages for each source language in R4. At first, each language output included only the sense that passed a certain matching percentage threshold. Later on, it was decided to also include entries that constitute exact matches at the headword and POS level (i.e., not found as inflections), even if none of the senses passed the initial threshold.

Prior to the matching phase, there were a few issues that were taken into consideration. Similarly to the initial creation of R2, text containing slashes, commas or semicolons that separated two or more values was handled to find matches for each value separately. Further, definite articles and prepositions were cleaned from the text. Any additional information that usually accompanies the main headword and found inside parentheses was removed. Diacritics, stress and case or capital letters (uppercase vs lowercase) were disregarded. Conversion tables provided for each L1 facilitated the normalisation and mapping process.

5. Conclusions and Future Work

The endeavor of converging and transforming existing lexicographic datasets into a brand-new resource requires substantial effort. The initial manual editing is tedious yet necessary; this process encompasses the initial shift from English as the main source language to a new language that is now at the front. What was previously a target language, embodying the lexicographic resolutions of translating that which cannot always be directly translated, is now at the forefront. The following stage of post-editing

enabled a combination of automated and manual processes to facilitate much of the manual labor. This also embodied a learning curve wherein insights were extracted from the work on each dataset and improved for the reiteration of the next step. In that sense, the incremental workflow, whereby each step enabled evaluation and later revision of previous steps, allowed for a flexible pipeline and immediate repairing of errors.

Some improvements of KIET could be derived from the automated post-editing pipeline. The admin tool could be enhanced with more automated features and functionalities, thus eliminating the need to perform these tasks in the next post-editing phase. For example, when post-editing revealed many duplicate entries that existed in datasets generated by previous versions of KIET, a feature to alert about possible duplications was added to both the editorial and administrative interfaces. Other processes such as the normalisation of POS and grammatical information could be added as a preliminary phase inside KIET, voiding the need for an extra step in the following automated pipeline. Generally speaking, the post-editing pipeline could be reduced to manual editing accompanied by particular automation as required, and anything that could be described as a general rule could be added to the KIET backend.

The process of merging datasets also relied heavily on automated checks, which could be further improved by expanding the arsenal of tools that are used for such revision. The R4 resource was merged with the newly generated R3 resources in a matching process consisting of a direct string-based comparison with minimal clean-up. Indeed, MWEs were included as well, and a closed list of inflections and synonyms were added to expand the pool of words in which the search for matches was conducted. The downside to this is that variants, either spelling variants or other morphologically inflected forms, could be missed even though the POS is identical, and the meaning is similar, which could result in fewer matches and a lower recall. However, many senses that may have been overlooked due to a small discrepancy in the headword form, or other small variations, might be detected with further adjustments. For example, this process could be improved by utilizing word embeddings that can provide an approximation of similarity between variants or differently spelled words. Similarly, the merge pipeline could be enhanced by employing sentence encoders to measure the similarity of two differently phrased definitions at the sense level.

The main benefit of the creation of new datasets and merging them with existing ones is the prospect of creating a larger, more extensive dataset, combining the strengths of different resources. The Global Series could be enhanced as well, not only by using the newly created R3 resource, but also taking external multilingual resources and applying the same pipeline, thus adding more components and enriching the data. In terms of evaluation, future work may include an exact documentation of numbers of matching instances for duplicate entries and mismatches that require manual review. A case could be made for the calculation of precision values for each language, as this information could be included in further identification of language-specific issues, but since this

project did not implement learning algorithms and its focus was the preparation of data for production and not the training of a model, we did not explicitly document these numbers, and the current information provided in this paper is based on a retrospective examination of logs.

6. References

Egorova, K. (2015). Editing an automatically-generated index with K Index Editing Tool. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom.* Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 268–280. https://elex.link/elex2015/proceedings/paper-17

Kaltenböck, M. & Kernerman, I. (2017). Introducing LDL4HELTA: Linked data lexicography for high-end language technology application. *Kernerman Dictionary News* (25), 2–3. https://kdictionaries.com/kdn/kdn25_2017.pdf.

Kernerman, I. (2015). A multilingual trilogy: Developing three multi-language lexicographic datasets. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom.* Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 372–383. https://elex.link/elex2015/proceedings/paper-24