Multilingual Editing of Linguistic Problems

Ivan Derzhanski
Department of Mathematical Linguistics
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
iad58g@gmail.com

Abstract

Multilinguality has been an essential feature of the International Linguistic Olympiad since its conception. Although deemed most desirable, the production of a problem set in several parallel versions and the verification of their equivalence is a time-consuming and error-prone task. This paper tells about the efforts to develop tools and methods which increase its efficiency and reliability.

1 Introduction

In September 2003 the 1st International Linguistics Olympiad (IOL, *née* International Olympiad in Theoretical, Mathematical and Applied Linguistics), an annual contest for secondary-school students in solving self-sufficient linguistic problems (Derzhanski, Payne 2009), took place in Bulgaria. Six countries were represented by a total of 33 participants. At the 10th instalment in 2012 the countries were 26, the contestants 131, and both numbers keep growing.

Since its launching, multilinguality has been a crucial feature of IOL. A linguistic problem depends more on the language in which it is formulated than a problem in, e.g., mathematics: not every problem can work in all languages, and even when it can, producing versions which give equal chances to all contestants is not always straightforward. For this reason at IOL, unlike many other international fora, there is no question of limiting the working languages to one or just a few. Accordingly their number has grown from five at IOL1 to fifteen at IOL10.[1] For the same reason the versions of the problem set in all working languages can't be created immediately before the contest, as is done at some of the other international science olympiads; they need to be prepared and verified well in advance.

The production of the multilingual package is a time-consuming and error-prone task, and it calls for the development of tools and methods to increase its efficiency and reliability.

2 The Past: IOL1

A linguistic problem is composed of language material and surrounding text; the language material in turn consists of data in unfamiliar languages and in Solverese[2] (usually translations of the unfamiliar language data). In a multilingual edition of the problem set it is imperative that the Solverese parts be equivalent and everything else be identical.

Figure 1 presents half a page from the Dutch and the English versions of the IOL1 problem set. It is easy to see that the formatting, the formulae and the Egyptian Arabic expressions had to be exactly the same.

In order to minimise the effort needed to edit the problems in all working languages and the chance that a technical mishap might create a divergence where none should occur, an *ad hoc* method was invented. The problem set was written in LaTeX, [3] with a master source file for each

[1] In fact at IOL1 and some subsequent early IOLs the versions that were made outnumbered the actual working languages by one, because an English version was made, although not used at the contest, for general reference and for advertising. At some of the recent IOLs, too, there have been more versions than working languages, as the versions in British and American English have been separate, though only differing in the format of the dates and in the spelling of a few words.

[2] On this term see (Bozhanov, Derzhanski 2013, fn. 2).

[3] This choice was made because TeX is not a mere typesetting system but a full-fledged programming environment, which enables some of the text to be computed rather than typed, greatly reducing the danger of typographic errors. The most powerful inspiration was (Knuth 1986, p. 218); see also (Derzhanski 2009, Appendix A).

Proceedings of the Fourth Workshop on Teaching Natural Language Processing, pages 27–34,
Sofia, Bulgaria, August 4-9 2013.

Opgave 2 (25 punten)

Hieronder staan rekenkundige vergelijkingen in het Egyptische dialect van het Arabisch[1]. Alle onderdelen voor en na het "=" teken zijn breuken waarin de tellers en noemers niet hoger zijn dan 10. (Alleen het rechterdeel van de laatste som is hierop een uitzondering.) Er is ook geen noemer, die gelijk is aan 1:

$$\begin{aligned} tumn + tumnēn &= talatt\ itmān && (1)\\ sabaʕt\ itlāt + suds &= ʕašart\ irbāʕ && (2)\\ tusʕēn + tusʕ &= sudsēn && (3)\\ xamast\ ixmās + subʕ &= tamant\ isbāʕ && (4)\\ subʕēn + xumsēn &= \frac{24}{35} && (5) \end{aligned}$$

Opdracht 1. Noteer deze vergelijkingen in cijfers.

Opdracht 2. In de vergelijking $rubʕ + ʕašart\ itsāʕ = sabaʕt\ isdās$ ontbreekt één teken. Welk teken is dat?

Noot: De letters x en $š$ worden ongeveer als de Nederlandse *ch* en *sj* uitgesproken; $ʕ$ is een specifieke Arabische medeklinker. Het streepje boven een klinker geeft lengte aan.

(Ivan Derzhanski)

Problem 2 (25 marks)

Below you see arithmetic equalities written in Egyptian Arabic[1]. All summands, as well as all sums except the last one, are represented as fractions in which neither the numerators nor the denominators are greater than 10, nor is any denominator equal to 1:

$$\begin{aligned} tumn + tumnēn &= talatt\ itmān && (1)\\ sabaʕt\ itlāt + suds &= ʕašart\ irbāʕ && (2)\\ tusʕēn + tusʕ &= sudsēn && (3)\\ xamast\ ixmās + subʕ &= tamant\ isbāʕ && (4)\\ subʕēn + xumsēn &= \frac{24}{35} && (5) \end{aligned}$$

Assignment 1. Write these equalities in figures.

Assignment 2. The equality $rubʕ + ʕašart\ itsāʕ = sabaʕt\ isdās$ is missing a sign. Which one?

Note: The letter $š$ is pronounced as English *sh*, x as the *ch* in *loch*; $ʕ$ is a specific Arabic consonant. A bar above a vowel indicates length. (Ivan Derzhanski)

Figure 1. Half a page from the Dutch and the English versions of the IOL1 problem set.

```
\newpage
\problem {25}
%
Hieronder staan rekenkundige
vergelijkingen in het Egyptische dialect van het Arabisch%
\footnote{[…]}.
%
Alle onderdelen voor en na het ``='' teken zijn breuken waarin de tellers
en noemers niet hoger zijn dan~$10$. (Alleen het rechterdeel van de laatste
som is hierop een uitzondering.) Er is ook geen noemer, die gelijk is aan
$1$:
%
\fracdata
%
\assignment Noteer deze vergelijkingen in cijfers.
\assignment In de vergelijking \hfill \fractest \hfill ontbreekt \'e\'en
teken.\\ Welk teken is dat?
\comment
De letters \wipa x en \wipa{\sh} worden ongeveer als de Nederlandse
\word{ch} en \word{sj} uitgesproken;
\wipa C is een specifieke Arabische medeklinker.
Het streepje boven een klinker geeft lengte aan.
\by{(Ivan Derzhanski)}
```

```
\def \probword {Opgave}
\def \asgtword {Opdracht}
\def \comment {\paragraph {Noot:}}
```

```
\newpage
\problem {25}
%
Below you see arithmetic equalities
written in Egyptian Arabic%
\footnote{[…]}.
%
All summands, as well as all sums except the last one, are represented as
fractions in which neither the numerators nor the denominators are greater
than~$10$, nor is any denominator equal to~$1$:
%
\fracdata
%
\assignment Write these equalities in figures.
\assignment The equality \hfill \fractest \hfill is missing a sign.\\ Which
one?
\comment
The letter \wipa{\sh} is pronounced as English \word{sh}, \wipa x as the
\word{ch} in \word{loch};
\wipa C is a specific Arabic consonant.
A bar above a vowel indicates length.
\by{(Ivan Derzhanski)}
```

```
\def \probword {Problem}
\def \asgtword {Assignment}
\def \comment {\paragraph {Note:}}
```

```
\newcommand \problem [1]{\section*{\probword\ \stepcounter
  {section}\thesection\ (#1 \pontword)}}
\newcommand \assignment {\stepcounter {assignment}\paragraph
  {\asgtword~\theassignment.}}

\def \fractest {$\egar{rubC} + \egar{Ca{\sh}art its\A C} \; = \;
  \egar{sabaCt isd\A s}$}
```

Figure 2. Some excerpts from the Dutch and the English master files and the macro file.

language version and a file of common macro definitions, input by all master files.

Figure 2 shows how this works. The excerpts from the master files generate the text of the problem seen in Figure 1. Both master files refer to the shared macro file for the set of equalities in the data (`\fracdata`) and the equality in the assignment (`\fractest`). The macro file also takes care of the uniformity of the formatting of problems and assignments, although the words for 'Problem' and 'Assignment' in the respective languages are defined in the master files (`\probword` and `\asgtword`).

The same technique saves repetition within each text, for example, when handling a very common form of assignment:

```
\def \fordword #1{Vertaal in het #1}

\assignment \fordword {Nederlands}:      [twice (in Problem 1 and in Problem 4)]
\assignment \fordword {Baskisch}:
\assignment \fordword {Adygisch}:
\assignment \fordword {Adygisch}, op alle mogelijke manieren:
```

```
\def \fordword #1{Translate into #1}

\assignment \fordword {English}:         [twice (as above)]
\assignment \fordword {Basque}:
\assignment \fordword {Adyghe}:
\assignment \fordword {Adyghe} in all possible ways:
```

Figure 3. Some more excerpts from the Dutch and the English master files for IOL1.

The system made the production of the six parallel problem sets significantly more efficient and reliable than if six separate documents had been written. Still, much material is shared by the source files, and as can be seen from Figure 1, the texts in Dutch and in English differ more than they need to.

3 The Present: IOL6 and onwards

The problem sets for IOL2–5 were prepared in Microsoft Word as separate documents, and the identity of the unknown language material as well as the equivalence of the Solverese texts was checked entirely by human eye and hand. By the time the LaTeX-based multilingual system was revived (in 2008), things had changed in several respects. The number of participants in IOL had grown significantly, as had the quantity and diversity of the working languages; IOL itself had become more mature, and harder problems were being assigned; most importantly, the awareness of IOL's Problem Committee of the need to invest more time and attention into the preparation of the problem set (Derzhanski *et al.*, 2004) had increased. But with only so many days in the year, this all meant that the multilingual process often had to start before the content of the problems had been finalised, with changes sometimes proving necessary as an effect of this process, as it emerged that some problems (or parts of them), especially problems involving word semantics, would be easier, or certain explanations make more sense, in some languages than in others.[4] And having to make the same content change in several parallel texts is undesirable, for obvious reasons.

Therefore when the system came back to life in the weeks before IOL6, it did so as its own antithesis. In the new version, which has been in use ever since, the main source files for the individual Solverese versions are very brief. Apart from setting the paper size and the encodings and invoking the Babel package (Braams, 2008) with the appropriate language settings, each inputs two other files. One is composed entirely of macro definitions; this is effectively a pseudocode-to-Solverese dictionary. The other is the text of the problems (statements and solutions), the same for all versions, written entirely in the said pseudocode.

[4] Several early versions of Problem IOL10#5 (on Rotuman, by Boris Iomdin and Alexander Piperski) required the solver to make the conjecture that in Rotuman the word for 'grey' is derived from the word for 'ashes', but this word was removed from the assignment at the final stage, when it was brought to the Problem Committee's attention that the same is true of three of IOL10's working languages.

The canonical solution of Problem IOL5#3 (on Georgian verb morphology, by Yakov Testelets), first composed in Russian, suggested that *predsedatel'stvovat'* was too long a word to gloss a suppletive Georgian verb; this was crossed out because the corresponding verb in English, *chair*, is arguably only two phonemes long.

The original Russian text of Problem IOL1#1 (on Jacob Linzbach's Transcendental Algebra, by Ksenia Gilyarova) glossed the verb ♥ in the same way (*ljubit'*) whether it referred to loving people or liking things, but the final version used different expressions because in Estonian there was no other choice.

Opgave Nr 3 (20 punten). Gegeven zijn enkele zinnen in het Baskisch evenals hun vertalingen in het Nederlands in willekeurige volgorde. Een van de Nederlandse zinnen correspondeert met twee zinnen in het Baskisch:

ahaztu ditut, ahaztu zaizkit, ahaztu zaizu, hurbildu natzaizue, hurbildu zait, lagundu ditugu, lagundu dituzu, lagundu dute, lagundu nauzue, mintzatu natzaizu, mintzatu gatzaizkizue, mintzatu zaizkigu, ukitu ditugu, ukitu naute

jij vergat hem, zij spraken met ons, ik naderde jullie, ik sprak met jou, wij hielpen hen, jullie hielpen mij, hij naderde mij, wij raakten hen aan, zij raakten mij aan, jij hielp hen, zij hielpen hem, wij spraken met jullie, ik vergat hen

(a) Bepaal wat bij elkaar hoort.

(b) Vertaal naar het Baskisch: jij raakte mij aan, zij naderden mij.

(c) Vertaal naar het Nederlands: ***lagundu dut***, ***hurbildu gatzaizkizu***.

(d) Een van de Nederlandse zinnen kan ook op een andere manier worden vertaald naar het Baskisch. Bepaal welke zin dat is en geef de andere mogelijke vertaling.

—*Natalya Zaika*

Problem #3 (20 points). Here are some sentences in Basque as well as their English translations in arbitrary order. One of the English sentences corresponds to two sentences in Basque:

ahaztu ditut, ahaztu zaizkit, ahaztu zaizu, hurbildu natzaizue, hurbildu zait, lagundu ditugu, lagundu dituzu, lagundu dute, lagundu nauzue, mintzatu natzaizu, mintzatu gatzaizkizue, mintzatu zaizkigu, ukitu ditugu, ukitu naute

you$_{\text{sg}}$ forgot him, they talked to us, I approached you$_{\text{pl}}$, I talked to you$_{\text{sg}}$, we helped them, you$_{\text{pl}}$ helped me, he approached me, we touched them, they touched me, you$_{\text{sg}}$ helped them, they helped him, we talked to you$_{\text{pl}}$, I forgot them

(a) Determine the correct correspondences.

(b) Translate into Basque: you$_{\text{sg}}$ touched me, they approached me.

(c) Translate into English: ***lagundu dut***, ***hurbildu gatzaizkizu***.

(d) One of the English sentences can be translated into Basque in one more way. Identify this sentence and give the other possible translation.

—*Natalya Zaika*

בעיה מס׳ 3 (20 נקודות). להלן משפטים בשפה הבסקית ותרגומיהם לעברית בערבוביה. אחד המשפטים בעברית תואם שני משפטים בבסקית:

ahaztu ditut, ahaztu zaizkit, ahaztu zaizu, hurbildu natzaizue, hurbildu zait, lagundu ditugu, lagundu dituzu, lagundu dute, lagundu nauzue, mintzatu natzaizu, mintzatu gatzaizkizue, mintzatu zaizkigu, ukitu ditugu, ukitu naute

שכחת ממנו, הם דיברו אתנו, ניגשתי אליכם, דיברתי אתך, עזרנו להם, עזרתם לי, הוא ניגש אלי, נגענו בהם, הם נגעו בי, עזרת להם, הם עזרו לו, דיברנו אתכם, שכחתי מהם

(א) קבעו את ההתאמות הנכונות.

(ב) תרגמו לשפה הבסקית: נגעת בי, הם ניגשו אלי.

(ג) תרגמו לעברית: ***lagundu dut***, ***hurbildu gatzaizkizu***.

(ד) אחד המשפטים בעברית ניתן לתרגום לשפה הבסקית בדרך אחת נוספת. מצאו את המשפט הזה וציינו את התרגום האפשרי השני.

—נטליה זאיקה

Figure 4. Half a page from the Dutch, English and Hebrew versions of the IOL10 problem set.

```
\problem \givesent {\inlgEus} \andtrans {\tothislang} \chaotict. \pasoreus:
%
\begin{center}
\bord{ahaztu ditut, ahaztu zaizkit, ahaztu zaizu, hurbildu natzaizue,
hurbildu zait, lagundu ditugu, lagundu dituzu, lagundu dute,
lagundu nauzue, mintzatu natzaizu, mintzatu gatzaizkizue,
mintzatu zaizkigu, ukitu ditugu, ukitu naute}\medskip

\ahazty 23, \mintzaty 64, \hurbildy 15, \mintzaty 12, \lagundy 46,
\lagundy 51, \hurbildy 31, \ukity 46, \ukity 61, \lagundy 26,
\lagundy 63, \mintzaty 45, \ahazty 16
\end{center}
%
\begin{assgts}
\item \corrcorr.
\item \fordinto {\tolgEus}: \ukity 21, \hurbildy 61.
\item \fordinto {\tothislang}:
  \bord{lagundu dut}, \bord{hurbildu gatzaizkizu}.
\item \formahat {\tolgEus}. \findtran.
\end{assgts}
%
\by{—\NZname}
```

```
\def \givesent #1{Gegeven zijn enkele zinnen in het #1}
\def \andtrans #1{evenals hun vertalingen in het #1}
\def \chaotict{in willekeurige volgorde}
\def \fordinto #1{Vertaal naar het #1}

\def \inlgEus{Baskisch}
\def \tolgEus{Baskisch}
\def \tothislang{Nederlands}

\def \mintzaty #1#2{\iN{#1} sprak\iJ{#1} met \iA{#2}}
\def \ukity #1#2{\iN{#1} raakte\6#1(,,,n,n,n) \iA{#2} aan}

\def \iN #1{\6#1(ik,jij,hij,wij,jullie,zij)}
\def \iA #1{\6#1(mij,jou,hem,ons,jullie,hen)}
\def \iJ #1{\6#1(,,,en,en,en)}
```

```
\def \givesent #1{Here are some sentences in #1}
\def \andtrans #1{as well as their #1 translations}
\def \chaotict{in arbitrary order}
\def \fordinto #1{Translate into #1}

\def \inlgEus{Basque}
\def \tolgEus{Basque}
\def \tothislang{English}

\def \mintzaty #1#2{\iN{#1} talked
to \iA{#2}}
\def \ukity #1#2{\iN{#1} touched \iA{#2}}

\def \iN #1{\6#1(I,y\ous,he,we,y\oup,they)}
\def \iA #1{\6#1(me,y\ous,him,us,y\oup,them)}
```

In the English master file:

```
\def \ous {ou$_{\textrm{\small sg}}$}
\def \oup {ou$_{\textrm{\small pl}}$}
```

Figure 5. Excerpts from the pseudocode source and the Dutch and English dictionaries.

Figure 4 presents half a page from the Dutch, English and Hebrew versions of the IOL10 problem set; Figure 5, the text of this problem in pseudocode and some excerpts from the Dutch and English dictionary files.

How much granularity is desirable depends on the variety of the data and the regularity of the relevant fragments of the grammars of the featured and the working languages. Breaking down a sentence such as *He ate the fish* (and its equivalents) into subject, verb and object and generating each by its own macro makes the most sense if the same constituents also appear elsewhere in the text, but it always makes verification easier.

One important advantage of this approach over the others was already noted: if a content change in some problem (adding, replacing or deleting some item in the data or the assignments) is required, it is made in one place only, reducing the danger of error. Another lies in the making of the dictionaries. Those are prepared by filling the cells of a spreadsheet, with all languages in parallel columns. Figure 6 shows a screenshot containing part of the spreadsheet for IOL10 (several rows and six of the 15 working languages).

	RU	NL	EN	RO	HU	HE
givesent #1	Даны предложения н	Gegeven zijn enkele z	Here are some sentenc	Sunt date propoziţii în	Adva vannak mondato	להלן משפטים ב#1
andtrans #1	и их переводы на #1	evenals hun vertalinge	as well as their #1 trar	şi traducerile lor în #1	valamint fordításuk #1	ותרגומיהם ל#1
chaotict	в перепутанном поря	in willekeurige volgor	in arbitrary order	în ordine arbitrară	véletlen sorrendben	בערבוביה
fordinto #1	Переведите на #1	Vertaal naar het #1	Translate into #1	Traduceţi în #1	Fordítsa le #1	תרגמו ל#1
inlgEus	баскском языке	Baskisch	Basque	limba bască	baszk nyelven	שפה הבסקית
tolgEus	баскский язык	Baskisch	Basque	limba bască	baszkra	שפה הבסקית
tothislang	русский язык	Nederlands	English	română	magyarra	עברית
mintzaty #1#	\iN{#1} говорил\ifru	\iN{#1} sprak\iJ{#1}	\iN{#1} talked to \iA{	\iN{#1} \iJ{#1} vorbi	\iN{#1} beszélt\iJ{#1	\iN{#1}דיבר\iJ{#1}ת
ukity #1#2	\iN{#1} дотронул\ifr	\iN{#1} raakte\6#1(„	\iN{#1} touched \iA{	\iN{#1} \iA{#2}-\iJ{	\iN{#1} \iA{#2} nyúl	\iN{#1}נגע\iJ{#1}ב\i
iN #1	\6#1(я,ты,он,мы,вы,с	\6#1(ik,jij,hij,wij,julli	\6#1(I,y\ous,he,we,y\c	\6#1(eu,tu,el,noi,voi,e	\6#1(én,te,ő,mi,ti,ők)	\ifxהוא3#1 \else\ifxהמ
paani #1#2	\Uppcase \iN{#1} съе	\Uppcase \iN{#1} he\	\Uppcase \iN{#1} ate	\Uppcase \iN{#1} \iJ{	Megett\6#1(em,ed,e,ü	\iN{#1}אכל\iJ{#1} #2
bonaiana	рыбу	de vis	the fish	peştele	a halat	את הדג
bonaover	кокос	de kokosnoot	the coconut	nuca de cocos	a kókuszt	את הקוקוס

Figure 6. A screenshot of part of the multilingual spreadsheet.

This makes it easy to compare words or sentences in any two languages and to find mismatches and imbalances. Also, since the ordering of the rows of the spreadsheet is immaterial, they can be arranged and rearranged to group certain words or sentences in close rows in order to make similarities or differences stand out.[5]

A final advantage is the move away from the model (disadvantageous for more than one reason[6]) in which the version of a problem in one working language is the original and the other versions are translations. The parallel production of all Solverese versions from the same pseudocode source and with use of dictionaries made from a table where all working languages are uniformly situated creates the effect of (machine) translation from pseudocode to all languages, which in turn makes all languages equal. At a contest such as IOL, where all contestants are to have the same chances regardless of their working languages, this is of vital importance.

The method has been tested and proven to work with two Cyrillic-written and 12 Roman-written languages, as well as Korean (at IOL7) and Hebrew (at IOL10), with hardly any technical difficulties. It remains to be seen whether it will meet just as cheerfully the predictable further growth of the number and diversity of IOL's working languages, but it is certain that its potential has not yet been fully explored.

[5] One of the phenomena in Problem IOL10#1 (on Dyirbal, by Artūrs Semeņuks) was factitive morphology; it was illustrated by several deadjectival verbs, which could be translated as lexical factitives (*bent → bend, healthy → heal*) or as periphrastic ones (*fat → make fat, sleep → make fall asleep*), but which were which differed from one working language to the other. In order to guarantee the equal difficulty of the problem in all versions it was necessary to ensure that each language used factitives of several types, which was facilitated by the summary character of the spreadsheet.

[6] At IOL5, where some versions of the problem set were made by translating the English one, the sentence 'Knowledge of English is not necessary for solving the problem' was supposed to be present in one of the problems, but was omitted from the English version (because of its obvious inappropriateness there) and therefore didn't make it into the other ones either; this was considered a grave mishap.

References

B. Bozhanov and I. Derzhanski. 2013. Rosetta Stone Linguistic Problems. In this volume.

Johannes Braams. 2008. Babel, a multilingual package for use with LaTeX's standard document classes. CTAN:macros/latex/required/babel/.

I.A. Derzhanski, A.S. Berdichevsky, K.A. Gilyarova, B.L. Iomdin, E.V. Muravenko, and M.L. Rubinstein. 2004. On the Translatability of Linguistic Problems: the Lessons of the First International Linguistics Olympiad. In: I.M. Kobozeva, A.S. Narinyani, and V.P. Selegey (eds.), *Proceedings of the International Conference Dialogue 2004: Computational Linguistics and Intellectual Technologies*. Nauka, Moscow, 166–171 (in Russian).

Ivan A. Derzhanski. 2009. *Linguistic Magic and Mystery*. Union of Bulgarian Mathematicians, Sofia.

I.A. Derzhanski and T.E. Payne. 2009. The Linguistics Olympiads: Academic competitions in linguistics for secondary school students. In: K. Denham and A. Lobeck (eds.), *Linguistics at School: Language Awareness in Primary and Secondary Education*, Cambridge University Press, Cambridge, UK, 213–226.

Donald E. Knuth. 1986. *The TEXbook*. Addison–Wesley, Reading, MA.