EUROCALL 2023: CALL for all Languages
15-18 August 2023, University of Iceland, Reykjavik
DOI: https://doi.org/10.4995/EuroCALL2023.2023.16952

Making ICALL's intelligence accessible and understandable: evaluation, validation, and future directions of ICALL ecosystems

Jasper Degraeuwe[a] and Patrick Goethals[b]
[a]Department of Translation, Communication and Interpretation, Ghent University, , Jasper.Degraeuwe@UGent.be and
[b]Department of Translation, Communication and Interpretation, Ghent University, , Patrick.Goethals@UGent.be

How to cite: Degraeuwe, J.; Goethals, P. (2023). ICALL ecosystems: evaluation, validation and future directions. In *CALL for all Languages - EUROCALL 2023 Short Papers.* 15-18 August 2023, University of Iceland, Reykjavik.
https://doi.org/10.4995/EuroCALL2023.2023.16952

Abstract

In this paper, we will reflect upon the revised and expanded version of the Intelligent Computer-Assisted Language Learning (ICALL) ecosystem developed in Degraeuwe and Goethals (2022a). We portray the upgraded version of the ecosystem and perform a reflective analysis on it following the same methodology as presented in Degraeuwe and Goethals (2022a). The analysis is carried out by means of a case study with 22 participants who study Spanish as a foreign language and who work with the ecosystem in a blended writing course focused on vocabulary learning and lexical ambiguity. Students' attitudes towards engaging in the ICALL ecosystem are gauged through a questionnaire, which revealed a slight but non-significant positive change in attitude (compared to a statistically significant positive change in Degraeuwe and Goethals 2022a). However, it should be noted that the initial attitudes before engaging with the ecosystem were already very positive (5.32 on an 8-point Likert scale), which might have made it more challenging for the ecosystem to have a considerable impact. Additionally, an analysis at question level showed that sparking curiosity about language technology and providing a good user experience remain challenging aspects.

Keywords: *Intelligent Computer-Assisted Language Learning, Natural Language Processing, reflective analysis, user attitudes.*

1. Introduction

With applications such as example sentence selection systems (Pilán et al., 2016), difficulty classifiers (Tack, 2021) and exercise generation tools (Bodnar, 2022; Heck & Meurers, 2022), the integration of Intelligent Computer-Assisted Language Learning (ICALL) in language learning courses can be a valuable addition to the arsenal of teaching methods, for example as a complement to on-campus vocabulary learning activities (Ruiz et al., 2021). Nevertheless, using ICALL also comes with its limitations. Recognising lexically ambiguous items and distinguishing between their senses is one of those pending issues in ICALL research, especially for languages other than English (Degraeuwe & Goethals, 2022b). In fact, the Natural Language Processing (NLP) driven technique of Word Sense Disambiguation (WSD) is rarely integrated in corpus query tools or in the development of computer readable resources for didactic purposes, such as graded word lists (Tack, 2021). Additionally, if end users (which may refer to students, teachers or even course book designers) want to fully exploit the potential of ICALL environments, they should possess certain technological metaskills that allow them to decide which queries are relevant and feasible to perform and which are not (Schweinberger, 2021).

In Degraeuwe and Goethals (2022a), we analysed how both abovementioned aspects could be addressed by designing an "ICALL ecosystem" (Section 2.1). Working in such ecosystem led to a statistically significant positive change in attitude towards ICALL, but this did not automatically mean that the users also enjoyed working with the computer more, or that the ecosystem sparked their interest in learning more about language technology. With the present study, we aim to corroborate the findings of Degraeuwe and Goethals (2022a) by performing a similar case study based on a revised and expanded version of the ecosystem (see Section 2.3). Furthermore, this follow-up study will also briefly touch upon the use of generative Artificial Intelligence (AI), as the adequate and efficient use of AI driven applications is bound to become an essential skill in education.

2. Method

2.1. Ecosystem design

As described in Degraeuwe and Goethals (2022a), the ecosystem enables users to generate and use customised learning materials **(Aspect_1)** and helps them gain technological metaskills by stimulating their curiosity and promoting their autonomy **(Aspect_2)**. In the meantime, all user activities are saved in a structured database, which can then be used for improving the NLP driven methods integrated into the environment **(Aspect_3)**.

2.2. Case study design

The study follows the same design as Degraeuwe and Goethals (2022a). The 22 participants (Dutch-speaking students of Spanish as a foreign language enrolled in a 3rd bachelor B2+ level writing course at university) work with the online learning environment of the Spanish Corpus Annotation Project [1] (SCAP) (Goethals, 2018), which includes a corpus consultation component, exercise generator and collaborative section for research purposes.

Part of the course consists of completing a blended vocabulary learning module, which encompasses five hours of on campus classes and an online module on lexical ambiguity. During the classes, the students learn to use the corpus consultation and vocabulary learning functionalities of the SCAP platform **(Aspect_1)**. For the online module, they consider lexical ambiguity from the perspective of the computer by watching knowledge clips [2] and develop their own WSD models by making interactive exercises on lexically ambiguous vocabulary items in the collaborative section of the platform **(Aspect_2)**. The responses to those exercises are collected in a database and used to develop the actual WSD method integrated into the environment **(Aspect_3)**.

2.3. Modified aspects of the ecosystem

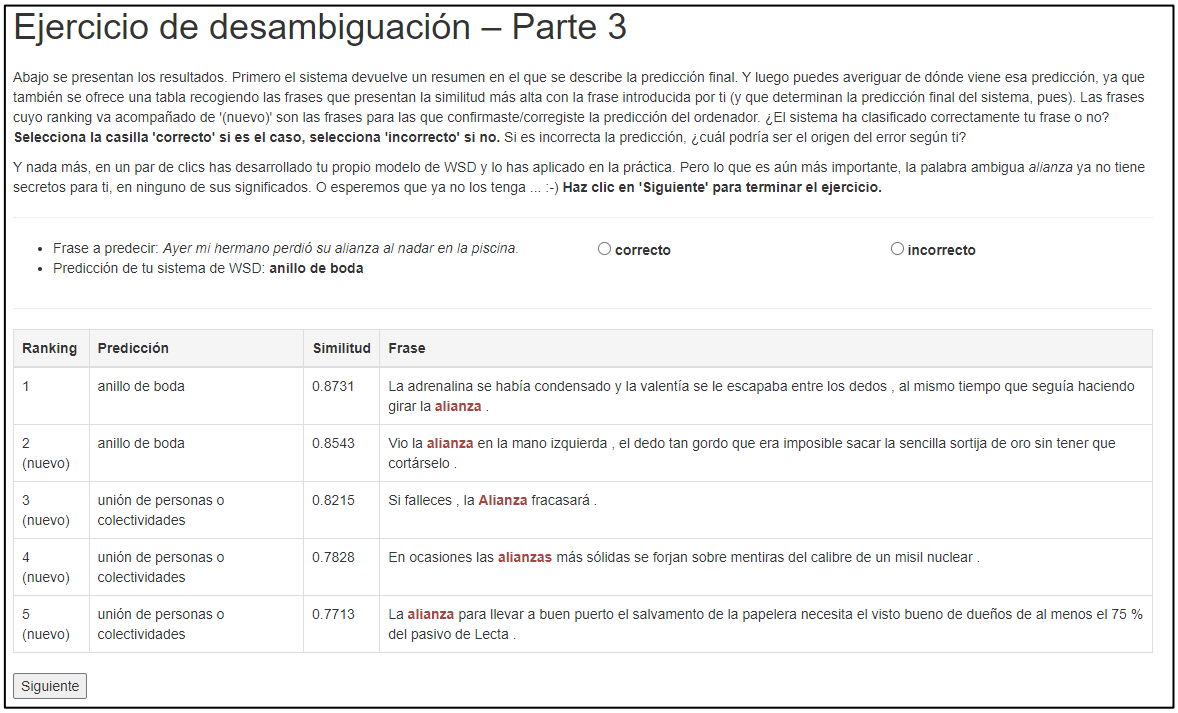
From Degraeuwe and Goethals (2022a) it could be concluded that, although the ecosystem led to enhanced insights into NLP and increased confidence in the computer as a learning assistant, the students did not necessarily enjoy working with the computer more or wanted to learn more about language technology. To address these shortcomings, we made the following adjustments to the ecosystem:

- Integration of short quizzes into the knowledge clips
- Addition of a free text component in the interactive exercises on lexical ambiguity (see Figure 1)
- Addition of "odd one(s) out" as a new type of exercise in the collaborative section of the platform: in this exercise, students are presented with six sentences for one particular meaning of a lexically ambiguous word after which they have to identify the sentence(s) which do not belong to that particular meaning

[1] Publicly available version of the platform accessible through scap.ugent.be. Demo video of the in-house version available at https://www.youtube.com/watch?v=RFaIWEEZcVM.

[2] Complete video available at https://www.youtube.com/watch?v=-ev56uEpIkA.

- Possibility to download customised learning materials created with NLP models (e.g. the WSD models) trained on the exercise response data collected in the ecosystem



Ejercicio de desambiguación – Parte 3

Abajo se presentan los resultados. Primero el sistema devuelve un resumen en el que se describe la predicción final. Y luego puedes averiguar de dónde viene esa predicción, ya que también se ofrece una tabla recogiendo las frases que presentan la similitud más alta con la frase introducida por ti (y que determinan la predicción final del sistema, pues). Las frases cuyo ranking va acompañado de '(nuevo)' son las frases para las que confirmaste/corregiste la predicción del ordenador. ¿El sistema ha clasificado correctamente tu frase o no? **Selecciona la casilla 'correcto' si es el caso, selecciona 'incorrecto' si no.** Si es incorrecta la predicción, ¿cuál podría ser el origen del error según ti?

Y nada más, en un par de clics has desarrollado tu propio modelo de WSD y lo has aplicado en la práctica. Pero lo que es aún más importante, la palabra ambigua *alianza* ya no tiene secretos para ti, en ninguno de sus significados. O esperemos que ya no los tenga ... :-) **Haz clic en 'Siguiente' para terminar el ejercicio.**

- Frase a predecir: *Ayer mi hermano perdió su alianza al nadar en la piscina.* ○ **correcto** ○ **incorrecto**
- Predicción de tu sistema de WSD: **anillo de boda**

Ranking	Predicción	Similitud	Frase
1	anillo de boda	0.8731	La adrenalina se había condensado y la valentía se le escapaba entre los dedos , al mismo tiempo que seguía haciendo girar la **alianza** .
2 (nuevo)	anillo de boda	0.8543	Vio la **alianza** en la mano izquierda , el dedo tan gordo que era imposible sacar la sencilla sortija de oro sin tener que cortárselo .
3 (nuevo)	unión de personas o colectividades	0.8215	Si falleces , la **Alianza** fracasará .
4 (nuevo)	unión de personas o colectividades	0.7828	En ocasiones las **alianzas** más sólidas se forjan sobre mentiras del calibre de un misil nuclear .
5 (nuevo)	unión de personas o colectividades	0.7713	La **alianza** para llevar a buen puerto el salvamento de la papelera necesita el visto bueno de dueños de al menos el 75 % del pasivo de Lecta .

Siguiente

Figure 1. Illustration of free text component added to exercise in which users can develop, use, and evaluate their own WSD models (for a translation, see Appendix 1). The table at the bottom presents the similarity values based on which the model obtained its prediction ("Predicción de tu sistema de WSD") for the sentence introduced by the user ("Frase a predecir")

2.4. Questionnaire

To gain insights into the potential of the upgraded ecosystem, the students are administered the same questionnaire as in Degraeuwe and Goethals (2022a), namely an adapted version of the Attitude towards CALL questionnaire (A-CALL; Vandewaetere & Desmet, 2009). The questionnaire contains 15 eight-point Likert scale questions (each of them representing a specific attitude towards ICALL; see Table 1) and is filled out by the students before and after completing the vocabulary learning module. Additionally, we include an extra multiple-choice question to analyse which technology assisted tools the participants use in their language learning process (see Figure 2 at the end of Section 3 for the multiple choice options).

3. Results

3.1. Attitudes towards ICALL

Table 1 reports the mean scores and Standard Deviation (SD) values of the 15 questions, with the scores from Degraeuwe and Goethals (2022a), hereafter referred to as the 'previous case study', being added between parentheses. The overall pre- and post-scores show a small positive change in attitude of 0.08, which did not appear to be statistically significant (paired samples t-test, p=0.45). However, it should be highlighted that the pre-scores were already quite elevated, which might have made it more challenging for the ecosystem to significantly affect students' attitudes. In fact, the overall pre-score of 5.32 more or less equals the overall post-score of 5.36 obtained in the previous case study. In summary, the overall results indicate that the ecosystem and its upgrades are not able to further improve attitudes which are already fairly to very positive before the interaction with the ecosystem takes place.

When breaking down the results at question level, the same tendency as in the previous case study can be observed: working in the ecosystem enhances students' insights into NLP (questions 1, 5 and 9) and increases their confidence in the computer as a vocabulary learning assistant (4 and 6), but this does not necessarily go hand in hand with an increased curiosity (2) and a better user experience (11). This tension also appears from the statistical significance at question level (paired samples t-test, $p<0.05$): the increases for questions 1 and 6 are significant, the decreases for questions 2 and 11 as well.

Table 1. Questionnaire results, with questions for which the scores were reversed being marked with $^{(rev)}$ and post-means with a statistically significant difference being marked with *.

Nr	Question	Pre (n=22)		Post (n=22)	
		Mean	SD	Mean	SD
1	The computer is able to analyse the grammatical characteristics of words and link words to their corresponding part of speech (noun, verb, adjective, etc.).	5.95 (5.12)	1.21 (1.62)	6.64* (7.18)	1 (0.86)
2	I am interested in knowing more about the technology which enables computers to automatically create vocabulary exercises and resources.	5.32 (4.16)	1.52 (2.2)	4.73* (3.89)	2.07 (1.87)
3	The computer only sees sequences of letters which are combined into words, it is not able to see meanings and concepts behind these sequences of letters.$^{(rev)}$	4.91 (4.91)	1.69 (1.51)	4.68 (5.96)	1.81 (1.48)
4	I have confidence in computer created vocabulary exercises and tests.	5.27 (4.69)	1.28 (1.31)	5.33 (5.86)	1.53 (1.3)
5	If I introduce a large collections of texts on a certain domain into a specific application, I think that this application will be able to return a keyword list with the most typical words for the domain.	6.41 (5.56)	1.3 (1.37)	6.68 (6.71)	0.95 (1.05)
6	The computer is able to generate vocabulary exercises and resources tailored to my proficiency level.	5.5 (5.34)	1.5 (1.21)	6.55* (6.68)	1.06 (1.22)
7	The teacher's attitude and enthusiasm towards and knowledge of computer-assisted vocabulary learning determine to a large extent my attitude towards using computers for vocabulary learning purposes.$^{(rev)}$	3.09 (3.22)	1.41 (1.77)	3 (3.54)	1.69 (2.12)
8	Computer-assisted vocabulary learning offers more flexibility to learning vocabulary in Spanish.	5.68 (5.33)	0.84 (1.5)	6.05 (5.61)	1.17 (1.89)
9	The computer is able to analyse the syntactic structure of sentences, and assign the correct syntactic function (subject, direct object, etc.) to words.	5.32 (4.53)	1.32 (1.27)	5.64 (5.61)	1.33 (1.47)
10	Computer assisted vocabulary learning is as valuable as traditional methods for vocabulary learning in Spanish.	5.77 (4.28)	1.48 (1.49)	5.95 (4.82)	1.43 (1.49)

11	I (would) like to learn Spanish vocabulary with the help of the computer.	6.77 (5.28)	1.07 (2.1)	5.95* (4.89)	1.21 (1.91)
12	I find it easier to accept an error committed by a language teacher than an error committed by the computer.(rev)	3.64 (4.06)	1.92 (1.78)	4 (4.07)	1.83 (2.02)
13	People who learn Spanish vocabulary through computer assisted learning methods are less proficient in Spanish than people who learn Spanish vocabulary through traditional paper and pencil methods.(rev)	6.5 (6)	1.19 (1.93)	6.32 (6.29)	1.52 (1.41)
14	Computer assisted vocabulary learning is a valuable extension of traditional learning methods for vocabulary learning in Spanish.	6.73 (6.09)	1.08 (1.47)	6.59 (6.07)	1.1 (1.74)
15	Vocabulary exercises and resources created automatically by an application cannot contain errors.(rev)	2.91 (3.34)	1.57 (2.06)	2.82 (3.18)	1.59 (1.93)
		5.32 (4.81)	**1.36 (1.64)**	**5.4 (5.36)**	**1.42 (1.58)**

3.2. Use of technology-assisted tools

Finally, we briefly discuss the results of the additional multiple-choice question on the use of technology-assisted tools by the students (Figure 2). The bar chart shows that language learning applications and machine translation systems are widely used by students, and CALL environments to a lesser extent (only 45% of the participants had experience with them before the start of the course). As for the use of generative AI, the results clearly show that AI-driven tools will inevitably become a part of language learning: before taking the course, none of the students had used tools like ChatGPT (the course started 3 months after the release of GPT-3.5), but by the end of the course this number had already risen to 8 of the 22 participants. In other words, this finding suggests that the integration of generative AI will become an important aspect of future ICALL research.

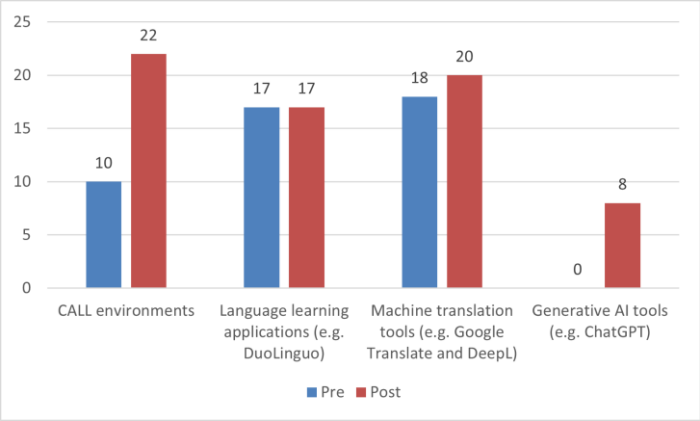


Figure 2. Bar chart presenting which technology-assisted tools are used by the 22 students

4. Discussion and conclusions

In this paper, we presented and reflected upon the upgraded version of the "ICALL ecosystem" introduced in Degraeuwe and Goethals (2022a). By using the same case study design we aimed to corroborate the findings of Degraeuwe and Goethals (2022a) which reported a significantly positive change in students' attitudes towards ICALL after having engaged with the ecosystem. However, this year's questionnaire results revealed only a small

(and statistically non-significant) positive change in attitude. It does have to be highlighted, though, that the initial attitudes of this year's participants were already very positive (average score of 5.32 on an 8-point scale), which might have made it more challenging for the ecosystem to achieve the same impact (note that last year's post-score amounted to 5.36, coming from a pre-score of 4.81). Additionally, despite the upgrades integrated into the ecosystem the findings again underlined the area of tension between what students consider to be the value, quality, and/or potential of learning methods and the user experience these methods provide/the interest they spark. Finally, the additional question on the use of technology assisted tools showed that these kinds of applications are widely used by students, with generative AI as an emerging new source.

To conclude, finding alternative ways to improve the user experience (e.g. the integration of generative AI) will require further research, although the lack of corroboration for the ecosystem's positive impact on user attitudes might also be considered an incentive to first redesign some of its core aspects. Furthermore, we also plan to perform experiments with teachers as a new type of target audience, for example by asking them to prepare a vocabulary learning class using the ecosystem. Finally, grouping the questionnaire items into different variables (e.g. curiosity and quality of learning assistance) and performing a correlation analysis on them could be another interesting avenue for future research.

Acknowledgements

This research has been carried out as part of a PhD fellowship on the IVESS project (file number 11D3921N), funded by the Research Foundation – Flanders (FWO). Additionally, we want to express our gratitude to the reviewers for their valuable feedback, and to Dr. Koen Plevoets for providing support on the statistical analysis.

References

Bodnar, S. (2022). The instructional effectiveness of automatically generated exercises for learning French grammatical gender: Preliminary results. In D. Alfter, E. Volodina, T. François, P. Desmet, F. Cornillie, A. Jönsson, & E. Rennes (Eds.), *Proceedings of the 11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022)* (Vol. 190, pp. 10–22). https://doi.org/10.3384/ecp190002

Degraeuwe, J., & Goethals, P. (2022a). ICALL ecosystems: making ICALL's intelligence both accessible and understandable. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny, & M. J. Whelpton (Eds.), *Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022* (pp. 89–94). https://doi.org/10.14705/rpnet.2022.61.1440

Degraeuwe, J., & Goethals, P. (2022b). Interactive word sense disambiguation in foreign language learning. In D. Alfter, E. Volodina, T. François, P. Desmet, F. Cornillie, A. Jönsson, & E. Rennes (Eds.), *Proceedings of the 11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022)* (Vol. 190, pp. 46–54). https://doi.org/10.3384/ecp190005

Goethals, P. (2018). Customizing vocabulary learning for advanced learners of Spanish. In T. Read, B. Sedano Cuevas, & S. Montaner-Villalba (Eds.), *Technological innovation for specialized linguistic domains: Languages for digital lives and cultures, proceedings of TISLID'18* (pp. 229–240). Éditions Universitaires Européennes.

Heck, T., & Meurers, D. (2022). Generating and authoring high-variability exercises from authentic texts. In D. Alfter, E. Volodina, T. François, P. Desmet, F. Cornillie, A. Jönsson, & E. Rennes (Eds.), *Proceedings of the 11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022)* (Vol. 190, pp. 61–71). https://doi.org/10.3384/ecp190007

Pilán, I., Volodina, E., & Borin, L. (2016). Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Revue Traitement Automatique Des Langues*, *57*(3), 67–91.

Ruiz, S., Rebuschat, P., & Meurers, D. (2021). The effects of working memory and declarative memory on instructed second language vocabulary learning: Insights from intelligent CALL. *Language Teaching Research*, *25*(4), 510–539. https://doi.org/10.1177/1362168819872859

Schweinberger, M. (2021). *An evaluation of computational, corpus-based approaches to language learning* [Webinar]. International Perspectives on Corpus Technology for Language Learning - Seminar Series, University of Queensland, Australia. https://languages-cultures.uq.edu.au/event/session/7415

Tack, A. 2021. *Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers*. Louvain-la-Neuve, Belgium: UCLouvain & KU Leuven PhD thesis

Vandewaetere, M., & Desmet, P. (2009). Introducing psychometrical validation of questionnaires in CALL research: The case of measuring attitude towards CALL. *Computer Assisted Language Learning*, *22*(4), 349–380. https://doi.org/10.1080/09588220903186547

Appendix 1. Translation of Figure 1

Below the results are presented. First, the system gives back an overview in which the final prediction is described. And then you can check where this prediction comes from because a table is added containing the sentences that show the highest similarity with the sentence introduced by you (and which thus determine the final prediction of the system). The sentences whose ranking is accompanied by '(new)' are the sentences for which you confirmed/corrected the prediction of the computer. Did the system classify your sentence correctly or not? **Tick the box 'correcto' if it's the case, tick 'incorrecto' if not.** If the prediction is incorrect, what could have caused the error according to you?

And that's it, in just a few clicks you've developed your own WSD model, and put it into practice. But what's even more important, the ambiguous item *alianza* doesn't hold any further secrets from you, in any of its meanings. Or let's hope that it doesn't … :-) **Click on 'Siguiente' to finish the exercise.**

- Sentence to predict: *Yesterday my brother lost his wedding ring while swimming in the pool*
- Prediction of your WSD system: **wedding ring**

Ranking	Prediction	Similarity	Sentence
1	wedding ring	0.8731	The adrenaline had been condensed and courage slipped through his/her fingers as he/she kept spinning the wedding ring.
2 (new)	wedding ring	0.8543	He/she saw the wedding ring on the left hand, the finger being so fat it was impossible to take off the simple gold ring without having to cut it off.
3 (new)	union of people or groups	0.8215	If you die, the Alliance will fail.
4 (new)	union of people or groups	0.7828	Sometimes the most solid alliances are forged based on lies with the calibre of a nuclear missile.
5 (new)	union of people or groups	0.7713	The alliance for the rescue of the paper business needs the approval of owners of at least 75% of Lecta's liabilities.