# Cohesion Grading Decisions in a Summary Evaluation Environment: A Machine Learning Approach

**Iraide Zipitria (iraide.zipitria@ehu.es)**
Psychology Faculty (UPV/EHU), Tolosa Etorbidea, 70
Donostia, E-20018

**Basilio Sierra, Ana Arruarte and Jon A. Elorriaga (b.sierra, a.arruarte, jon.elorriaga@ehu.es)**
Computer Science Faculty (UPV/EHU), Manuel Lardizabal Pasealekua
Donostia, E-20010

## Abstract

The work presented in this paper has been carried out in the context of a summary writing environment provided with automatic grading. Regarding summarisation discourse, some of the most relevant variables identified in previous work are comprehension, adequacy, use of language, coherence, and cohesion. This work is focused on cohesion. The described exploratory study starts from basic automatic measures of cohesion to further analyse which of them best reflects human expert overall cohesion grades for learner summaries written in the Basque language. For this purpose, 45 basic cohesion measures are compared to overall human cohesion grades. Machine Learning techniques are used to select the best combination for cohesion grading.

**Keywords:** Cohesion grading, machine learning, automatic scoring.

## Introduction

A summary is a short clear description that provides the main facts or ideas about a given topic. In educational contexts, a summary is an overview of the most important information on the studied theme. Summarising requires active meaning construction to a much greater degree than choosing a response in a multiple-choice test, or even than writing short answers to isolated open questions. Thus, not only is summary writing an effective means to construct and integrate new knowledge, it is also a more efficient method for assessing what students do and do not understand than traditional comprehension tests (E. Kintsch, Steinhart, Stahl, & the LSA Research Group, 2000). Thus, summaries are widely used in traditional teaching as an educational diagnostic strategy to infer comprehension, or how much information from the reading text is retained in memory (Bartlett, 1932; Garner, 1982; W. Kintsch, Patel, & Ericsson, 1999).

However, evaluating and grading summaries is a complex and time consuming task for teachers. Human judges have certain variance on summary grading. So, there is a need to systematise written summary evaluation for students. Researchers have sought to develop applications that automate summary grading and evaluation in a way that a given summary will always gain the same score.

Most of the work carried out in Computer Assisted Assessment has tried to infer the student's knowledge comprehension by analysing and comparing the answer generated by the student either explicitly represented in the system –mostly multiple choice questions– or with answers that could be obtained using the knowledge represented in the system. The automatic evaluation of open-ended text, e.g. summaries, is a complex task strongly conditioned by text comprehension methods; statistical modelling, and Natural Language Processing (NLP) techniques. The open-ended assessment mode, although less accurate than the close-ended mode, has been present in Artificial Intelligence and Education since the very early work in Socratic dialogues systems (Clancey, 1982; Ford, 1988; Woolf, 1988; Winkels & Breuker, 1989). After these first works, there was a period when open-ended approaches had a lower profile, but new developments in NLP and cognitive modelling have seen a revival with a variety of approaches in various applications: dialogue systems (Schulze et al., 2000; A. Graesser, Person, & Harter, 2001; Zinn, Moore, & Core, 2002), feedback on narratives (Robertson & Wiemer-Hastings, 2002), and so on.

The work presented in this paper has been carried out in the context of a learner oriented summary writing environment provided with automatic grading, LEA (Zipitria, Arruarte, & Elorriaga, 2008b). Relevant variables identified when producing a summarisation environment are: text related (text type, text present/absent, theme and text length), discourse related (comprehension, adequacy, use of language, coherence, and cohesion), learner related (learner level and learner's prior knowledge) and available aid tools (dictionaries, spell and grammar check, theory in summarisation strategies, concept maps, schema, etc.). Those variables have been identified after an in-depth study of both the state of the art in summary grading and an empirical study carried out to observe human summary grading performance to model their criteria (Zipitria, Larrañaga, Armañanzas, Arruarte, & Elorriaga, 2008a).

In the context of this work, the global summary grading decisions are gained by means of a Bayesian Network based modelling approach, based on measures such as: comprehension, adequacy, use of language, coherence, and cohesion (Zipitria et al., 2008b).

- Comprehension. Comprehension measures the level of understanding that can be inferred from each summary.

- Adequacy. It refers to the use of adequate register and terminology in the written summary.

- Use of language. It looks at orthographic, syntactic and lexical errors (Cassany, 1993).

- Cohesion and coherence. Coherence and cohesion are closely related and often used as synonyms. A way of distinguishing both concepts is suggested by A. C. Graesser, McNamara, Lowerse, and Zhiqiang (2004), who refer to coherence as a psychological construct, whereas cohesion is referred to as a textual construct. Similarly, Todd, Khongput, and Darasawang (2007) in a connective cohesion study say that cohesion refers to explicit connective links, whereas coherence refers to implicit connections. Therefore, coherence would exist in the way that people interpret text rather than in the texts themselves, while cohesion would be provided by the text features. Cohesion has been defined by Halliday and Hasan (1976) as a set of resources for constructing relationships in discourse transcending grammatical structure (reference, ellipsis, substitution, conjunction, lexical cohesion, etc.). Hence, the aim of cohesive studies is to measure the way text discourse is tied in language. Cohesion features have been measured in this study to resemble human global cohesion grades.

In LEA, comprehension and coherence are modelled based on Latent Semantic Analysis (Zipitria, Arruarte, & Elorriaga, 2006) and adequacy and use of language are computed based on surface measures gathered from tagged text and statistical analysis. The present study describes the procedure followed searching for the best available approach to model overall cohesion grading of learner summaries written in Basque language[1].

The paper is organised as follows. Section 1 is a summary of previous work that includes measures of cohesion and Section 2 describes the cohesion grading experimental setting, results and discussion.

## Previous work measuring cohesion

Cohesion has already been automatically measured under different approaches and for a variety of purposes.

Morris and Hirst (1991), in a domain independent approach, analyse lexical cohesion in text. Lexical cohesion is measured as a result of chains of related words that contribute to the continuity of lexical meaning. These lexical chains are a direct result of units on the same topic. A thesaurus is used as the knowledge base for computing lexical chains. Lexical chains are also used to determine text structure. E. Kintsch et al. (2000) took an LSA approach to cohesion, gaining sentence to sentence paraphrasing measures for learner summary grading purposes. Alonso and Fuentes (2003) describe the integration of cohesive properties with coherence for automatic summarisation purposes. An account for cohesive formation is gained by means of diagnosis of lexical cohesive chains as extra-strong, strong and medium-strong. A. C. Graesser et al. (2004) present a wide account in cohesion and coherence measures, producing over 200 measures –over 50 types of cohesion– based on surface linguistic features in a tool called Coh-Metrix. Siddharthan (2006) refers to work in automatic text production that applies a syntactic simplification process as a way to reduce comprehension complexity and maintain cohesiveness. Adequate sentence ordering, cue word selection, referring expression generation, determiner choice and pronominal use are resolved to preserve cohesiveness. Carenini, Ng, and Zhou (2008) work in the context of automatic summarisation of e-mail conversations. Cohesive measures are collected in the form of clue words or word co-occurrences between adjacent fragments, semantic similarity or subsequent sentence similarity measures based on WordNet and cosine – using TF(Term Frequency) and IDF (Inverse Document Frequency), local and global weights respectively – or segment to segment cosine similarity. Finally, Vechtomova and Karamuftuoglu (2008) measure lexical cohesion between query terms in the context of IR (Information Retrieval) term proximity. Both short distance and long distance collocation relations are measured.

## Cohesion grading experiment

As part of the modelling process to obtain global grades for each summary, the summary grading decision making model (Zipitria et al., 2008b) requires global cohesion grades. The goal of this study is to obtain a model which shows which combination of specific cohesion measures best predict cohesion. In other words, which cohesion features predict the decision of global cohesion of graders in comparison to a real-life cohesion grading task. Human cohesion grades are predicted by automatic measures of discourse cohesive features.

### Procedure

17 human experts were asked to grade the level of cohesion of summaries that had previously been gathered from university students, second language learners and primary and secondary school pupils. Experts were university lecturers or primary and secondary and L2 teachers who had been teaching summarisation strategies for more than a decade. A total of 17 summaries were written in Basque language. The goal was to obtain a wide range of different scenarios involving cohesion in summarisation. Grades were gathered on a 1 to 10. Each of the 17 raters produced grades for every summary with a between-rater agreement $r = 0.7$ and $p < 0.05$. Finally, all the grades were discretized into $Fail, Pass$ and $Distinction$.

The task for expert grading participants consisted of reading the text based on which the summaries were written. Next, they were expected to read each summary to produce global cohesion grades. In order to avoid misconception, verbal and written definitions on cohesion were provided to experts.

In parallel, cohesion measures were automatically modelled using NLP techniques. The mean scores of the graders were compared to cohesion measures in order to observe the amount of information explained by the cohesion measures.

[1] Non Indo-European language spoken in the north of Spain and south of France. Grammatically complex, it is an agglutinative, order free and verb final language. A complete English description of the Basque grammar can be found in Hualde and Ortiz de Urbina (Hualde & Ortiz de Urbina, 2003).

**Cohesion measures ($X$)** Cohesion measures were created (see Table 1) based on theory on English discourse cohesion (Halliday & Hasan, 1976; Schiffrin, Tannen, & Hamilton, 2001) and language specific differences for Basque (Hualde & Ortiz de Urbina, 2003). In addition, previous modelling work has also been taken into account (Baayen, 2001; A. C. Graesser et al., 2004). 45 markers aiming to word variability, text structure, lexical cohesion, conjunctions and verbal cohesion have been studied:

*Word variability*

A total of 14 measures which refer to vocabulary variability related information: $X1$ Size of the sample in word tokens, $X2$ Number of distinct lemmas, $X3$ Number of distinct word tokens, $X4$ Distinct concept proportion in text, $X5$ Concept proportion among word variability, $X6$ Mean number of letters per word, $X11$ Measures on how single word measures deviate from the central word mean tendency, $X12$ Mean of word tokens to number of distinct word types, $X13$ Word proportion in text and $X14$ Lemma proportion in text.

*Text Structure*

This refers to the cohesion which is inherent to the textual structure as narrative, formal correspondence, sonnet, etc. (Halliday & Hasan, 1976). Four surface structure measures have been measured to reflect structure: $X7$ Mean sentences per paragraph, $X8$ Number of paragraphs, $X9$ Number of sentences and $X10$ Average words per sentence.

*Lexical cohesion*

In lexical cohesion the same word is repeated and has the same referent in both cases. It is not necessary for the second instance to be an exact repetition of the same word (Halliday & Hasan, 1976).

Two measures emulate lexical cohesion indices measured by means of overlapping concepts in subsequent sentences. Overlapping concepts are measured as word overlap and lemma overlap: $X15$ Cosine of overlapping words in subsequent sentence comparison and $X16$ Cosine of overlapping lemmas in subsequent sentence comparison.

*Conjunction and connectors*

Conjunctive elements are cohesive by means of their specific meaning. They express meaning which presuppose the presence of other components in discourse. It is based on the assumption that there are forms of systematic relationships between sentences (Halliday & Hasan, 1976). 16 indices have been measured with the aim of capturing the cohesion provided by conjunctive relations: $X17$ Average commas per sentence, $X18$ Measures on how single comma measures deviate from the central word mean tendency per sentence, $X19$ A rule based approach to the adequate use of the comma, $X20$ Amount of commas, $X21$ Number of connectives, $X22$ Number of additives, $X23$ Additive type-token ratio, $X24$ Number of quantifiers, $X25$ Connector type-token ratio, $X26$ Number of adversatives, $X27$ Adversative type-token ratio, $X28$ Number of distributive connectors, $X29$ Distributive tokens between connective tokens, $X30$ Connective tokens between word tokens, $X31$ Number of types of connectors and $X32$ Connective tokens times connector variety.

*Verbal cohesion*

Verbal forms in Basque provide important ties in discourse cohesion. Verbs can consist of single words (synthetic) or consist of a participial form and an auxiliary (analytical). Auxiliaries can also be used as the main verb. Participles carry aspectual information whereas auxiliaries convey information about argument, structure, tense and mood. Auxiliaries vary in four different tenses/aspects: present, past, hypothetical and imperative (Hualde & Ortiz de Urbina, 2003). Ties provided by verbal forms are measured by a total of 13 indices: $X33$ Average number of words before verb, $X34$ Measures on how single measures of word occurrences before verb deviate from the central tendency, $X35$, verbs per sentence, $X36$ verbs per sentence by verb variability, $X37$ Number of Verbs, $X38$ Number of distinct Verbs, $X39$ Verb type/token ratio, $X40$ Number of Transitive Verbs, $X41$ Number of distinct Transitive Verbs, $X42$ Transitive Verb type-token ratio, $X43$ Number of auxiliary verbs, $X44$ Distinct auxiliary verbs and $X45$ Auxiliary verb type/token.

The process from text to cohesion measure implementation starts with: (1) Text splitting and tagging. Next, (2) Texts are automatically analysed using POS (Part Of Speech) tagging with a morphosyntactic analyser (Aduriz et al., 2004) and a dependency parser (Bengoetxea & Gojenola, 2009). (3) Finally, there is a statistical processing to obtain the cohesion measures.

The human and automatic cohesion grades obtained were discretized to be analysed under several Machine Learning classification strategies.

## Results

This Section describes the ML analysis followed in this study.

**Experimental Design** In order to detect relevant cohesion measures (variables), we first describe how a Feature Subset Selection (FSS) can be performed in an automatic way. After applying different FSS approaches, Feature Selection allows to find the relations between the selected cohesion measures (variables) and the global cohesion grade. The relation is measured based on a set of classifiers. Finally, the goodness of the measure is considered based on the obtained grading accuracy.

In addition to the filtered and wrapper variable selections, the classifiers have also been applied to measure the cohesion for the eight most common variables in previous cohesion measures and the combination of all the 45 measures. The next Section, introduces the description of the variable sorting approach taken for this dataset in the filter approach.

It is worth mentioning that all the experiments have been carried out using the Leave One Out validation technique; this implies learning the classifier with all but one example, and then applying the obtained classifier to the example which has been left out. This process is repeated 17 times (once by example) for each classifier and feature set.

Table 1: Cohesion measures' predictive effect sizes

| Type | *Measure* | $f^2$ | $R^2$ | *Sig.* |
|---|---|---|---|---|
| Word variability | X1 | .161 | .078 | .155 |
| | X2 | .154 | .072 | .163 |
| | X3 | .183 | .095 | .131 |
| | ***X4*** | ***.602*** | ***.331*** | ***.012*** |
| | ***X5*** | ***.602*** | ***.331*** | ***.012*** |
| | **X6** | **.162** | **.78** | **.154** |
| | X11 | .028 | .042 | .538 |
| | X12 | .014 | -.056 | .662 |
| | X13 | .046 | -.025 | .438 |
| | ***X14*** | ***.291*** | ***.171*** | ***.063*** |
| Text structure | X7 | .029 | -.4 | .527 |
| | X8 | .094 | .021 | .27 |
| | X9 | .136 | .057 | .189 |
| | X10 | .016 | -.055 | .644 |
| Lexical cohesion | X15 | .237 | .134 | .090 |
| | X16 | .138 | .06 | .184 |
| Conjunction and Connectors | X17 | .007 | -.063 | .750 |
| | X18 | .117 | .042 | .22 |
| | X19 | .121 | .044 | .215 |
| | X20 | .107 | .032 | .242 |
| | X21 | .078 | .007 | .311 |
| | X22 | .057 | -.013 | .385 |
| | X23 | .001 | -.071 | .933 |
| | X24 | .001 | -.07 | .887 |
| | X25 | .097 | .024 | .261 |
| | X26 | .049 | -.021 | .418 |
| | X27 | .008 | -.063 | .737 |
| | X28 | .179 | .091 | .136 |
| | X29 | .077 | .006 | .314 |
| | X30 | .103 | .029 | .248 |
| | X31 | .041 | -.068 | .826 |
| | X32 | .091 | .019 | .276 |
| Verbal cohesion | X33 | .007 | -.064 | .756 |
| | X34 | .003 | -.068 | .831 |
| | X35 | .094 | .021 | .27 |
| | ***X36*** | ***.404*** | ***.237*** | ***.032*** |
| | X37 | .18 | .094 | .134 |
| | X38 | .169 | .084 | .146 |
| | *X39* | *.27* | *.157* | *.072* |
| | X40 | .052 | -.18 | .406 |
| | X41 | .18 | .098 | .127 |
| | X42 | .169 | .084 | .146 |
| | ***X43*** | ***.31*** | ***.183*** | ***.05*** |
| | ***X44*** | ***.291*** | ***.171*** | ***.063*** |
| | ***X45*** | ***.322*** | ***.19*** | ***.05*** |

**Filters** The use of classifiers requires sorting the variables prior to being classified.

In order to perform the experiment and evaluate the adequateness of the new approach, statistical measures have been used to search for the most salient variables for the cohesion problem. The formulas used with this purpose are well-known metrics in Feature Selection and behavioural research methods: *Gain Ratio*, *One Rule*, *Recursive Elimination of Features (RELIEF)*, *Support Vector Machines (SVM)*, *Chi-square* ($\chi^2$), *Principal Component Analysis (PCA)* and *Effect size* ($f^2$). Selected cases are marked in Table 1.

Table 2: Variable ordering obtained for each of the statistical metrics

| *Metric* | $f^2$ | *GR* | *OneR* | *Relief* | *SVM* | $\chi^2$ | *PCA* |
|---|---|---|---|---|---|---|---|
| First | X4 | X15 | X35 | X44 | X41 | X15 | – |
| Second | X5 | X16 | X14 | X35 | X9 | X16 | – |
| Third | X6 | X12 | X36 | X36 | X19 | X12 | – |
| Forth | X14 | X14 | X10 | X14 | X44 | X14 | – |
| Fifth | X36 | X13 | X25 | X8 | X8 | X13 | – |
| Sixth | X43 | X20 | X5 | X9 | X23 | X20 | – |
| Seventh | X44 | X22 | X15 | X43 | X7 | X22 | – |
| Eighth | X45 | X21 | X17 | X4 | X26 | X21 | – |

**Variable selection based on filtering strategies** Variable sorting under the previously described filtering strategies can be found in Table 2; it should be noticed that the PCA approach does not give a ranking among the variables, but a set of polynomials with linear combinations of some features, this is the reason why the PCA column in Table 2 is empty.

The ML experimental phase has been organised in the following way: First, each of the five selected classifiers has been used to measure the impact of the cohesion measures for the set of student summary global cohesion grades. As shown in Table 3, the first variable in the ordering given for each metric was taken into account first. Next, a second variable is included for each metric, and the accuracy obtained with these two variables is tested with all the classifiers. The same process is run including the third, fourth, and so on variables until a decrease in the accuracy is obtained. Results in Table 3 show that the variable number for each filter is different depending on the moment when an error increase appears.

Table 3: Number of errors obtained by each approach for each subset of variables. PCA approach does not use individual variables but a linear combination of some of them.

| *Metric* | *N* | *Variables* | *BN* | *NB* | *K-NN* | *SVM* | *ANN* |
|---|---|---|---|---|---|---|---|
| 6**f2* | 1 | X4 | 8 | 12 | 12 | 10 | 8 |
| | 2 | +X5 | 8 | 12 | 12 | 10 | 8 |
| | 3 | +X14 | 9 | 8 | 11 | 11 | 10 |
| | 4 | +X36 | 9 | 9 | 11 | 11 | 10 |
| | 5 | +X39 | 9 | 9 | 11 | 10 | 11 |
| | 6 | +X43 | 9 | 8 | 12 | 11 | 10 |
| 4**GainR* | 1 | X15 | 7 | 10 | 10 | 8 | 8 |
| | 2 | +X16 | 7 | 13 | 12 | 8 | 8 |
| | 3 | +X12 | 7 | 10 | 12 | 8 | 10 |
| | 4 | +X14 | 9 | 10 | 11 | 11 | 8 |
| 8**OneR* | 1 | X35 | 8 | 6 | 7 | 9 | 7 |
| | 2 | +X14 | 9 | 5 | 7 | 8 | 7 |
| | 3 | +X36 | 9 | 6 | 6 | 6 | 9 |
| | 4 | +X10 | 9 | 6 | 6 | 6 | 9 |
| | 5 | +X25 | 9 | 6 | 6 | 6 | 11 |
| | 6 | +X5 | 9 | 6 | 9 | 6 | 10 |
| | 7 | +X15 | 9 | 8 | 10 | 6 | 12 |
| | 8 | +X17 | 9 | 8 | 10 | 6 | 10 |
| | 9 | +X22 | 9 | 8 | 10 | 6 | 13 |
| | 10 | +X23 | 9 | 7 | 10 | 6 | 11 |
| | 11 | +X19 | 9 | 8 | 9 | 7 | 9 |
| 7**Relief* | 1 | X44 | 8 | 5 | 5 | 13 | 5 |
| | 2 | +X35 | 8 | 5 | 8 | 6 | 6 |
| | 3 | +X36 | 9 | 6 | 6 | 6 | 8 |
| | 4 | +X14 | 9 | 6 | 7 | 6 | 8 |
| | 5 | +X8 | 9 | 6 | 6 | 6 | 12 |
| | 6 | +X9 | 9 | 6 | 7 | 6 | 9 |
| | 7 | +X43 | 9 | 7 | 7 | 7 | 9 |
| 6**SVM* | 1 | X41 | 8 | 9 | 8 | 8 | 9 |
| | 2 | +X9 | 9 | 10 | 9 | 7 | 7 |
| | 3 | +X19 | 9 | 7 | 7 | 7 | 11 |
| | 4 | +X44 | 9 | 5 | **4** | 6 | 10 |
| | 5 | +X8 | 9 | 7 | 5 | 7 | 6 |
| | 6 | +X23 | 9 | 8 | 8 | 8 | 9 |
| 4**chi*$^2$ | 1 | X15 | 7 | 10 | 10 | 8 | 8 |
| | 2 | +X16 | 7 | 13 | 12 | 8 | 8 |
| | 3 | +X12 | 7 | 10 | 12 | 8 | 10 |
| | 4 | +X14 | 9 | 10 | 11 | 11 | 8 |
| 8**PCA* | 1 | – | 8 | 10 | 13 | 9 | 12 |
| | 2 | – | 8 | 8 | 8 | 9 | 7 |
| | 3 | – | 8 | 8 | 10 | 9 | 8 |
| | 4 | – | 8 | 7 | 10 | 8 | 9 |
| | 5 | – | 8 | 7 | 13 | 8 | 9 |
| | 6 | – | 8 | 7 | 12 | 9 | 11 |
| | 7 | – | 8 | 7 | 11 | 9 | 9 |
| | 8 | – | 9 | 8 | 11 | 9 | 11 |

The variable subset with the best results in the filter approach is composed of the first four variables selected with the SVM filtering metric for the K-NN classifier. It shows an accuracy of 4 errors and the combination is compound by the next variables: $X41$, *transitive verb types*, **verbal cohesion**.

$X9$, *number of sentences* **text structure**. $X19$ *excessive use of commas* **connectives**. Finally, $X44$ *distinct auxiliary verbs* **verbal cohesion**. The same combination of variables obtains the best result with the NB paradigm.

Table 4: Number of errors obtained by each approach using the wrapper FSS.

| | *BN* | *NB* | *K-NN* | *SVM* | *ANN* |
|---|---|---|---|---|---|
| *Errors* | 7 | 5 | 2 | 6 | 4 |
| *Variables* | X15 | X44 | X11, X44 | X41, X44 | X33, X44 |
| *ALL* | 9 | 9 | 10 | 8 | 12 |
| *Experts* | 9 | 9 | 10 | 5 | 11 |

**Variable selection based on wrapper strategies** The variable subset with the best results in the wrapper approach is composed of the first two variables selected with the K-NN classifier. It shows an accuracy of 2 errors and the combination is compound by the next variables: $X11$ *single word measures deviation from the central word mean tendency* taken from the **Word variability** related variable set, and $X44$ *distinct auxiliary verbs*, taken from the **verbal cohesion** feature set.

**Variable selection based on some previously used cohesion measures** Previous research (see some examples in Section ) has measured similar factors to account for cohesion. We have selected the next factor combination to observe how they account for cohesion.: $X3$ (word types), $X6$ (mean letters per word), $X13$ (type-token ratio), $X15$ (sentence overlap), $X21$ (number of connectives), $X33$ (average number of words before verb), $X35$ (verbs per sentence) and $X36$ (verb variability). The combination of the eight previously studied measures (named EXPERT) has been tested under the different classifiers. It should be noticed that in this case the variables are listed using an ascending index. The reason is that the variable ordering is not known. In other words, there is no previous record of one being more relevant than another.

Results are shown in Table 4. The best approximation is provided by the *SVN* variable combination with an accuracy of 5 errors. The results are almost as accurate as the best option based on FSS filter strategies. However, they are still far from the best measures under the wrapper approach and *K-NN*.

**Using ALL the cohesion measures** Another approach is to use all the available indicators to search for cohesion grades. Here, the ALL variable option tests the 45 variables combination for classification purposes.

As shown in Table 4, the ALL variable option does not show accurate results. The accuracy shown is equal to or greater than 8 errors. Again, from the classification paradigms, *SVN* shows the best results.

## Discussion

The goal of this study was to know which measures best predict global cohesion grades. A total of 45 measures were compared to overall cohesion human grades with no previous record on which one was most relevant. The modelling analysis allows searching for the best modelling approach for Basque cohesion grading.

According to the observed results considering all the available information is not the best option for global cohesion grading decision making. The reason for this is redundancy. The EXPERT approach, which combines the most commonly used cohesion measures, has produced a good approximation under SVM classification. Nonetheless, the use of a wrapper approach and K-NN classifier seems to be the best fit for the Basque case.

The difference with the EXPERT combination is probably due to language grammar specific differences. In terms of the variable combination, the amount of auxiliary verb type ($X44$) is the most recurrent one in the best models. This is probably due to Basque grammar morphology. The Basque auxiliary verb carries a lot of grammatical information. Each auxiliary verb provides information about the subject, the two object forms – direct object and indirect object –, as well as tense and aspect. Therefore, the number of auxiliary verb types probably shows how syntactically connected the discourse is. In addition, there are other measures for text structure, word variability, and verbal cohesion which have also been salient.

The obtained model for global cohesion grading will be used as part of a summary evaluation environment (Zipitria et al., 2008a). In order to gain an overall grade for a summary each overall discourse measure is fed into the grading decision making Bayes net (Zipitria et al., 2008b). But, there still are many questions to be answered. Would results be very different if we had measured cohesion indicators that are not included in this study? Does the Basque language require further language specific analysis to better account for cohesion? Would results be very different in another language? Are there interactions among predictors?

We expect that language morphology might be responsible for language differences for cohesion. In future, we aim to analyse the impact that different languages and their morphology make in terms of results. In addition, some of the obtained results might by tied to the particular language under which the study was run. Future work will look at testing the grading scheme under more languages –e.g. Spanish and English– providing LEA with a multilingual approach.

In addition, searching for a greater scope of cohesion measures might also make differences in the results. More theoretically relevant cohesion features (Halliday & Hasan, 1976; Schiffrin et al., 2001) could be automatically modelled and empirically analysed for Basque (e.g. anaphora resolution, ellipsis, etc). A wider collection of measures and further Natural Language Processing tools could allow more in-depth analysis of discourse cohesion and probably a greater accuracy.

## Acknowledgments

This work is supported by the University of The Basque Country (EHU09/09), the Spanish Ministry of Education (TIN2009-14380), and the Basque Government (IT421-10).

## References

Aduriz, I., Aranzabe, M., Arriola, J., Diaz de Ilarraza, A., Gojenola, K., Oronoz, M., et al. (2004). A cascaded syntactic analyser for Basque. In *Proceedings of computational linguistics and intelligent text processing* (pp. 124–135).

Alonso, L., & Fuentes, M. (2003). Integrating cohesion and coherence for automatic summarization. In *Proceedings of the 11th meeting of the european chapter of the association for computational linguistics (eacl2003)* (pp. 1–8).

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Kluwer Academic Publishers.

Bartlett, F. C. (1932). *Remembering; a studty in experimental and social psychology*. Cambridge University Press.

Bengoetxea, K., & Gojenola, K. (2009). Exploring treebank transformations in dependency parsing. In *Recent advances in natural language processing, ranlp 2009.*

Carenini, G., Ng, R. T., & Zhou, X. (2008). Summarizing emails with conversational cohesion and subjectivity. In *Acl-08: Hlt: Proceedings of the 46th annual meeting of the association for computational linguistics: Human language technologies* (pp. 353–361).

Cassany, D. (1993). *Didáctica de la corrección de lo escrito* (Vol. 108). Spain: Editorial Graó, de IRIF SL. (In Spanish)

Clancey, W. J. (1982). Tutoring rules for guiding a case method dialogue. In D. Sleeman & J. S. Brown (Eds.), *Intelligent tutoring systems* (pp. 201–226). London: Academic press, inc.

Ford, L. (1988). The appraisal of an icai system. In *Artificial intelligence and human learning* (pp. 109–123). London: Chapman and Hall, Ltd.

Garner, R. (1982). Efficient text summarization. costs and benefits. *Journal of Educational Research*, *75*(5), 275–279.

Graesser, A., Person, B., & Harter, D. (2001). Teaching tactics and dialog in autotutor. *International Journal of Artificial Intelligence in Education*, *12*, 257–279.

Graesser, A. C., McNamara, D. S., Lowerse, M. M., & Zhiqiang, C. (2004). Coh-metrix: Analysis of text on cohesion of language. *Behavior Research Methods*, *36*, 193–202.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. Longman.

Hualde, J. A., & Ortiz de Urbina, J. (2003). *A grammar of Basque*. Mouton de Gruyter.

Kintsch, E., Steinhart, D., Stahl, G., & the LSA Research Group. (2000). Developing summarization skills through the use of lsa-based feedback. *Interactive learning environments*, *8*(2), 87–109.

Kintsch, W., Patel, V., & Ericsson, K. (1999). The role of long-term working memory in text comprehension. *Psychologia*, *42*, 186–198.

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, *17*, 21–48.

Robertson, J., & Wiemer-Hastings, P. (2002, June). Feedback on children's stories via multiple interface agents. In *Proceedings of the 6th international conference its* (pp. 923–932).

Schiffrin, D., Tannen, D., & Hamilton, H. E. (Eds.). (2001). *The handbook of discourse analysis*. Blackwell Publishing.

Schulze, K. G., Shelby, R. N., Treacy, D., Wintersgill, M. C., VanLehn, K., & Gertner, A. (2000). Andes: A coached learning environment for classical newtonian physics. *The Journal of Electronic Publishing*, *1*(6).

Siddharthan, A. (2006, June). Syntactic simplification and text cohesion. *Research on Language and Computation*, *4*(1), 77–109.

Todd, R. W., Khongput, S., & Darasawang, P. (2007). Coherence, cohesion and comments on students' academic essays. *Assessing Writing*(12), 10–25.

Vechtomova, O., & Karamuftuoglu, M. (2008). Lexical cohesion and term proximity in document ranking. *Information Processing and Management*, *44*, 1485–1502.

Winkels, R., & Breuker, J. (1989). Discourse planning in intelligent help systems. In C. Frasson & G. Gauthier (Eds.), *Intelligent tutoring systems: At the crossroad of artificial intelligence and education* (pp. 124–139). Norwood, New Jersey: Abblex Publishing Corporation.

Woolf, B. P. (1988). Representing complex knowledge in an intelligent machine tutor. In J. Self (Ed.), *Artificial intelligence and human learning* (pp. 3–28). London: Chapman and Hall, Ltd.

Zinn, C., Moore, J. D., & Core, M. G. (2002, June). A 3-tier planning architecture for managing tutorial dialogue. In S. A. Cerri, G. Gouardéres, & F. Paraguau (Eds.), *Proceedings of the 6th international conference on intelligent tutoring systems its* (pp. 574–584). Biarritz, France and San Sebastian, Spain: Springer-Verlag.

Zipitria, I., Arruarte, A., & Elorriaga, J. A. (2006). Observing lemmatization effect in LSA coherence and comprehension grading of learner summaries. In K. Ashley & M. Ikeda (Eds.), *Proceedings of intelligent tutoring systems.* Jhonghli, Taiwan: Springer.

Zipitria, I., Arruarte, A., & Elorriaga, J. A. (2008b). LEA: Summarization web environment based on human instructors' behaviour. In *Proceedings of 8th international conference of advanced learning technologies* (pp. 564–568).

Zipitria, I., Larrañaga, P., Armañanzas, R., Arruarte, A., & Elorriaga, J. A. (2008a). What is behind a summary-evaluation decision? *Behavior Research Methods*, *40*(2), 597–612.