Generación automática de informes en lenguaje natural en una plataforma de e-learning

A. Ramos-Soto, M. Lama, B. Vázquez-Barreiros, A. Bugarín, M. Mucientes, and S. Barro

Centro de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela
{alejandro.ramos, manuel.lama, borja.vazquez, alberto.bugarin.diz, manuel.mucientes, senen.barro}@usc.es

Resumen En este trabajo describimos el servicio SoftLearn Activity Reporter (SLAR), que genera automáticamente informes textuales en lenguaje natural acerca de la actividad de los estudiantes en la plataforma de aprendizaje virtual SoftLearn. Mediante esta aproximación mostramos la utilidad de las descripciones lingüísticas en la generación de dichos informes y cómo estos pueden complementar la información que facilitan los cuadros de mando visuales habituales en este tipo de plataformas. Se muestran ejemplos de aplicación sobre datos reales de un curso de grado gestionado a través de la plataforma Softlearn, que evidencian el interés en incluir la generación automática de informes en lenguaje natural en los 'learning analytics dashboards.'

Keywords: Descripciones lingüísticas de datos, Generación de Lenguaje Natural, Sistemas 'Data to text', learning analytics dashboards

1. Introducción

Una de las áreas más activas del ámbito del 'learning analytics' [1] es el desarrollo de interfaces de usuario que permiten a docentes y estudiantes comprender (y optimizar) la actividad que sucede en los cursos gestionados por plataformas de e-learning. En este contexto, han surgido los cuadros de mando o 'learning analytics dashboards' (LAD) [2,3], entendidos como aplicaciones que permiten visualizar de formas muy diversas los datos recabados en un entorno de aprendizaje. Con frecuencia los LAD se orientan hacia contextos de aprendizaje muy específicos y suelen por tanto incluir herramientas de tipo gráfico con propósitos muy diversos, tales como detectar estudiantes aislados [4], comprender los procesos de colaboración entre estudiantes en entornos sociales [5] o visualizar los diferentes indicadores (numéricos o de otro tipo) sobre los que se evalúa el progreso y el desempeño de los estudiantes [6].

La práctica totalidad de los LAD se basan completamente en visualizaciones gráficas que, con frecuencia, no son fácilmente interpretables por los docentes y estudiantes, sobre todo cuando la cantidad de datos a visualizar es muy elevada (por ejemplo, interacciones entre estudiantes en entornos colaborativos,

series temporales con gran número de indicadores, ...). En este trabajo proponemos el desarrollo de herramientas y técnicas 'Data to text' (D2T) que generen automáticamente informes textuales en lenguaje natural que contengan la información mas relevante acerca de los datos que se visualizan en los LAD. Los informes en lenguaje natural permiten superar, al menos parcialmente, las dificultades de interpretación indicadas anteriormente y no son, en ningún caso, una alternativa a los LAD, sino una herramienta complementaria que ayuda a los usuarios a comprender mejor lo que están viendo.

El ámbito de investigación de los sistemas D2T ([7,8]) es una parte de los sistemas de Generación de Lenguaje Natural (NLG) que trata acerca de la generación automática de narrativas expresadas en lenguaje natural que presenten a los usuarios la información mas relevante (habitualmente escondida o implícita) existente en los datos. Los modelos y técnicas de computación flexible se utilizan, en este contexto, principalmente en las etapas iniciales del diseño de los sistema D2T (determinación de contenido), con el objetivo de proporcionar una descripción lingüística (lenguaje intermedio) sobre cuya base se construye la narrativa final [9]. Los sistemas D2T se han aplicado en ámbitos de aplicación muy diversos [9], como los pronósticos meteorológicos, la e-salud, economía, entre muchos otros, pero, al menos hasta donde alcanza nuestro conocimiento, las únicas aplicaciones D2T en tecnologías del aprendizaje son [10], para la generación de informes a los estudiantes basados en factores externos que afectan a su rendimiento y [11], orientado hacia la evaluación por parte del docente sobre una actividad de aprendizaje específica. No hay por tanto, referencias de aplicación de sistemas D2T datos generados por la actividad docente en entornos de aprendizaje virtual, como el servicio SoftLearn Activity Reporter (SLAR) [12] que describimos en este trabajo. SLAR se ha integrado en la plataforma SoftLearn [13] (plataforma basada en minería de procesos que facilita la evaluación de los estudiantes por parte de los docentes) para extraer la información relevante de los datos y producir descripciones lingüísticas intermedias, utilizando conjuntos difusos, variables lingüísticas, y referencias temporales que finalmente dan lugar a informes en lenguaje natural acerca de la actividad de los estudiantes. SLAR ha sido probado con datos de 72 estudiantes que cursan la asignatura 'Tecnología Educativa' del grado en Pedagogía de la Universidad de Santiago de Compostela.

El trabajo se ha organizado de la siguiente manera: la sección 2 describe el servicio SLAR para la generación automática de informes; la sección 3 presenta diversos ejemplos reales de dichos informes; la sección 4 resume las contribuciones más relevantes del trabajo.

2. Informes textuales automáticos en SoftLearn

SoftLearn [13] [12] es una plataforma de evaluación que opera como uno de los servicios de 'learning analytics' de una arquitectura 'big data' que captura, almacena y pone a disposición en tiempo real los datos producidos por la actividad de los estudiantes de una asignatura. La arquitectura incluye sensores de las actividades de aprendizaje, que capturan los eventos relevantes producidos

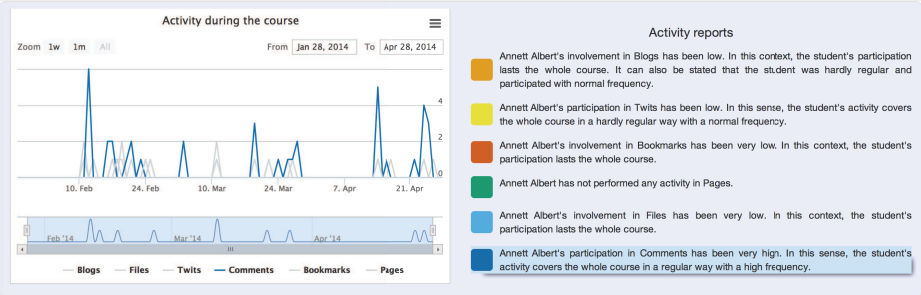

Figura 1. Vista general del cuadro de mandos de SoftLearn para docentes, accesible en [12]: información visual e informes textuales acerca de la actividad de un estudiante.

por la participación de los estudiantes en las actividades del curso. Los datos se almacenan en una base de datos orientada a grafos mediante interfaces Experience API y son enviados a los servicios de análisis de la arquitectura, que se encargan de procesar los datos para extraer información valiosa sobre los procesos de aprendizaje. SoftLearn permite a los docentes evaluar el rendimiento de los estudiantes, les proporciona información sobre el proceso de aprendizaje y sobre su comportamiento y actividad durante el curso. La interfaz gráfica de usuario de SoftLearn [13] permite a los docentes (i) comprender el comportamiento de los estudiantes mediante la visualización de las 'rutas de aprendizaje' que han seguido en el proceso de aprendizaje y (ii) también facilita la evaluación de las actividades de aprendizaje realizadas por los estudiantes en la asignatura. Respecto a estas últimas, el cuadro de mandos (LAD) proporciona de forma gráfica diferentes indicadores estadísticos aobre los estudiantes y su actividad en los diferentes elementos del portfolio del curso, como son las entradas realizadas en blogs, los comentarios publicados, bookmarks seleccionados, 'twits' emitidos o ficheros publicados (Fig. 1). [1]

La funcionalidad del LAD se ha extendido y mejorado con la inclusión del servicio SLAR, que proporciona informes en lenguaje natural generados automáticamente a partir de los datos de cada una de las actividades del portfolio de cada estudiante. Estos informes permiten al docente hacerse una mejor idea de la actividad del estudiante (que también puede visualizarse con gráficas temporales), lo que a su vez facilita la evaluación de los estudiantes mediante rúbricas basadas en el proceso de aprendizaje, su interacción social, motivación, colaboración con otros estudiantes y contenidos publicados. SLAR se basa en técnicas borrosas para la descripción lingüística de datos (LDD) y en técnicas data-to-text (D2T) para la generación de lenguaje natural. En particular, sigue una aproximación similar a la de GALiWeather [14], un servicio de generación de pronósticos meteorológicos en lenguaje natural.

[1] En la figura se muestra el funcionamiento real de SoftLearn (en su versión en inglés, única disponible por el momento), accesible en [12], con datos y alumnos reales, cuyos nombres se han anonimizado al objeto de mantener su privacidad.

2.1. Arquitectura del Servicio SLAR

Nuestra aproximación sigue un proceso 'pipeline' de dos etapas (Fig. 2), en la primera de las cuales se extrae la información relevante de la actividad de cada estudiante, en forma de términos lingüísticos independientes del idioma, denominadas descripciones lingüísticas intermedias [14]. Esta información básica sirve como entrada a la segunda etapa, en la cual se generan los informes textuales finales utilizando plantillas de lenguaje natural.

SoftLearn distingue diferentes elementos del portfolio de cada estudiantes. Estos incluyen blogs, ficheros, twits, comentarios, bookmarks y páginas. Para cada uno de ellos, se registra diariamente el nivel de actividad de cada estudiante. Como consecuencia, el docente puede visualizar la implicación de los estudiantes en la asignatura utlizando el cuadro de mandos de SoftLearn. Por ejempo, la Fig. 1 muestra la actividad real de un estudiante en el elemento 'Comentarios' durante todo el curso.

El sistema de generación de informes proporciona información textual de los elementos de cada portfolio individual a partir de los datos asociados a cada actividad.

Primera etapa: método de generación de descripciones lingüísticas (LDD) La primera etapa de nuestra solución obtiene una descripción lingüística intermedia de la actividad de cada estudiante referida a los elementos de su portfolio en una ventana temporal: semestre, curso completo, ... Cada descripción es un conjunto de etiquetas lingüísticas y datos relevantes extraidos de las series temporales de datos de cada actividad del estudiante (Fig. 1) sobre diversas características relevantes:

- *Nivel de participación.* Proporciona información sobre la participación absoluta del estudiante.
- *Regularidad.* Proporciona información sobre la regularidad de la participación en las actividades de cada estudiante, es decir, cuanto tiempo se desvía del promedio la duración de sus períodos de inactividad.

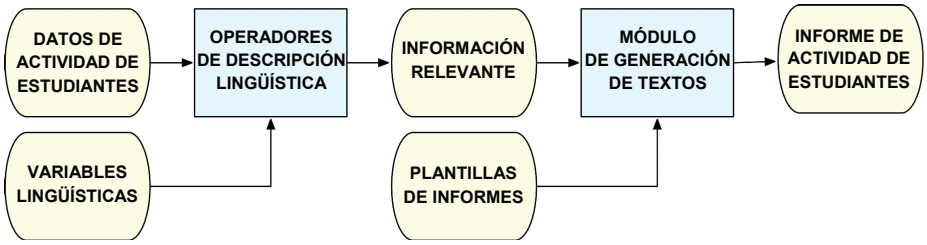

Figura 2. Esquema general del servicio SLAR (SoftLearn Activity Reporter) para la generación de informes en lenguaje natural.

- *Frecuencia*. Proporciona información sobre cómo de frecuente es un/a estudiante en su actividad. Cuando menos tiempo transcurre entre las actividades registradas mayor es su frecuencia de actividad.
- *Alcance del tiempo de actividad*. Proporciona información sobre la ventana temporal de actividad del estudiante en el periodo del curso: los instantes temporales de inicio y fin de su participación.
- *Períodos de inactividad*. Proporciona información sobre la duración del período de mayor inactividad.
- *Impacto*. Proporciona información sobre el impacto de la actividad del estudiante a partir del número de 'likes' y comentarios recibidos de otros estudiantes como resultado de su actividad.

Para cada una de estas características se han definido un conjunto de términos o etiquetas que categorizan las diferentes posibilidades que interesa considerar y se han implementado los operadores asociados que seleccionan la etiqueta lingüística mas adecuada para cada característica. La Tabla 1 muestra las etiquetas consideradas para describir cada característica, que pueden definirse y configurarse independientemente para cada elemento del portfolio (por ejemplo, la definición *HIGH* referida a la participación en 'Comentarios' puede ser diferente de la definición en 'Blogs', ya que la participación en esta última suele requerir un mayor esfuerzo por parte del estudiante. La Figura 3 muestra un ejemplo de la partición borrosa utilizada en la evaluación del indicador 'Alcance del tiempo de actividad.'

Los operadores de características reciben como entrada los datos de actividad y los correspondientes conjuntos de etiquetas (excepto el operador 'Inactividad', que no utiliza etiquetas) y realizan diversas operaciones de evaluación del grado de cumplimiento de las etiquetas respecto de los datos, para determinar cual de ellas describe mejor los datos de entrada originales. En la Tabla 2 se muestra una descripción lingüística resultante de este proceso, que contiene la información relevante relativa a los datos de actividad mostrados en la Fig. 1.

Tabla 1. Etiquetas definidas para cada uno de los indicadores de SoftLearn.

Característica	Etiquetas
Nivel de participación	*VERY LOW, LOW, NORMAL, HIGH, VERY HIGH*
Regularidad	*STRICTLY REGULAR, REGULAR, HARDLY REGULAR, IRREGULAR, VERY IRREGULAR*
Frecuencia	*VERY LOW, LOW, NORMAL, HIGH, VERY HIGH*
Alcance del tiempo de actividad	*BEGINNING, HALF, END* (del período)
Inactividad	Numérico
Impacto	*VERY LOW, LOW, NORMAL, HIGH, VERY HIGH*

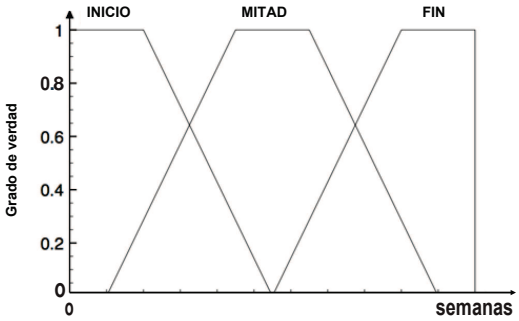

Figura 3. Partición temporal utilizada en el indicador 'Alcance del tiempo de actividad' descrito en la Tabla 1

Segunda etapa: Generación de Lenguaje Natural (NLG) La etapa de Generación de Lenguaje Natural de esta aplicación consiste en un módulo dividido en tres componentes lógicos que, desde una perspectiva global, reciben las descripciones lingüísticas intermedias (Tabla 2) y generan los informes textuales. El primer componente trata con la existencia y el nivel de actividad del estudiante proporcionado por el operador de la característica 'Nivel de participación'. La segunda proporciona información adicional sobre esta actividad, a través de la agregación de la información proporcionada por los operadores de características 'Regularidad', 'Frecuencia', 'Alcance del tiempo de actividad' e 'Inactividad' . Por último, el tercer componente genera los informes acerca del impacto de la actividad del estudiante sobre otros estudiantes.

Los informes producidos por estos tres componentes se integran en un único informe que es el finalmente mostrado en el cuadro de mandos de SoftLearn. La Figura 4 amplía el esquema de la Fig. 2, para mostrar una descripción gráfica mas detallada sobre las relaciones entre los componentes de las etapas LDD y NLG.

Tabla 2. Ejemplo de una descripción lingüística para los datos de actividad de la Fig. 1.

Nivel de participación	VERY HIGH
Regularidad	REGULAR
Frecuencia	HIGH
Alcance del tiempo de actividad	BEGINNING - END
Inactividad	16
Impacto	LOW

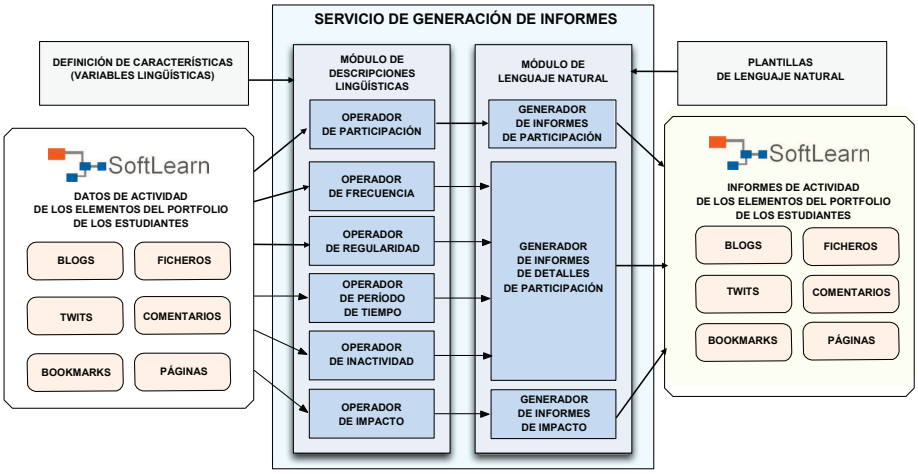

Figura 4. Correspondencia de los diferentes módulos de las etapas de descripción lingüística y generación de informes.

Hemos definido plantillas de lenguaje específicas (ficheros de texto estructurado) que contienen las frase genéricas en lenguaje natural para cada componente. Dichas plantillas son cargadas por el servicio y proporcionadas a su correspondiente componente NLG. Utilizando la información contenida en las descripciones lingüísticas intermedias, cada componente NLG identifica y mapea los términos de la descripción en expresiones en lenguaje natural.

Esta tecnología nos permite proporcionar informes en diferentes idiomas sin mas que cambiar las plantillas de salida, a la vez que incluir diferentes plantillas alternativas para proporcionar diferentes formas de expresar la misma información, lo cual resulta de ayuda para evitar redundancias cuando se generan varios informes a la vez.

3. Ejemplos de informes

Hemos aplicado el servicio SLAR sobre datos reales anonimizados extraídos de la participación de 72 estudiantes en las actividades de la asignatura 'Tecnología Educativa' del Grado en Pedagogía de la Facultad de Ciencias de la Educación de la Universidad de Santiago de Compostela. Esta asignatura se desarrolló en un modo de aprendizaje semipresencial con actividades virtuales que los estudiantes realizaron a través de un e-portfolio social con blogs, micro-blogging, marcas de favoritos, páginas web, etc. Específicamente, los datos registraron diariamente el número de veces que cada estudiante realizaba una actividad en el elemento del portfolio 'Comentarios', así como el número de comentarios y 'likes' que obtenía de otros estudiantes. De entre el elevado número de situaciones diferentes a que da lugar la diversidad de datos, presentamos en esta sección tres ejemplos de informes de estudiantes con diversos patrones de actividad. Todos

los datos mostrados son reales, aunque los nombres de los estudiantes se han anonimizado para mantener su privacidad. La plataforma se encuentra plenamente accesible (en su versión en inglés, única desarrollada por el momento) en el demostrador [12], siendo posible generar y visualizar directamente en ella los informes individualizados reales de la actividad de los seis elementos del portfolio de cada uno de los estudiantes.

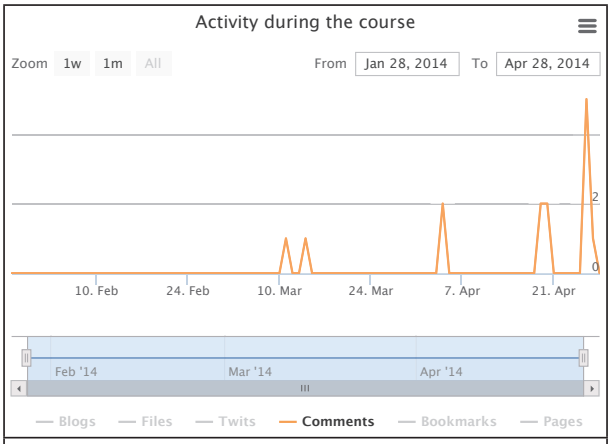

Figura 5. Ejemplo de informe sobre un/a estudiante de perfil algo inactivo.

Por ejemplo, la Fig. 5 muestra un informe de un/a estudiante con una baja actividad, que comienza hacia la mitad del curso, en el que se destaca un período importante de inactividad. Las Figuras 6 y 7 muestran informes sobre estudiantes con un comportamiento opuesto a los dos ejemplos anteriores. Aunque en una primera inspección los gráficos de actividad dan a entender que los patrones de actividad son visualmente similares en ambos casos, existen diferencias importantes entre ellos. La participación en el caso de la Fig. 6 es normal, mientras que la actividad del caso de la Fig. 7 es muy alta. La incoherencia aparente se explica si observamos que las escalas de las figuras son diferentes, pudiéndose apreciar que el estudiante con participación normal interviene un máximo de dos veces por día, mientras que en el otro caso se alcanzan a menudo participaciones de hasta cuatro actividades por día de participación.

En este sentido, y de manera especial en estos últimos ejemplos, vemos cómo los informes textuales proporcionan una forma coherente de generar información

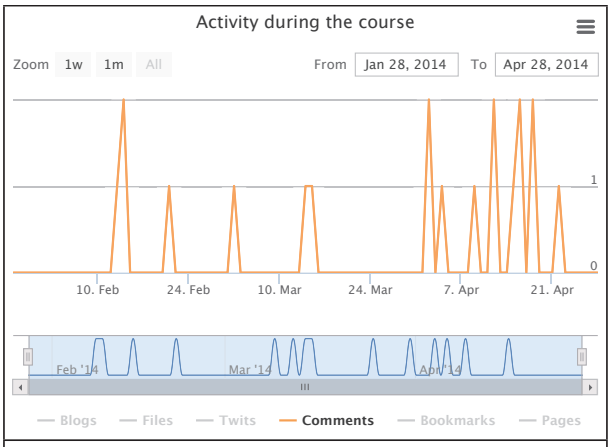

Figura 6. Ejemplo de informe sobre un/a estudiante activo.

objetiva que puede complementar los cuadros de mando visuales y ayudar a los docentes a percibir y valorar de una manera mas directa (lenguaje natural) el comportamiento de los estudiantes.

4. Conclusiones

Hemos descrito el servicio SoftLearn Activity Reporter (SLAR) [12] que genera automáticamente informes textuales acerca de los estudiantes que participan en las actividades de aprendizaje en entornos virtuales. SLAR combina técnicas de descripción lingüística de datos con una aproximación de Generación de lenguaje natural basada en plantillas. SLAR se ha integrado en la plataforma SoftLearn para complementar y mejorar la información proporcionada por su cuadro de mandos visual con información textual que ayude a docentes y estudiantes a comprender el comportamiento de estos durante el proceso de aprendizaje. Hemos probado nuestra solución con los datos reales generados por 72 estudiantes de la asignatura 'Tecnología Educativa' del Grado en Pedagogía de la Universidad de Santiago de Compostela.

Agradecimientos Esta investigación ha sido financiada por el Ministerio de Economía y Competitividad (proyecto TIN2014-56633-C3-1-R, cofinanciado por el Programa FEDER) y por la Xunta de Galicia (proyectos EM2014/012 y CN2012/151, cofinanciados por el Programa FEDER). A. Ramos-Soto está fi-

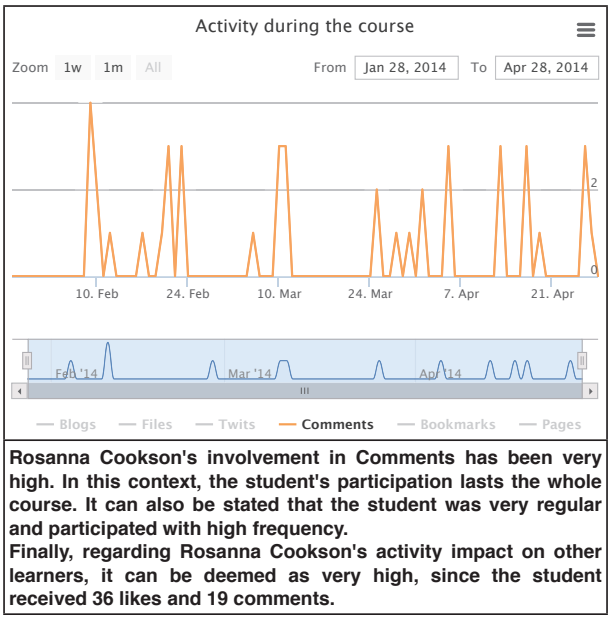

Figura 7. Ejemplo de informe sobre un/a estudiante muy activo.

nanciado por el Ministerio de Economía y Competitividad mediante el programa de contratos predoctorales FPI.

Referencias

1. Siemens, G., Gasevic, D.: Guest editorial - learning and knowledge analytics. Educational Technology & Society **15**(3) (2012) 1–2
2. Verbert, K., Duval, E., Klerkx, J., Govaerts, S., Santos, J.L.: Learning analytics dashboard applications. American Behavioral Scientist **57**(10) (2013) 1500–1509
3. Verbert, K., Govaerts, S., Duval, E., Santos, J.L., Assche, F.V., Parra, G., Klerkx, J.: Learning dashboards: an overview and future research opportunities. Personal and Ubiquitous Computing **18**(6) (2014) 1499–1514
4. Dawson, S., Bakharia, A., Heathcote, E.: SNAPP: Realising the affordances of realtime sna within networked learning environments. In Dirckinck-Holmfeld, L., Hodgson, V., Jones, C., de Laat, M., David, D.M., Ryberg, T., eds.: 7th International Conference on Networked Learning, Lancaster, Lancaster University (2010) 125–133
5. Maldonado, R.M., Kay, J., Yacef, K., Schwendimann, B.: An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In Cerri, S.A., William, J.C., Papadourakis, G., Panourgia, K., eds.: 11th International Conference on Intelligent Tutoring Systems (ITS 2012). Volume 7315 of Lecture Notes in Computer Science., Springer (2012) 482–492
6. Govaerts, S., Verbert, K., Duval, E., Pardo, A.: The student activity meter for awareness and self-reflection. In Konstan, J.A., Chi, E.H., Höök, K., eds.: Con-

ference on Human Factors in Computing Systems CHI 2012, Extended Abstracts Volume, ACM (2012) 869–884
7. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press (2000)
8. Reiter, E.: An architecture for data-to-text systems. In Busemann, S., ed.: Proceedings of the 11th European Workshop on Natural Language Generation. 97–104
9. Ramos-Soto, A., Bugarín, A., Barro, S.: On the role of linguistic descriptions of data in the building of natural language generation systems. Fuzzy Sets and Systems. Accepted. (2015)
10. Gkatzia, D., Hastie, H., Janarthanam, S., Lemon, O.: Generating Student Feedback from Time-Series Data Using Reinforcement Learning. In: Proceedings of the 14th European Workshop on Natural Language Generation. Association for Computational Linguistics (2013) 115–124
11. Sánchez-Torrubia, M., Torres-Blanc, C., Trivino, G.: An approach to automatic learning assessment based on the computational theory of perceptions. Expert Systems with Applications **39**(15) (2012) 12177–12191
12. Vázquez-Barreiros, B.: Softlearn demo, http://tec.citius.usc.es/softlearn/ (Last visited October15th 2015)
13. Vázquez-Barreiros, B., Lama, M., Mucientes, M., Vidal, J.C.: Softlearn: A process mining platform for the discovery of learning paths. In: 14th International Conference on Advanced Learning Technologies (ICALT 2014), IEEE Press (2014) 373–375
14. Ramos-Soto, A., Bugarín, A., Barro, S., Taboada, J.: Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. IEEE Transactions on Fuzzy Systems **23**(1) (2015) 44–57