ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

Andrea
Cascallar Fuentes

Tese de doutoramento

Fuzzy Quantified Protoforms for Data-To-Text Systems: a new model with applications

Santiago de Compostela, 2022

Programa de Doutoramento en Investigación en Tecnoloxías da Información



TESE DE DOUTORAMENTO

FUZZY QUANTIFIED PROTOFORMS FOR DATA-TO-TEXT SYSTEMS: A NEW MODEL WITH APPLICATIONS

Andrea Cascallar Fuentes

ESCOLA DE DOUTORAMENTO INTERNACIONAL DA UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PROGRAMA DE DOUTORAMENTO EN INVESTIGACIÓN EN TECNOLOXÍAS DA INFORMACIÓN

SANTIAGO DE COMPOSTELA
2022

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

DECLARACIÓN DO AUTOR/A DA TESE

D./Dna. Andrea Cascallar Fuentes

Título da tese: Fuzzy Quantified Protoforms for Data-To-Text Systems: a new model with applications

Presento a miña tese, seguindo o procedemento axeitado ao Regulamento, e declaro que:

1. *A tese abarca os resultados da elaboración do meu traballo.*
2. *De ser o caso, na tese faise referencia ás colaboracións que tivo este traballo.*
3. *Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.*
4. *A tese é a versión definitiva presentada para a súa defensa e coincide a versión impresa coa presentada en formato electrónico.*

E comprométome a presentar o Compromiso Documental de Supervisión no caso de que o orixinal non estea na Escola.

En Santiago de Compostela, 23 de Febreiro de 2022

Asdo. Andrea Cascallar Fuentes

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AUTORIZACIÓN DO DIRECTOR/TITOR DA TESE

Fuzzy Quantified Protoforms for Data-To-Text Systems: a new model with applications

Dr. Alejandro Ramos Soto

Dr. Alberto J. Bugarín Diz

INFORMAN:

Que a presente tese correspóndese co traballo realizado por D/Dna. ***Andrea Cascallar Fuentes****, baixo a miña dirección/titorización, e autorizo a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como director desta non incorre nas causas de abstención establecidas na Lei 40/2015.*

De acordo co indicado no Regulamento de Estudos de Doutoramento, declara tamén que a presente tese de doutoramento é idónea para ser defendida en base á modalidade de Monográfica con reproducción de publicacións, nos que a participación do/a doutorando/a foi decisiva para a súa elaboración e as publicacións se axustan ao Plan de Investigación.

En Santiago de Compostela, 23 de Febreiro de 2022

Asdo. Alejandro Ramos Soto
Director/a tese

Asdo. Alberto J. Bugarín Diz
Director/a tese

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

A mi padre

Mi mayor inspiración, siempre

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

There are two means of refuge from the miseries of life: music and cats.

Albert Schweitzer

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecimientos

En primer lugar, quiero expresar mi agradecimiento a mis directores de tesis, Alejandro Ramos Soto y Alberto Bugarín Diz, por el apoyo brindado durante estos años y la confianza depositada en mí.

También quisiera agradecer al Departamento de Electrónica y Computación, así como al Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS), de la Universidade de Santaigo de Compostela, por el gran ambiente de trabajo y los recursos proporcionados que hicieron posible el desarrollo de esta tesis. Gracias también al Prof. Paulo Félix, director adjunto del CiTIUS durante gran parte de mi tesis por su apoyo y por hacer del centro un lugar agradable en el que investigar, como a la dirección actual del centro. Aprovecho también para dar gracias al personal de apoyo de CiTIUS por su disponibilidad y ayuda, facilitando la realización de mi trabajo en el centro.

Quiero aprovechar para agradecer a MeteoGalicia por su apoyo en los proyectos llevados a cabo en colaboración, en concreto a Juan Taboada y Anthony Saunders, cuya ayuda ha sido indispensable durante el desarrollo de esta tesis. También a Javier Gallego y a Carlos Heble, con quien ha sido un placer trabajar juntos. Y por último, agradecer a Carlos Peña, cardiólogo en el Hospital Universitario de Santiago de Compostela, por su ayuda imprescindible.

Por supuesto, también me gustaría agradecer a todas las entidades que han hecho posible el desarrollo de esta tesis con su financiación. En primer lugar, a la Consellería de Cultura, Educación e Ordenación Universitaria (ref. R2016/011), así como a la Axencia Galega de Innovación (ref. IN855A 2018/42). También a la Agencia Estatal de Investigación (ref. TIN2017-84796-C2-1-R), así como a MINECO-Plan Estatal (ref. TIN2015-73566-JIN) y a la empresa Balidea Consulting and Programming S.L (ref. 2016-CE078, 2018-CE097).

En lo personal, no puedo olvidarme de mis compañeros y compañeras del laboratorio LS02 que han hecho que esta experiencia sea inolvidable, tanto a los que me acogieron cuan-

USC UNIDADE

do llegué como a los que se fueron incorporando, haciendo de él un sitio único (ahora en nuestros recuerdos). En especial agradecer a Adrián, David y Víctor, por tantísimos momentos compartidos (y los que nos quedan).

Agradecer a mis amigos, tanto a los de siempre como a los de la carrera, porque siempre puedo contar con ellos y ni el tiempo ni la distancia han cambiado nuestra amistad. En especial a Uxía y María, por estar siempre unidas todos estos años. También agradecerle a Gabriel su apoyo y confianza en mí.

Por último, y lo más importante, quiero agradecer profundamente a mi familia. En primer lugar, a mis padres, que se han esforzado para que yo pudiera elegir mi camino y me han apoyado siempre. A mi hermana, mi abuela Mercedes, mis padrinos y mis primas María y Silvia, por estar unidos en los buenos y los malos momentos, no puedo estar más agradecida por mi familia - *somos los que somos*.

Febreiro de 2022

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Resumo

Hoxe en día, xérase unha gran cantidade de datos cada segundo, tanto por parte de organizacións como usuarios. Nos últimos anos, desenvolvéronse diversas técnicas para almacenar estes datos co obxectivo de posibilitar e facilitar a súa dispoñibilidade. Dispoñer destes datos é moi importante pola información que conteñen, que pode ser de grande utilidade en moitos ámbitos, por exemplo, como soporte nos procesos de toma de decisións.

Para poder empregar esta información e, polo tanto, que resulte de utilidade, débese realizar un proceso de extracción da información a partir dos datos. Porén, a análise e a interpretación destes datos en ambientes profesionais estase volvendo unha tarefa moi difícil debido a que a cantidade de datos que se deben procesar medra de xeito exponencial, excedendo as capacidades dos expertos. Nesas circunstancias, xorde unha alta demanda de métodos e sistemas computacionais que poidan realizar estas tarefas. Tradicionalmente, estas tarefas lévanse a cabo mediante ferramentas de visualización e técnicas estatísticas, sen embargo, ofrecen unha información moi básica e xeral sobre a información contida nos datos que resulta insuficiente xa que, en moitos campos, precísase obter a información detallada oculta nos datos.

Neste contexto, xorden ferramentas que permiten a comunicación dos resultados obtidos na análise de datos dun xeito comprensible por humanos. No eido da Intelixencia Artificial, o campo de Xeración de Linguaxe Natural permite a xeración de textos comprensibles a partir de diversas fontes de datos [192]. En concreto, dentro desta área, os sistemas data-to-text son unha tecnoloxía emerxente de indubidable utilidade, capaces de procesar grandes conxuntos de datos numéricos ou simbólicos e convertelos en textos lexibles que conteñen información relevante e comprensible por humanos.

Por outra banda, relacionado co campo da Xeración de Linguaxe Natural, no campo da Lóxica Borrosa propuxéronse diversas aproximacións co obxectivo de xerar información lingüística significativa como soporte ao manexo da incerteza, coñecida como resumos lingüísticos

USC UNIVERSIDADE

empregando termos lingüísticos. O obxectivo destas ferramentas é resumir de xeito lingüístico os valores de variables empregando o concepto de protoforma introducido por Zadeh [261], en concreto en forma de descricións cuantificadas borrosas, que representan información cuantitativa sobre a cantidade de elementos que cómpren unha determinada propiedade de interese. As sentenzas cuantificadas borrosas son unha ferramenta versátil e moi útil para modelar expresións en linguaxe natural, empregadas en moitas areas diferentes. Na bibliografía, estas propostas están principalmente enfocadas a construír resumos lingüísticos a partir de diversas fontes de datos, como por exemplo, no ámbito da medicina [175], catalogación de textos [265], series temporais [36, 71, 111, 112, 114] ou toma de decisións con múltiples criterios [46, 115, 244, 246].

Na linguaxe humana empréganse cuantificadores lingüísticos acotío, sendo unha ferramenta moi potente para representar e describir o coñecemento sobre a cantidade de elementos que cómpren unha determinada propiedade [51]. Existe unha gran variedade de cuantificadores na linguaxe, sendo os absolutos (“dous ou máis”, “algúns”,...) e os relativos (“a metade”, “a maioría”,...) os máis empregados. Na bibliografía existen diversas propostas baseadas en estruturas de descricións cuantificadas borrosas, porén, a maioría das aproximacións existentes están baseadas en sentenzas cuantificadas borrosas unarias (“Case todos os traballadores teñen un bo salario”) e binarias (“Case todos os traballadores novos teñen un bo salario”) [117, 122, 145, 208, 268].

Obter o conxunto de descricións lingüísticas máis axeitado a partir dun conxunto de datos nun escenario concreto pode considerarse un problema de busca onde as sentenzas candidatas deben ser xeradas e avaliadas con respecto a o conxunto de datos correspondente. Sen embargo, en moitas aplicacións as características do problema (tamaño do conxunto de datos, cantidade de posibles descricións a xerar,...) fan imposible abordalos xerando todas as posibles sentenzas. Polo tanto, algúns autores propuxeron solucións baseadas en algoritmos tanto heurísticos [36, 182, 183] como meta-heurísticos [6, 38, 65] co obxectivo de atopar o conxunto máis relevante de sentenzas de acordo cuns criterios previamente establecidos sen necesidade de explorar todo o espazo de busca, o cal sería moi ineficiente.

Ademais, para obter as sentenzas máis representativas é preciso avalialas con respecto ao conxunto de datos. Esta avaliación ten un gran impacto no comportamento do algoritmo de busca xa que dirixe a súa execución, polo que debe realizarse coidadosamente. O obxectivo de avaliar unha sentencia cuantificada borrosa é obter o grado de verdade no intervalo [0, 1] para un determinado conxunto de datos. Esta medida comprende o cálculo da compatibilidade

USC UNIVERSIDADE

entre o número de elementos no referencial que cómpren a sentenza e o cuantificador presente na sentenza. Polo tanto, este cálculo depende *i)* dos datos, *ii)* da definición dos cuantificadores, *iii)* da definición dos termos lingüísticos definidos a partir das propiedades presentes no referencial e *iv)* do método de cuantificación borrosa empregado para avaliar as descricións.

Aínda que o uso de propostas baseadas en descricións lingüísticas de datos pode ser empregado de xeito directo de forma que a información resultante sexa consumida directamente polos usuarios, estas estruturas poden carecer da expresividade requirida en moitos ámbitos. No entanto, estes elementos poden ser útiles empregándoos como linguaxe intermedia que pode ser abstraída e combinada con outras pezas de información co obxectivo de xerar un texto en linguaxe natural final mediante sistemas de xeración de linguaxe natural que poda ser consumido por usuarios e lles resulte informativo [243, 247, 258, 259, 260]. Neste contexto, as solucións Data-To-Text propostas baseadas en descricións lingüísticas de datos adoitan estar formados por estruturas demasiado básicas que poden resultar insuficientes na súa aplicación noutros ámbitos, especialmente en casos reais. Polo tanto, o obxectivo desta tese é investigar o emprego de técnicas baseadas en Intelixencia Artificial para mellorar os resultados na xeración de descricións lingüísticas de datos no contexto de grandes conxuntos de datos. En concreto, este obxectivo xeral comprende as seguintes metas:

- Estender a fase de determinación de contido en sistemas Data-To-Text, aplicando técnicas da Intelixencia Artificial para a inclusión de dimensións moi relevantes nalgúns ámbitos, por exemplo, a información xeográfica no ámbito da meteoroloxía. Ademais, propoñer tamén solucións baseadas en algoritmos meta-heurísticos para a extracción do conxunto de sentenzas cuantificadas máis representativo a partir de grandes conxuntos de datos.

- Analizar o impacto da selección do método de cuantificación borrosa na avaliación de descricións cuantificadas borrosas, comprobando se se atopan diferenzas significativas nos seus comportamentos que poidan condicionar o sistema no que se empregan ou se, polo contrario, os métodos mostran comportamentos similares e, polo tanto, a súa selección pode ser condicionada por outros aspectos coma o rendemento.

- Deseñar e desenvolver un modelo que comprenda a arquitectura Data-To-Text para describir series temporais, propoñendo estruturas de sentenzas máis complexas que conteñan a información necesaria e poidan resultar de utilidade. Ademais, neste tipo de dato a dimensión temporal é moi importante, polo tanto, neste obxectivo compréndе-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

se o desenvolvemento dunha ontoloxía temporal borrosa que comprenda os conceptos necesarios para manexar a imprecisión inherente da linguaxe.

Esta tese está desenvolvida seguindo o método científico [89], que consiste nun proceso iterativo formado polas seguintes fases:

- Definición de obxectivos: esta primeira etapa, levada a cabo no Capítulo 1, consiste en definir os obxectivos que serán abordados durante o desenvolvemento desta tese. Nesta fase tamén se realiza un estudo en profundidade destes obxectivos coa finalidade de identificar sub-obxectivos que poidan ser cumplidos.

- Estudo dos traballos relacionados: esta fase lévase a cabo no Capítulo 2, realizando un estudo da investigación relacionada para avaliar os modelos e as técnicas xa propostas, coas súas fortalezas e as súas debilidades.

- Desenvolvemento da investigación: nos Capítulos 3, 4 e 5, despois destas dúas fases previas, trabállase no desenvolvemento dos modelos e na súa aplicación en casos de uso reais. Estes casos permiten acceder a datos cos que avaliar os métodos propostos e establecen unha serie de requirimentos que se usan como guía durante este proceso.

- Probas: tamén nestes tres capítulos, a última fase é a realización de probas para validar os métodos propostos, unha vez están desenvolvidos. Estas probas son realizadas con datos reais seleccionados en fase anteriores.

- Aplicación en casos de aplicación reais: nesta fase, os modelos propostos desenvolvidos no Capítulo 5 son aplicados e validados en problemas reais, dentro de proxectos de innovación e transferencia, nos ámbitos da saúde e a información ambiental.

O traballo desenvolvido nesta tese deu lugar a diversas publicacións, entre as que destacan:

- *Empirical Study of Fuzzy Quantification Models for Linguistic Descriptions of Meteorological Data* presentado na *2020 IEEE International Conference of Fuzzy Systems (FUZZ-IEEE)* actualmente clasificado como Tipo A-, Clase 2 no ranking GII-GRIN-SCIE (GGS).

- *An experimental study on the use of fuzzy quantification models for linguistic descriptions of data*, presentado na *24th European Conference of Artificial Intelligence (ECAI)* actualmente clasificado como Tipo A, Clase 2 no ranking GII-GRIN-SCIE (GGS).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- *Automatic generation of textual descriptions in data-to-text systems using a fuzzy temporal ontology: application in air quality index data series*, aceptado na revista Applied Soft Computing en febreiro de 2022. Revista con un factor de impacto 6.725 (JCR 2020).

- *Un sistema para la descripción automática en lenguaje natural de gráficas de sectores: aplicación en datos de calidad del aire* presentado no 2021 XX Congreso Español sobre Tecnologícas y Lógica Fuzzy (ESTYLF).

- *Meta-heuristics for generation of linguistic descriptions of weather data: experimental comparison of two approaches*, aceptado na revista Fuzzy Sets and Systems en febreiro de 2022. Revista con un factor de impacto 3.343 (JCR 2020).

Nesta memoria detállanse as contribucións do traballo desenvolvido e os resultados máis relevantes da experimentación levada a cabo:

En concreto, o Capítulo 2 analiza en profundidade a investigación previa relacionada con esta tese. En primeiro lugar, esta revisión inclúe unha revisión xeral da bibliografía que cubre as propostas dentro do campo da Xeración de Linguaxe Natural, dende as aproximacións clásicas ata as empregadas nas actualidade. Nesta revisión tamén se inclúe un estudo do estado da investigación dentro do campo da Descrición Lingüística de Datos, profundando ademais na aplicación destas técnicas dentro do campo da Xeración de Linguaxe Natural e, en concreto, nos sistemas Data-To-Text para a extracción de información lingüística manexando a incerteza e a imprecisión. Ademais, neste capítulo inclúese unha revisión en profundidade do estado da investigación relacionada con cada uns dos sub-obxectivos propostos nesta tese.

O primeiro obxectivo que abordamos nesta tese foi a xeración de descricións lingüísticas sobre datos meteorolóxicos con propiedades xeográficas. O Capítulo 3 está destinado a describir o traballo realizado con respecto a extracción do conxunto de sentenzas cuantificadas borrosas máis axeitado, centrándonos en comprobar a eficiencia de distintas aproximacións meta-heurísticas para explorar a explosión combinatoria de descricións resultantes de considerar varias propiedades (temperatura, vento, estado do ceo, área xeográfica) simultaneamente.

Para isto, empregáronse dous tipos diferentes de conxuntos de datos meteorolóxicos, datos de observación en tempo real e datos de predicción, facilitados pola Axencia Meteoroloxica de Galicia (MeteoGalicia) [154], que describen a situación meteorolóxica en Galicia respecto a tres variables meteorolóxicas. Neste capítulo, co obxectivo de mellorar a fase de

USC UNIVERSIDADE

extracción do conxunto de descricións máis adecuadas, propuxéronse dúas solucións meta-heurísticas baseadas no algoritmo Simulated Annealing [224] e nun Algoritmo Xenético [73]. En concreto, a experimentación realizada neste capítulo para analizar o comportamento das solucións meta-heurísticas propostas está formada principalmente por dúas fases: *i)* execución das aproximacións meta-heurísticas para os conxuntos de datos para as dúas fontes de datos consideradas obtendo, en cada execución, unha clasificación de descricións, e *ii)* avaliación dos resultados das aproximacións para comparar o seu rendemento na procura da clasificación máis representativa con respecto aos datos.

Para determinar a representatividade das descricións candidatas, tamén no Capítulo 3 propóñense un conxunto de métricas tanto para a avaliación individual de cada candidata con respecto aos datos como a nivel da clasificación resultante de cada algoritmo. Ademais, neste capítulo tamén se inclúe unha avaliación por expertos das estruturas de sentenzas propostas para determinar se son adecuadas para describir este tipo de datos. En xeral, destes dous tipos de avaliación pódese concluír, en xeral, que a solución baseada no Algoritmo Xenético ten un mellor rendemento e, para o caso binario, as solucións xeradas por este algoritmo tamén son mellores en canto son máis adecuadas para describir estes datos.

Con respecto ao segundo obxectivo desta tese, relacionado coa análise do impacto do método de cuantificación na avaliación de sentenzas cuantificadas borrosas, no Capítulo 4 descríbense os detalles do estudo realizado sobre o comportamento de métodos de cuantificación borrosa á hora de avaliar sentenzas cuantificadas borrosas unarias e binarias formadas por cuantificadores relativos e coherentes por parte dun conxunto de seis métodos de cuantificación borrosa moi empregados na investigación relacionada. A selección deste tipo de cuantificadores é un requisito para poder empregar algúns dos métodos de cuantificación deste estudo. En primeiro lugar, neste capítulo analizamos os métodos de cuantificación seleccionados en canto ás propiedades teóricas analizadas na investigación relacionada tanto para o caso da avaliación unaria como binaria.

O primeiro dos experimentos levados a cabo neste capítulo foi un estudo da correlación entre os métodos considerados empregando un conxunto de datos meteorolóxicos, centrándonos en tres variables meteorolóxicas: ceo, vento e temperatura. Este estudo estaba composto de dúas fases: *i)* xeración das descricións lingüísticas, onde se xeran as sentenzas e se avalían empregando os métodos de cuantificación comparados e *ii)* execución de tests estatísticos co fin de analizar a correlación no comportamento entre pares de métodos. Os resultados destes experimentos mostraron unha correlación practicamente absoluta entre todos os pares de

USC UNIVERSIDADE

métodos de cuantificación na avaliacións de sentenzas cuantificadas unarias. No caso da avaliación binaria, os resultados mostran unha correlación máis baixa en xeral, podendo clasificar os métodos de cuantificación en tres grupos segundo mostran unha correlación suficiente no seu comportamento.

Tamén no Capítulo 4 estendeuse o primeiro dos experimentos levando a cabo un estudo estatístico para detectar diferenzas significativas entre pares de métodos de cuantificación. Neste experimento empregáronse 15 conxuntos de datos empregados en tarefas relacionadas coa Intelixencia Artificial xerando variables lingüísticas a partir tanto de variables numéricas como categóricas. Igual que no estudo previo, neste diferenciáronse dúas etapas: *i)* xeración das descricións lingüísticas, onde se xeran as sentenzas e se avalían empregando os métodos de cuantificación comparados e *ii)* execución de tests estatísticos co fin de atopar diferenzas significativas, se existen, no comportamento entre pares de métodos. Os resultados deste experimento probaron que non existen diferenzas significativas entre os métodos de cuantificación considerados, sendo consistente co estudo de correlación. Polo tanto, neste escenario a selección do método de cuantificación podería estar guiado por outros criterios. Porén, na avaliación de descricións cuantificadas binarias, detectáronse diferenzas significativas entre tódolos pares de métodos polo menos no 50 % dos experimentos. Polo tanto, neste caso, a selección do método de cuantificación ten gran impacto na avaliación deste tipo de sentenzas.

Por último, o Capítulo 5 está centrado no terceiro dos obxectivos propostos nesta tese, propoñendo o deseño dun modelo Data-To-Text baseando en descricións lingüísticas de datos co obxectivo de compensar as deficiencias dos modelos propostos na investigación relacionada en canto a que empregan estruturas simples que resultan pouco informativas nalgúns escenarios, como casos de aplicación reais. A estrutura deste modelo Data-To-Text está baseada na arquitectura proposta por Reiter [188, 192, 193], aceptada como a arquitectura clásica estándar. A arquitectura proposta está formada polas seguintes fases: *i)* determinación de contido, *ii)* planificación do documento, *iii)* micro-planificación e *iv)* realización lingüística.

Ademais, neste modelo pretendeuse incluír ferramentas para o manexo da dimensión temporal, presente en moitos tipos de datos e de gran importancia en moitos casos reais. En concreto, levouse a cabo o deseño dunha ontoloxía temporal que inclúe a estrutura deste modelo temporal e as definicións dos conceptos incluídos para manexar esta dimensión.

Como aplicación real deste modelo proposto, desenvolvemos dous sistemas Data-To-Text en dous ámbitos reais. O primeiro dos sistemas, no eido da información ambiental, é ICA2TEXT, desenvolvido en colaboración con expertos da Rede Calidade do Aire de Gali-

cia (MeteoGalicia [154]) para a descrición de series temporais de datos de calidade do aire. O índice de calidade do aire é un indicador utilizado por axencias meteorolóxicas de todo o mundo para medir a presenza e densidade dun conxunto de contaminantes na atmosfera. Esta información é moi importante tanto para expertos como para os cidadáns, xa que ten un gran impacto na saúde das persoas. Polo tanto, despois da fase de validación por expertos pasou a unha fase de probas, previa ao seu despregamento na páxina oficial de MeteoGalicia.

O segundo caso de aplicación deste modelo é no campo da medicina, en concreto no eido da rehabilitación cardíaca. Neste ámbito desenvolvemos o proxecto SMART CARDIA[1], que recibiu Premio "Conocimiento", no VI Forum RIES Retos Internacionales del Ecosistema Salud del Clúster Saúde de Galicia (CSG) o 10 de Novembro de 2021. O obxectivo deste proxecto é manter un seguimento dos pacientes que se encontren neste programa, por ter sufrido un evento cardíaco.

Dentro deste proxecto, xorde o interese de xerar reportes en linguaxe natural a partir dos indicadores clínicos (colesterol, tensión,...) dos que se realiza un control para falicitar esta tarefa aos profesinoais sanitarios. Deste xeito, a partir do modelo Data-To-Text incluído neste capítulo, desenvolvemos o módulo de reportes para describir automaticamente series de datos a partir de 19 indicadores médicos en colaboración cos especialistas do Servicio de Cardioloxía do Hospital Clínico Universitario de Santiago de Compostela. Debido á natureza deste proxecto, a validación deste sistema centrouse en contrastar a calidade e a corrección dos textos xerados automaticamente mediante a revisión dun conxunto de casos xerados automáticamente por expertos. Malia que o número de expertos dispoñibles para esta avaliación non foi elevado, permitiunos verificar a idoneidade desta proposta.

Finalmente, como resultado do traballo desenvolvido nesta tese de investigación, identificamos algunhas direccións de investigación de interese, en concreto, para mellorar a expresividade de sistemas Data-To-Text baseados en descricións lingüísticas de datos e ampliar a súa aplicabilidade a casos de aplicación reais. Dentro da fase de determinación de contido, con respecto á extracción do conxunto máis representativo de sentenzas cuantificadas, problema que non se abordou en profundidade na investigación; a maioría das propostas existentes están condicionadas polo caso de aplicación e son deseñadas ad hoc. Polo tanto, o traballo futuro nesta liña debería centrarse no estudo de novas solucións propoñendo novas aproximacións que poidan aplicarse de xeito sinxelo en diferentes casos de uso para poder realizar

[1] https://citius.usc.es/investigacion/proxectos/listado/plataforma-intelixente-preditiva-soporte-xestion-clinica-proceso-rehabilitacion

comparativas empregando diversas fontes de datos.

Tamén dentro desta fase dentro da arquitectura Data-To-Text clásica, nesta tese estudamos a avaliación das sentenzas cuantificadas borrosas, realizada mediante métodos de cuantificación borrosa, demostrando que a selección do método de cuantificación para a avaliación de descricións unarias pode ser guiada por criterios de rendemento, entre outros, xa que non se atoparon diferenzas significativas no seu comportamento. Pola contra, o traballo futuro debería afondar na análise do comportamento dos métodos de cuantificación na avaliación de sentenzas cuantificadas borrosas binarias, xa que neste caso os experimentos levados a cabo nesta tese mostran diferenzas significativas no seu comportamento. Ademais, este estudo podería ser ampliado en diversos aspectos, por exemplo, *i)* engadindo novos métodos de cuantificación máis recentemente propostos na investigación, como F^A [58], que permiten avaliar diversos tipos de cuantificadores como, por exemplo, cuantificadores semi-borrosos; *ii)* realizando unha comparativa con outros tipos de cuantificadores, xa que os estudo desta tese está centrado en cuantificadores relativos e coherentes; ou *iii)* ampliando a experimentación con estruturas de sentenzas máis complexas.

Outro aspecto a ter en conta é que, frente ás aproximacións existentes, a nosa proposta dun modelo de descricións lingüísticas permite o deseño de estruturas máis complexas. Deste xeito, pódense construír resumos lingüísticos máis representativos con respecto aos datos descritos coa finalidade de poder adaptarse a casos de uso reais. Ademais, propuxemos un modelo temporal para manexar a imprecisión existente en diversos tipos de datos. O traballo futuro debería seguir dentro desta liña, investigando novos modelos baseados en estruturas máis complexas e comprendendo dimensións críticas en moitos escenarios.

Por último, avaliar os resultados dun sistema de xeración de linguaxe natural é unha tarefa complicada xa que a dispoñibilidade expertos adoita ser limitada e os métodos de validación automática demostraron non ser suficientes aplicados en solitario, ainda que poden ser útiles como validación complementaria á validación realizada por expertos. Neste sentido, o traballo futuro debería centrarse tamén na investigación de técnicas de avaliación para sistemas Data-To-Text que sexa factible aplicar en escenarios reais cuns resultados fiables.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Contents

Resumo **xiii**

1 Introduction **1**
1.1 Motivation 1
1.2 Hypothesis and objectives 6
1.3 Contributions 7
1.4 Methodology 12
1.5 Dissertation structure 13

2 Related work **15**
2.1 Obtaining the most representative set of descriptions in Linguistic Descriptions of Data (LDD) 27
2.2 Obtaining the truth degree of fuzzy quantified sentences using fuzzy quantified methods 28
2.3 Fuzzy linguistic models for the management of the uncertainty in time series . 34
2.4 General remarks about the application of fuzzy techniques in the Natural Language Generation field 36

3 Using meta-heuristics for searching Linguistic Descriptions of Data **39**
3.1 Linguistic Descriptions of Data in weather observation and forecasting . . . 40
3.2 Modeling of solutions 44
3.3 Proposal of meta-heuristic approaches for LDD extraction 52
3.4 Experimental results 68
3.5 Discussion and final remarks 79

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

4 Empirical study of the behavior of fuzzy quantification methods for linguistic descriptions of data **83**
4.1 Analysis of the considered fuzzy quantification methods 84
4.2 Study of the correlation between fuzzy quantification methods 90
4.3 Statistical significance study between fuzzy quantification methods 97
4.4 Discussion and final remarks 105

5 Fuzzy model for the management of the imprecision in temporal knowledge **107**
5.1 Data-To-Text model for linguistically summarizing temporal data 108
5.2 ICA2TEXT: Linguistic descriptions of Air Quality Index (AQI) time series . 118
5.3 SMART CARDIA: Intelligent management of the cardiac rehabilitation process 128
5.4 Discussion and final remarks 143

6 Conclusions **145**
6.1 Future work 149
6.2 Final remarks 151

Bibliography **155**

List of Figures **183**

List of Tables **187**

List of Acronyms **191**

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER 1

INTRODUCTION

1.1 Motivation

It is well known that nowadays large amounts of data are currently generated and consumed every second by both users and organizations. For instance, in recent years, governments have made big efforts to make a multitude of open content available to users so that it can be consulted and analyzed without restrictions or control mechanisms, known as Open Data Paradigm [3]. Advances in technology mean the storage of these data and have also facilitated their easy availability. These data have real value insofar as they can be transformed into relevant and useful information, which should be conveyed in an effective manner to the information consumers (e.g., decision makers). Getting meaningful insights from these large amounts of data poses several challenges which cannot only be addressed with traditional statistics and graphical visualizations. In general, these tools are very useful for obtaining basic and general information from data, but for users to dig deeper and understand the real relevant information behind the data, it is necessary to employ techniques that are better suited to the specific needs of each domain and can scale up as the data accumulates.

Moreover, some types of data have very important characteristics that require special attention. For instance, the geographic dimension is very relevant in many contexts, such as Geographic Information Systems (GIS) [140], where data contain spatial and geographic information. These data are used in multiple technologies that have made our lives easier in many aspects, presenting them as maps. An example is Google Maps [93], which allows us to search for addresses and places of interest, calculate routes, etc.

On the other side, in many domains, such as medicine or meteorology, data have a critical

time dimension that must be carefully addressed, since an error could seriously harm people's safety. For example, if a mistake is made in the medication pattern of a hospitalized patient or if the behavior of a severe thunderstorm is incorrectly predicted over hours. Therefore, this situation motivates the interest in using other types of complementary techniques that allow to know the characteristics of the data in greater depth in a understandable manner for the users. In this context, the tools that allow to linguistically communicate the results of data analyses in a comprehensible manner to human decision makers are still under heavy development.

A promising Artificial Intelligence (AI) technique for this purpose is Natural Language Generation (NLG), which allows the generation of comprehensible texts with useful information from several sources and modalities of data (mainly numerical and textual) [192]. Within the more general area of the Natural Language Generation (NLG), Data-To-Text (D2T) systems are an emerging technology of undoubted usefulness, which can process large amounts of numerical or symbolical data, converting them into readable texts that contain relevant and understandable information for users. D2T systems include: *i)* a data analysis stage where the relevant information is extracted from data and *ii)* a generation stage where information is conveyed in natural language.

Human language is inherently imprecise, often employing imprecise terms in everyday life that do not have a strict formal definition and are therefore conditioned by many factors, such as the speaker's judgment. An example, speaking of the weather, is a person's perception of temperature, which can be high or low at the same instant for different people. Within the NLG field, the semantics of the terms are not usually modeled, especially if they are imprecise. This is a special disadvantage, since the absence of handling this type of terms causes a loss of naturalness in the discourse and, therefore, it does not totally resemble the way humans express themselves.

In this sense, in the Fuzzy Logic field several approaches were proposed as tools for generating meaningful linguistic information which supports uncertainty management allowing to model the semantic of imprecise terms, known as Linguistic Descriptions of Data (LDD) or linguistic descriptions using linguistic terms. LDD summarize in a linguistic form one or more numerical variables and their values using the general concept of protoform [261] and the more specific of fuzzy quantified statements, which provide quantitative information about the fulfillment of some properties of interest in a numerical data set (e.g., "Many students are tall"). Although the information included in LDD may, sometimes, be directly consumed by users, linguistic descriptions of data may lack in general the expressiveness of

USC UNIVERSIDADE

actual texts. Notwithstanding, these are useful information items, usually expressed in an intermediate language which can be abstracted and combined with other information sources in order to produce (after performing the planning stages) the final natural language narrative which is conveyed to and consumed by the users in NLG systems in general and D2T in particular [243, 247, 258, 259, 260].

The presence and use of linguistic quantifiers in human language is a very powerful tool for representing and describing knowledge about the quantity of elements that fulfill one or more properties [51]. Among the huge variety of determiners we use in language (lexical, proportional, absolute, exception, partitive, ...) [56, 90, 124], most of the attention in the related literature has been paid to absolute quantifiers, which express quantities over the total number of elements which fulfill the properties (e.g. "Two or more", "A few") within a given referential, and to relative quantifiers, which make the counting depending on the total number of elements of the referential (e.g. "A half", "Almost all"). Regarding the number of properties considered in the quantified statements (the n-arity of the quantifier) it usually ranges from one (unary quantifiers) to four (quaternary quantifiers), although literature has mostly focused on unary and binary quantifiers [117, 145, 268].

In general, quantified sentences are a versatile tool for modeling natural language expressions, used in a wide range of areas [51]. Most approaches in the literature are focused on the linguistic summarization of data from many fields such as text categorization [265], eldercare [239]. temporal series [36, 71, 111, 112, 114] or multiple-criteria decision-making [46, 115, 244, 246] where, for instance, fuzzy quantification models were proposed for aggregating the criteria according to their importance [19]. Another fruitful application is fuzzy querying on databases [23, 121], since natural language statements can be modeled by quantified sentences, being also suitable in the information retrieval area.

To create linguistic descriptions, it is necessary to define at least the following components: *i)* fuzzy quantifiers, such as "many", which are defined as fuzzy sets with a fuzzy membership function [258]; *ii)* linguistic variables, defined as fuzzy sets with an associated fuzzy membership function from categorical or numerical variables (see Figure 1.2); *iii)* aggregation operators, which allow to create more complex expressions composed by a combination of linguistic terms, such as "very hot and dry"; and *iv)* evaluation criteria to assess the suitability of a generated description to select the most representative sentences with regard to a data set, such as the truth value. These components allow the construction of fuzzy quantified sentences. Specifically, the unary descriptions are structured as follows:

"*Q Y* are *S*"

For instance, "Almost all workers are well paid" where *Q* is a quantifier (*"almost all"* in the example), *Y* is a referential set (*"workers"* in the example) and *S* a linguistic value of *Y* (*"well paid"* in the example).

Binary descriptions have the following structure:

"*Q KY* are *S*"

For example, "Almost all *young* workers are well paid", where an additional qualifier *K* is included (*"young"* in the example).

Approaches for LDD are mostly based on unary and binary fuzzy quantified sentences [117, 122, 145, 208, 268]. Despite being the most commonly used quantified propositions in the literature, these structures may be too simple when describing information from several domains, so it is necessary to propose different types of more complex quantified sentences to solve this problem. In other domains, several types of protoforms different from quantified propositions were proposed, for instance, in time series description [37, 120, 159], health [5, 268], process management [233], elder care [238] or probabilistic uncertainty consideration [8].

The problem of finding an appropriate LDD on a given data set can be considered a search problem where candidate protoforms need to be generated and assessed on the data set. In many applications, the data set size in conjunction with the number of linguistic elements in the protoform (variables, values, ...) make unfeasible to explore all combinations. Because of this, some authors have proposed approaches based on heuristics [36, 182, 183] or metaheuristics (genetic algorithms) [6, 38, 65] search techniques for finding relevant sentences (according to pre-defined criteria) in those cases where the number of possible descriptions is so large that the search space cannot be fully explored efficiently. Nevertheless, in general, this problem has not received much attention in contexts where the search space and/or the size of the data set becomes too large.

Furthermore, the evaluation of a quantified sentence may influence greatly the performance of the search algorithm, and thus needs to be laid out carefully. The aim of evaluating a quantified statement is to get a truth degree in the real interval [0, 1] for a given data set. From a qualitative perspective, the truth degree measures how well the statement fits the underlying data. This step involves calculating the compatibility between the number of elements in the referential which fulfill the sentence (its cardinality) and the quantifier in the sentence.

USC UNIVERSIDADE

Therefore, this compatibility measure depends on *i)* the data, *ii)* the quantifiers definition (Figure 1.1), *iii* the linguistic terms defined from the properties in the referential (Figure 1.2) and *iv)* the fuzzy quantification method used for evaluating the fuzzy quantified statement.

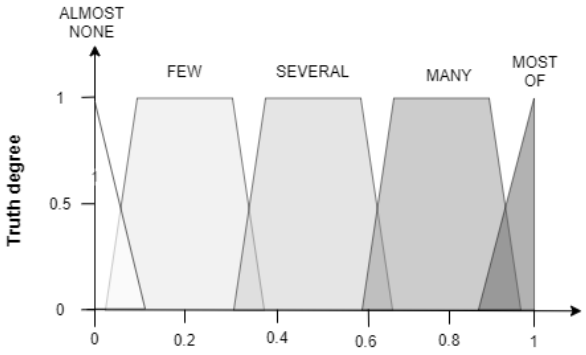


Figure 1.1: Definition of fuzzy quantifiers as fuzzy sets.

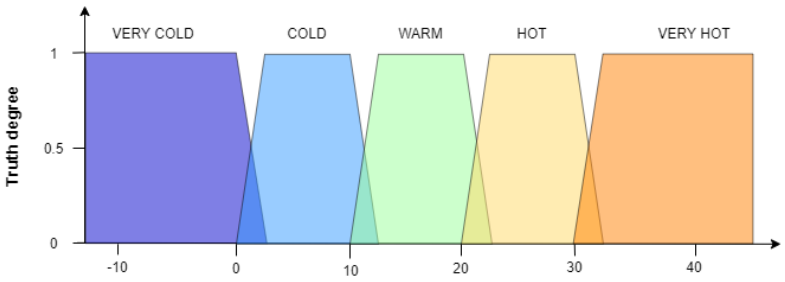


Figure 1.2: Linguistic variable composed of five labels defined as fuzzy sets.

Currently, in the literature within Data-To-Text systems based on Linguistic Descriptions of Data, the proposals cannot describe the information in a sufficiently expressive way as they do not reach the level of detail that the data features and the related domain demands. Specifically, this problem can be disaggregated into three aspects that need to be improved in the development of these models:

- In the content determination phase, the problem of extracting the most representative set of fuzzy quantified sentences with respect to a data set has not been addressed yet in depth in the literature, especially when the volume of data to be processed is too large.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Besides, most approaches (both heuristic and meta-heuristic) [6, 36, 38, 182] rely on a single algorithm, and no comparisons with other alternatives were carried out in terms of performance or relevance of the generated sentences. Furthermore, no significance analysis was reported in the comparison of the proposed algorithms.

- Also, at the stage of generating fuzzy quantified descriptions in the content determination phase, it is necessary to evaluate the generated candidates regarding the data and the defined linguistic terms to measure their adequacy in order to select the best set of sentences in terms of representativeness. In the literature a good deal of fuzzy quantification methods have been proposed and later studied from a theoretical point of view by analyzing an extensive list of properties they fulfill that help to characterize their behavior [10, 51, 52, 53, 56, 60, 61, 62, 90, 226]. From this perspective, all the fuzzy quantification models exhibit different behavior, since all of them fulfill different properties. Notwithstanding, the behavior of the models has not been studied yet from a practical or pragmatical perspective, by analyzing the real quantitative differences existing among them.

- In the literature, the proposed models are mainly based on unary and binary fuzzy quantified statements [117, 122, 145, 208, 268]. Although these structures are useful in many cases, the properties and requirements of certain areas and especially when working in real application use cases, the information they provide can be limited.

1.2 Hypothesis and objectives

The baseline hypothesis of this dissertation is as follows: *Protoform structures that are more expressive than fuzzy quantified statements and more efficient search methods are required in the design of Data-To-Text systems based on Linguistic Descriptions of Data to meet the needs identified in real use cases.*

Therefore, the overall aim of this thesis is to investigate the usage of Artificial Intelligence-based techniques for the generation of linguistic descriptions using fuzzy protoforms within Data-To-Text systems in large data set contexts. The objective of this thesis comprises the following sub-objectives:

- **O1.** To improve and extent the content determination phase in Data-To-Text systems, introducing new techniques based on Artificial Intelligence for the representation of

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

imprecise knowledge. Specifically, to propose new models formed by fuzzy protoforms including geographic information with the use of different meta-heuristic solutions for the extraction of the set of protoforms that best represent the data to be described in large search spaces.

- **O2.** To measure and compare the impact of the selection of the fuzzy quantification method. In particular, to analyze its behavior empirically when it is used for the evaluation of fuzzy protoforms to check if there are significant differences between the behavior of the different fuzzy quantification methods. If detected, the selection of the quantification method has to be conditioned by this feature or, if no significant differences are detected between them, the case in which the selection of the quantification method should be guided by other features.

- **O3.** To design and implement a model that covers the Data-To-Text pipeline [188, 192, 193] to describe data series proposing new types of sentences that are more complex than those usually used, so that they contain relevant information. Furthermore, in this type of data, the temporal dimension is critical, therefore, our objective is also to design a fuzzy temporal ontology that defines the temporal concepts necessary for managing temporal knowledge.

1.3 Contributions

The research developed in this thesis has led to the following contributions:

- **C1.** A design of fuzzy quantified sentences to handle the geographic dimension found in several types of data. This dimension is very important in geolocated data, but it is necessary to model the semantics corresponding to geographic references, which do not have a precise definition. Therefore, a fuzzy definition of these terms was proposed in this work.

- **C2.** A design and comparison of two meta-heuristic approaches for the exploration of large search spaces with the aim of extracting the set of the most representative unary and binary fuzzy quantified statements with regard to the data. Exploring a search space is a computationally expensive task, therefore, we proposed approaches that offer a balance between low execution time and a sufficiently representative set of descriptions.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- **C3.** An empirical study of the correlation in the behavior of the set of six widely used fuzzy quantification models in the literature in the evaluation of unary and binary quantified statements with meteorological data. In the related work, only the theoretical properties that these methods fulfill, or not, have been analyzed. However, this is not enough since, only attending to these properties, all the methods have a different behavior, since they fulfill different properties. Therefore, this study empirically tested whether there was a correlation between the behavior of these methods and therefore they behave similarly from a pragmatic point of view.

- **C4.** A statistical significance study of the behavior of six fuzzy quantification methods widely used in the literature in the evaluation of unary and binary fuzzy quantified sentences for a set of 15 different data sets used for different Artificial Intelligence related tasks. This contribution extends the work done in the previous one, performing a statistical significance analysis between quantification methods by expanding the number of data sets considered.

- **C5.** A Data-To-Text model based on computational intelligence and natural language generation for automatically explaining temporal series. This model allows the construction of more complex linguistic descriptions than those usually proposed in related work, based on unary and binary quantified sentences, which are uninformative in many environments, especially real use cases.

- **C6.** A temporal model composed of a fuzzy temporal ontology with temporal references that address the problem of managing the imprecision of the temporal knowledge. The temporal dimension is very important in many types of data within real environments, therefore, our proposed model offered support in modeling temporal references as needed.

- **C7.** The application of the Data-To-Text and the fuzzy temporal models in two real application areas developing two systems: ICA2TEXT in the meteorology field, with the aim of describing air quality time series from the Galician Meteorology Agency (MeteoGalicia) [154] data; and SMART CARDIA in the medicine field, in particular focused to describe time series of medical indicators about patients in the cardiac rehabilitation program.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.3.1 Publications

The contributions detailed above are included in the following publications, which constitute the scientific output of the research developed during the development of this thesis:

Conferences

- A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín, *Descripciones lingüísticas de datos de observación meteorológica usando temple simulado*. In proceedings of the 2018 XVIII Congreso Español sobre Tecnologías y Lógica Fuzzy (ESTYLF), pp. 469-474, Granada (Spain), 2018.

- C. Heble-Lahera, A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín, *Empirical Study of Fuzzy Quantification Models for Linguistic Descriptions of Meteorological Data*. In proceedings of the 2020 IEEE International Conference of Fuzzy Systems (FUZZ-IEEE), pp. 1-7, Glasgow (United Kingdom), 2020. ISSN: 1558-4739.
  Conference ranking (GRIN-SCIE): A-, Class 2.

- A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín, *An experimental study on the use of fuzzy quantification models for linguistic descriptions of data*. In proceedings of the 24th European Conference of Artificial Intelligence (ECAI), pp. 267-274, Santiago de Compostela (Spain), 2020. ISBN: 978-1-64368-100-9 (print), 978-1-64368-101-6 (online).
  Conference ranking (GRIN-SCIE): A, Class 2.

- A. Cascallar-Fuentes, J. Gallego-Fernández, A. Ramos-Soto, A. Saunders, A. Bugarín, *Un sistema para la descripción automática en lenguaje natural de gráficas de sectores: aplicación en datos de calidad del aire*. In proceedings of the 2021 XX Congreso Español sobre Tecnologícas y Lógica Fuzzy (ESTYLF), pp. 292-297, Málaga (Spain), 2021.

Journals

- A. Cascallar-Fuentes, Javier Gallego-Fernández, Alejandro Ramos-Soto, Anthony Saunders, Alberto Bugarín-Diz, *Automatic generation of textual descrptions in data-to-text systems using a fuzzy temporal ontology: application in air quality index data series*. Applied Soft Computing, vol. 119, 2022. ISSN: 1568-4946.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Journal Impact Factor: 6.725 (JCR 2020) 1.290 (SJR 2020). Journal ranked in JCR 2020, in Computer Science, Artificial Intelligence (Q1, 24/139), and D1 Computer Science, Interdisciplinary applications (D1, 11/111) categories.

- A. Cascallar-Fuentes, Alejandro Ramos-Soto, Alberto Bugarín-Diz, *Meta-heuristics for generation of linguistic descriptions of weather data: experimental comparison of two approaches*. Fuzzy Sets and Systems, 2022. ISSN: 0165-0114.
  Journal Impact Factor: 3.343 (JCR 2020), 0.902 (SJR 2020). Journal ranked in JCR 2020, in Computer science, Theory & Methods (Q1, 23/110), Mathematics, Applied (D1, 18/265), Statistics & Probability (Q1, 20/125) categories.

1.3.2 Transfer of results

In addition to the publications detailed above, during the development of this doctoral thesis, the results have been transferred to other institutions.

Transfer to the society/scientific community

- SimpleNLG-GL[2]: it is a Java API developed within the TECANDALI network to facilitate the tasks of linguistic realization in the Natural Language Generation task for the Galician language. It is a trilingual English/Spanish/Galician adaptation of the SimpleNLG v4.4.8 library [88], developed following the structure of the Spanish version [185]. SimpleNLG-GL was presented at the 11th International Conference on Natural Language Generation (INLG 2017) [31].

- MeteoText[3]: this demonstrator, developed in collaboration with Javier González-Corbelle, allows to generate descriptions from meteorological data in both Galician and Spanish for all the municipalities of Galicia using the Galician Meteorological Agency's (MeteoGalicia) [154] forecast data. It also allows to modify the meteorological data on demand. In this demonstrator we have developed a new version of GALiWeather [184], which allows to generate more dynamic texts, including synonyms for the terms used. Moreover, for the Galician language, it allows to select a linguistic area, which has particular characteristics in terms used for a certain phenomenon, and allows to limit

[2] https://github.com/citiususc/SimpleNLG-GL
[3] https://demos.citius.usc.es/meteotext/

the terms using only the normative ones or including also those of each linguistic variety, which are not recognized by the regulations but are more friendly with the way of speaking of the citizens.

Transfer to institutions

- PREBENCAR (Intelligent Virtual Advisor for Cardiovascular Prevention and Wellness): this project was led by the IT company Balidea and developed in collaboration with the cardiology area of the University Clinical Hospital of Santiago de Compostela. Within this assistant for cardiovascular prevention and wellness, we developed a Data-To-Text (D2T) system to generate reports in natural language from the data that a patient registers in the application on a set of medical variables of interest, related to sport and health (weight, heart rate,...).

- ICA2TEXT: this is a Data-To-Text (D2T) system to generate linguistic descriptions from air quality in collaboration with Javier Gallego-Fernández and experts from MeteoGalicia [154]. Specifically, this system generates descriptions from data on the distribution of air quality values in Galician meteorological stations and from hourly time series of both evolution and prediction for each station. This system is currently undergoing the last phase of testing, so it is very close to being deployed on MeteoGalicia's production server.

- SMART CARDIA (Intelligent and predictive platform to support the clinical management of the cardiac rehabilitation process): it is a project led by the company Balidea aimed at developing a software platform for the intelligent management of the cardiac rehabilitation process. Within this project, we have developed a Data-To-Text (D2T) system to generate descriptions from a large number of medical indicators of interest in the cardiac rehabilitation process of patients who have suffered from heart disease in collaboration with the cardiology area of the University Clinical Hospital of Santiago de Compostela. This project has been recognized with award "Conocimiento" in the VI Forum RIES Retos Internacionales del Ecosistema Salud del Clúster Saúde de Galicia (CSG), 10 de Noviembre de 2021.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.4 Methodology

This dissertation follows the scientific method [89], an iterative process composed by *i)* a first phase in which the objectives to be addressed during its development are defined, followed by *ii)* a stage of study of the related works proposed in the literature, *iii)* development of the research to meet the objectives that have been proposed, *iv)* a testing phase to validate the proposed methods and, if applicable, *v)* application of the developed methods in real cases. This set of phases has been followed in the process carried out to meet the objectives defined in this thesis. Nevertheless, it is important to highlight that these stages are not rigid and that they adapt to the conditions, for example, perform a set of tests before completing the development of the research to have references to its progress; or returning to a previous stage to make the improvements that are considered necessary. The phases described above are applied as follows in this thesis:

- Definition of objectives: the first phase, which is carried out in Chapter 1, consists of determining the objectives to meet. A deep analysis of these objectives is carried out in order to define them in more detail and identify the sub-objectives that make them up.

- Study of the related work: this stage is detailed in Chapter 2, performing an analysis of the state-of-the art to review all the related works proposed so far, which allows studying the current state of this field of research and evaluating the existing models and techniques, and their weaknesses and strengths. This will determine the needs that the proposed research should address.

- Research development: once the objectives have been defined and the current state-of-the research field has been analyzed, the proposal of Artificial Intelligence models application on real use cases is carried out. Chapters 3, 4 and 5 include the details of this experimentation. These cases must provide, on the one hand, valid data sets on which to apply the models, while also establishing the knowledge representation requirements that will guide the proposal of new models.

- Testing: as a last phase, also at these three chapters, once the approach has been developed, a series of tests are performed to validate the proposed methods. For this, the models generated are applied to the real examples of use selected in the previous phases and their quality is evaluated following the usual standards.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Application in real cases: in this stage, the models proposed in Chapter 5 are applied and validated in real problems, within innovation and transfer projects, in the domains of health and environmental information.

1.5 Dissertation structure

This Ph.D. dissertation is composed of six chapters that analyze the related prior work, describe in detail the contributions of this thesis and a discussion of the experimental results, and present the conclusions of this research. More concretely, the document structure is as follows:

- First of all, in Chapter 2, an exhaustive state-of-the-art review is presented. First, this chapter includes a general review of the literature which covers both Natural Language Generation and Linguistic Descriptions of Data fields, specifically, focusing on the use in the literature of LDD as a technique within the NLG in Data-To-Text systems for extracting linguistic information dealing with uncertainty and imprecision. Moreover, a deeper review of the state-of-the-art related to each of the above-mentioned objectives is presented.

- Chapter 3 is focused on the objective O1, describing the proposed approach to improve the content determination of Data-To-Text systems proposing fuzzy quantified descriptions that contain geographic information designing two meta-heuristic approaches applied to two different weather data sets which describes the weather situation in Galicia (NW Spain) with data provided by the Galician Meteorological Agency, MeteoGalicia [154] — contributions C1 and C2 —.

- In Chapter 4, the details of the performed experimental comparison studies of a set of six widely used fuzzy quantification methods behavior in the evaluation of unary and binary fuzzy quantified statements resulting from the objective O2 are presented. A preliminary empirical study of the correlation —contribution C3 — with meteorological data sets and a statistical significance study —contribution C4 — with many Artificial Intelligence-related tasks data sets of the behavior of the fuzzy quantification models are reported.

- Chapter 5 is focused on the objective O3, presenting the details of the proposed Data-To-Text pipeline to describe time series — contribution C5 —. This includes the pro-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

posal of a fuzzy temporal model — contribution C6 — defined through a fuzzy temporal ontology which addresses the problem of managing the imprecision of the temporal information. Moreover, this chapter details the application of these proposed models in two real application areas, meteorology, namely air quality index; and medicine, in particular cardiology — contribution C7 —.

- Finally, in Chapter 6, the concluding remarks of this thesis are detailed, including a summary and a discussion of the contributions of the research developed in this thesis and the experimental results. Moreover, this chapter includes the discussion of future lines of work that could extend the three objectives of this dissertation research, improving the results.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER 2

RELATED WORK

To solve the difficulties encountered in the task of extracting relevant information from various data sources in a human-understandable way, Artificial Intelligence provides users with tools to analyze data sets in order to extract useful information from them. Natural Language Processing (NLP) is a branch of Artificial Intelligence that researches and formulates computationally effective mechanisms that facilitate human-machine interaction and enable much more fluid communication than formal languages. NLP techniques have been applied in several application fields having a great impact, such as health [214, 266].

Within NLP there are two main lines of research: *i)* Natural Language Understanding (NLU), a branch that deals with machines understanding the content of texts, applied, for example in question answering [163] or dialogue systems [96], and *ii)* Natural Language Generation (NLG), which deals with the generation of texts from several data sources [48] in different applications fields such as health [176] or project management [231]. Natural Language Generation is a subfield within NLP that deals with the generation of text from various data sources to address the problem of generating relevant information in a form understandable to users using natural language [192]. Within this field, there are different research areas:

- Summarization: systems generate summarized texts from other texts. In this type of systems it is necessary to apply natural language comprehension techniques. An example of system is applied to summarize social media texts [125].
- Dialog systems: systems generate dialogues from texts made by people or bots. For example, to allow robots to interact with humans [81]. In this type of systems it is also

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

necessary to apply natural language understanding techniques.

- Narrative and poetics: these types of systems generate different types of narratives or poetry, using, for example, neural networks [82].
- Data-To-Text (D2T): this type of system generates text from sets or series of numerical or symbolic data, for example, from meteorological or health data [100, 161].

The Natural Language Generation field has existed since ELIZA [230], a natural processing program to carry out certain types of machine-human conversations created in 1966. Nevertheless, it was not until the 1980s that the first applications that translate data into text appeared [25, 127, 167]. Since then, interest in NLG has increased and systems with high impact have emerged in several fields, such as the medical area [101, 176, 225], meteorology [43, 91, 184, 210], air quality reports [27, 181, 229] or project management [231].

The aim of Natural Language Generation is to automatically generate texts from different types of data sources that cannot be easily distinguished from handwritten texts by users, while ensuring that the texts produced meet a set of requirements. Making texts produced by an NLG system indistinguishable by users is a problem that needs to be addressed in the NLG research as it is a challenge in design and evaluation, since the assessment of these texts must cover a wide set of requirements (e.g. quality, accuracy, clarity, etc.) [28, 223]. In the current state-of-the art, there is not a consensus on the design of a NLG system. In the literature, different types of architectures were proposed, although a system can combine features from over one of them:

- Modular architectures, which contain crisp division among sub-tasks with significant variations.
- Planning perspectives, with a less modular perspective on the sub-tasks.
- Integrated approaches, which cut across sub-task division with high reliance on statistical learning.

From the different architectures, the most widely used and accepted is Reiter and Dale's pipeline [188, 192, 193], which considers converting a certain type of input data into an output text as the main task of an NLG system. However, this task is composed of a set of sub-stages that define the necessary steps to generate the output. This pipeline is composed of three general tasks:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Text planning: focused on the strategic generation, i.e., “what to say”.
- Sentence planning: combines sentence aggregation, lexicalization and referring expression generation, i.e., “how to say it”.
- Linguistic realization: applies the syntactic and morphological rules to produce a text which is syntactically, morphologically and orthographically correct.

These three main tasks are divided into six sub-tasks (Figure 2.1 shows a diagram with the mapping between general phases and sub-tasks):

- Content determination: this is the process of deciding what information should be conveyed in the text. In this phase, the set of messages to be included are defined from the input data. These messages are defined with a formal language that defines entities, concepts and relationships of the domain to which the data belongs.
- Discourse planning: at this stage, an order and structure is provided to the messages in order to generate a well-structured text.
- Sentence aggregation: at this step, to improve the readability of the resulting text, many messages are grouped in a single sentence, if necessary.
- Lexicalization: this task is focused on selecting the specific words and expressions to be used in the text to describe the data.
- Referring expression generation: this phase is focused on selecting expressions to unequivocally identify each entity from the domain.
- Linguistic realization: this task is the same as defined in the general pipeline described above. This phase is focused on applying the syntactic and morphological rules to ensure that the generated text is correct.

Although this pipeline is considered the standard and these tasks are vital in the design of a correct NLG system, they can be modified in specific applications as required by the domain, for example, by splitting one task into more specific sub-tasks or merging two tasks into a general one.

USC UNIVERSIDADE

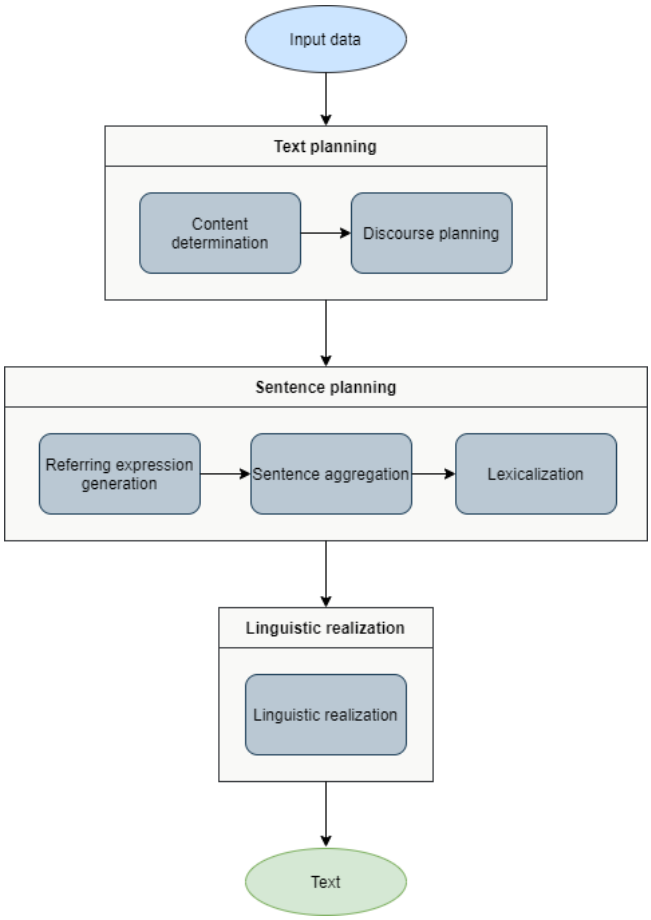


Figure 2.1: NLG pipeline defined in [188, 192, 193].

It is important to mention that the architectures described above are not mutually exclusive, i.e., one approach may contain features of two different architectures. The following is a description of the proposed approaches:

- Regarding to modular approaches, existing surveys [190, 192, 193] agree that the architecture originally introduced in [188] is the standard modular pipeline. A related architecture proposed by Levelt [132, 133] makes a similar distinction of what to say and how to say it. More recently, in [151] the "object-and-arrows" framework is proposed, where the information flows between the sub-tasks. In [189], Reiter proposes

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

an architecture to accommodate systems with raw data as input, which requires a preprocessing step. In order to solve some problems related to early decision making or generation under constraints among others, alternative architectures have been proposed, such as blackboard architectures [166], in which tasks are not rigidly organized; or revision-based architectures [104, 142], which allow feedback between modules having the possibility of changing unsatisfactory choices.

- In AI, the planning problem is defined as the process to achieve a specific objective that is composed of sub-goals, satisfied by actions with preconditions and effects. In the classical planning-based approach [78], tasks are represented as tuples of preconditions and effects. The connection between planning and NLG is achieving a communicative goal through the performance of a set of tasks [55, 84]. One approach inside this perspective is planning through the grammar, where recent approaches integrate planning into the grammar viewing linguistic structures as planning operators. These approaches are commonly couched in the formalism of Lexicalized Tree Adjoining Grammar (LTAG) [108], where pieces of linguistic structures can be coupled with information that specify the semantic preconditions and the pragmatic objectives (e.g., [85, 128] use this approach). An advantage of this type of approach is the availability of a significant number of off-the-shelf planners, however, they usually have efficiency problems.
  These approaches assume the relationship between an action and its consequences as fixed, but this is not realistic since the consequences of an output arise from several sources of uncertainty. Therefore, another approach is understanding this problem as a stochastic optimization problem [197, 198].

- Currently, Data-To-Text corpora exists for specific domains consisting of database elements paired with free text. A promising approach is introducing statistical techniques to automatically align data and text [129, 134].

- Another way to make decisions is in terms of classification, used in tasks such as content determination [69] or realization [79]. Considering that generation is a leveled decision-making process, one way to model it using classifiers is cascading, so that the output of one classifier is the input of the next. This approach maintains a separation between the NLG tasks, but it may experience the problem of error propagation.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Lately, there has also been a growing interest in understanding generation through probabilistic Context-Free Grammar (CFG). An example is [14], with a base generator hand-crafted. However, it is possible to extract rules from corpora [213, 232].

- In recent years, there has been a growing interest in the NLG field for applications with Neural Network (NN) architectures. The reason for such increased interest in these techniques is due to, firstly, advances in hardware, and because they can deal with sequences of varying lengths, avoiding parameter explosion. In recent times, many approaches have been proposed for D2T systems using neural networks [105, 123, 143]. Notwithstanding, it is important to highlight that this type of approach suffers from some problems: *i)* first, a lack of training data with text samples available [41] and *ii)* it suffers from the problem of information missing, repetition and "hallucination" [187, 207]. Some approaches have already emerged to solve these problems [42, 165, 205], although this is still an open problem.

Another challenge still open in the literature is the evaluation of NLG systems. The first problem we encounter is the lack of a standard in the input format, which complicates the comparison between systems. Moreover, even for the same input and a single system, there are a large number of possible correct outputs [86]. A methodological distinction exists between intrinsic and extrinsic evaluation methods [107]:

- An intrinsic evaluation method measures the performance of a system in terms of its effectiveness regarding to the users. In NLG, the intrinsic evaluation is based on human judgments, who assess them rating a set of criteria (e.g., readability [74, 156], accuracy [103, 211], etc.); or corpora [16, 74], that measures the extent to which texts generated by a system resemble those written by humans under a set of comparable conditions using automatically computed metrics [203], used in related fields as machine learning. In the literature, some papers have focused on reviewing the current state-of-the art in evaluation, analyzing the existing proposals, and discussing the problems encountered [7, 102, 222]. These works reveal that, in the literature, human-based evaluations are mainly based on Likert rating scales method, from 2-point to 10-point, being 5-point the most popular option followed by 2-point and 3-point scales [7, 222]. Furthermore, [222] proposes a series of best practices in the evaluation of NLG systems, [102] proposes quality criteria and definitions to understand what is being evaluated in a specific case, and [216] proposes a methodology to evaluate the accuracy of

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

texts generated by D2T systems. At the same time that the Likert scale has become more popular in the evaluation of D2T systems, its use has received more criticism for not being sufficiently consistent and reliable [40, 170, 248]. Regarding to the application of automatic metrics in NLG evaluation, such as BLEU [174] or ROUGE [135], some studies in the literature concluded that, although these measures have a considerable potential in specific cases, they usually do not correlate well with human evaluations, so they should be used cautiously as a complement to expert evaluations or in development phases to guide developers [15, 169, 191].

- Extrinsic evaluation methods are focused on measuring the effectiveness of a system in achieving a desired goal (e.g., persuasion and behavior change [195], purchasing decision [29], decision support in a medical setting [176], etc.)

Among the different areas of research within the field of NLG, in this review of related work, we focus here on Data-To-Text (D2T) systems. There are currently numerous D2T systems applied in various domains and developed for a wide range of purposes. The following is a review of the most relevant Data-To-Text systems in those domains where most efforts have been focused. In the field of meteorology, the application of D2T systems for the generation of texts from various data sources has always been of great interest, with many proposals in this area:

- FOG [91] is a bilingual (English and French) Data-To-Text report generator that produces daily forecasts from weather maps that cannot be distinguished from forecasts written manually by humans. Texts describe data from Canada, United States and United Kingdom, which until that had been done manually.
- MultiMeteo [43, 44] is a multi-language weather report Data-To-Text system developed by many European weather agencies including Instituto Nacional de Meteorología (Spain), Météo-France (France), Institut Royal Météorologique (Belgique) and Zentralanstalt fur Meteorologie und Geodynamik (Austria).
- SumTime-Turbine [251, 252] is a system that produces textual summaries of time series data from gas turbines in English. Every second 600 data points from over 250 channels are monitored, therefore, a manual analysis is not feasible.
- SumTime-Mousam [210] is a Data-To-Text weather forecast generator for offshore oil platforms from a multivariate time series that allows forecasters to adapt the output

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

text to meet their needs, so they have the control of the forecast content. In a later work [196], SumTime-Mousam was revised in order to determine if the generated texts were correct and the suitability of the selected words. From this analysis, the system was rebuilt, producing texts that avoid words that are only used in certain dialects or that can be confusing in conveying information because their meaning depends on the dialect in which they are used.

- The RoadSafe [219, 220] project aims to design a Data-To-Text system that generates textual weather forecasts for winter road maintenance application from spatio-temporal data.

- GALiWeather [184] is an NLG system which automatically generates short-term weather forecasts in natural language from prediction data of the Galician Meteorology Agency (MeteoGalicia) [154]. In a later work [181], this system functionality was extended, including a Data-To-Text system that produces textual forecasts from the air quality data.

- The MARQUIS system [229] is a Data-To-Text system that generate textual summaries from numerical air quality time series including simple time references (specific hours or intervals) for describing the last air quality value and information of the pollutants (concentration, archive information and forecast).

- [199] describes a Data-To-Text system that automatically generates textual descriptions from numerical data, in particular, weather prediction and air quality.

- [161] describes a Data-To-Text system to automatically generate weather comments from simulation results of numerical weather prediction models and past observation data.

Various systems have also been developed in the field of medicine, some of which have had a major impact on the sector:

- The STOP system [194, 195] aims to change the behavior of smoking patients by automatically generating personalized letters using as input the answers of a questionnaire previously filled in by them. This proposal did not obtain the expected results since it did not increase the percentage of patients who quit smoking with respect to manually

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

written letters. However, the results are not clear and it seems that the challenge is the complexity of the problem rather than the system.

- The Suregen-2 system [130] generates different types of medical reports, such as procedure reports, clinical findings or referral letters in German.
- A web-based NLG system is proposed in [64] to provide better access to relevant information that patients need to understand their medical condition to make more informed decisions. This approach is aimed at patients who are going to undergo reconstructive breast surgery, as the surgical procedures are very complex.
- BabyTalk [103, 176] is a project composed of a set of NLG systems which aims to generate several types of reports from Neonatal Intensive Care Unit (NICU) data, such as texts for parents or nurses. An example is BT-Nurse, which summarizes 12 hours of NICU data intending to support nurses in the beginning of a nursing shit. Other system is BT-45, which summarizes 45 minutes of NICU data with the aim of supporting decision-making in real time by the clinical staff.
- In [100] a Data-To-Text system is proposed to generate targeted reports for colorectal cancer patients providing information on quality of life implications after treatment in a personalized way.

As this review of related work has shown, the field of NLG has been the focus of attention in addressing the problem of generating natural language text from various data sources. Moreover, the interest in the field of NLG has grown considerably in recent years. Nevertheless, in this field, the semantics of the terms used are not modeled, which is a shortcoming when it comes to handling the imprecision of the language. Aiming to manage the uncertainty and imprecision inherent in language, in the Fuzzy Logic field [255, 262] several approaches have emerged that focus on the development of data descriptions modeling the semantic of imprecise language elements through fuzzy linguistic terms that deals with the inherent imprecision of the language.

Between 1980 and 1990, Yager provided a preliminary definition of linguistic description introducing the concept of quantified sentence [243, 247]. In the mid-1990s, seminal work by Zadeh introduced the Computing With Words (CWW) paradigm [258, 264, 263], where computations are performed on linguistic terms modeled as fuzzy sets, and its evolution, Computational Theory of Perceptions (CTP), also called computing with perceptions [259,

USC UNIVERSIDADE

260]. Although several approaches based on these paradigms were proposed, a useful technique to model the uncertainty of imprecise terms applied in the content determination stage within the NLG pipeline [180] is Linguistic Descriptions of Data (LDD) [116, 118, 217], which summarizes in a linguistic form one or more numerical variables and their values using the general concept of protoform [261] and the more specific of fuzzy quantified statements, which provide quantitative information about the fulfillment of some properties of interest in a numerical data set.

Considering that LDD is a fairly recent field of research when compared to the NLG domain because most of the research in this direction can be found from the year 2000 onwards, there is not yet a general approach to generate different types of linguistic descriptions for any type of application. Granular Linguistic Model of a Phenomemon (GLMP) [217] is a framework that supports the generation of linguistic descriptions applying fuzzy rules over these statements combining ideas from two fields: the Computational Theory of Perceptions (CTP) [259] and systemic functional linguistics [95]. It is based on a hierarchy of interconnected nodes called Perception Mappings (PM) with a set of Computational Perceptions (CP) as input. This framework has been used in several domains as healthcare, describing sensors' data or energy consumption, among others [45, 72, 147, 148, 152, 175, 177, 268]. In the literature, approaches are mainly based on unary and binary quantified statements [117, 122, 145, 208, 268]. However, these structures are not suitable in other domains, for instance, process management, health, elder care, etc. Thus, other types of structures have been proposed in other studies [5, 8, 37, 120, 159, 233, 238, 268].

The process of constructing linguistic descriptions is guided by the fuzzy techniques on which it is based. In this process, the following components must be defined:

- Fuzzy quantifiers, such as "many", "at least 5", "all except about 5", "twice as many", etc. Their meaning is associated with fuzzy sets, represented through a fuzzy membership function [258].

- Linguistic variables which can be based on categorical or numerical variables. For example, from a numerical variable "temperature" an associated linguistic variable can be composed of five labels defined as fuzzy sets: "very cold", "cold", "warm", "hot", "very hot" (see Figure 1.2). These labels are also associated with a fuzzy membership function.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

	Protoform structure	**Example**
[37]	Q of $D^S_{i,j}$ are A^S	Most days of year 2001, both series exhibit a local change with the same sign (0.714104)
[34]	Q of D X is A	At least 70% of the time with cold weather, the patient inflow is much higher in center A than in center B.
[35]	Q of DS are AS	Most days of the hot season, patient inflow was low or very low
[5]	Q y's with care P Q R y's with care P Q y's care P Qt time Q R y's with care P Qt time	Most patients with disease X have low blood pressure Most patients with disease X have a low heart rate most of the time
[234]	y's are getting PG	Resident Y is getting less active
[221]	Q of X in T1 is C to T2	About two thirds of the days the consumption in the mornings is lower than the consumption in the afternoons
[109]	Among all segments, Q are P Among all R segments, Q are P	Among all segments, most are slowly increasing Among all short segments, most are slowly increasing
[59]	Association between AXs and BXs is C in T	Association between risky high pressures and high temperatures has been very high in the last few seconds
[159, 160]	Prec1 TizmeCtx} the series is Pdty(PI) with a period of Prec2 pˆ units} -	The first two months, the series is highly periodic with a period of approximately 1 week
[233]	In Q cases sequence XYZ was P In Q cases, when condition R was fulfilled, there was P When condition R was fulfilled, sequence XYZ was P in Q cases	In most cases sequence "Registration, Analysis, Decision" was fast In most cases, when case was accepted, there was a sequence "Registration, Analysis, Decision" When case was accepted, sequence "Registration, Analysis, Decision" was done by the same employee in most cases
[148, 150, 175]	$V_r T_j Qk$	HR is high around 15 and 30 minutes with normal intensity

Table 2.1: Types of protoforms structures in LDD

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Aggregation operators, which allow to create more complex expressions composed of a combination of linguistic terms, such as “very hot and dry” or “young or blonde”.
- Evaluation criteria to assess the suitability of a generated description to select the most representative sentences with regard to a data set, such as the truth degree.

The evaluation of linguistic descriptions of data is a challenge in the literature, since each application domain is subject to its own criteria. Nonetheless, there is a strong consensus in the literature on the dimensions to be taken into account as a starting point when evaluating linguistic descriptions [26, 39, 63, 236, 237, 240]:

- Truth degree: this is the most important criteria to measure the quality of a fuzzy quantified statement.
- Relevance: it is related to the number of elements in the data set that must fulfill the condition in the description to be considered relevant.
- Accuracy: it measures the degree a description conveys a true reflection of the input data.
- Length: the size of the description may be important in some domains, for example, preferring to create shorter descriptions, losing some precision or not.
- Coverage: this criterion measures the extent to which the data are described and what data are not included in the description.
- Specificity: this criterion measures how general a description is, considering that the more general, the less informative it is.
- Compatibility of the descriptions: having two or more descriptions, it is necessary to know the degree of incompatibility between them.
- Ambiguity: two statements can be compatible but ambiguous so they can confuse users.

Other proposals to evaluate linguistic descriptions [26, 153] have been based on conversational Gricean Maxims [94]:

- Maxim of quality: to be truthful, not giving false information or not supported by evidence.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Maxim of quantity: to be as informative as possible, giving as much information as needed and no more.
- Maxim of relation: to be relevant, saying things pertinent to the situation
- Maxim of manner: to be clear, brief and orderly, avoiding ambiguity.

There are several challenges that must often be faced in the design of an LDD approach. For example, the definition of quantifiers and linguistic variables, which is a domain-related task, has the inherent imprecision associated with the use of fuzzy sets. Likewise, depending on the number of quantifiers and linguistic terms considered, the number of potential linguistic descriptions that will be generated could end up becoming too large, getting the set of the most suitable sentences for a given data set is considered a search problem. Thus, approaches proposed in LDD are based on search algorithms generating all possible fuzzy sentences so that the most representative set can be selected. For this, the candidate descriptions must be evaluated in order to accept those that are of interest and discard those that are not sufficiently representative. Nevertheless, in some scenarios, working with very large data sets and defining many linguistic variables make unfeasible the generation of all possible different quantified statements since the search space cannot be fully explored efficiently, therefore, in the literature both heuristic and meta-heuristic approaches were proposed to guide the linguistic description search process according to pre-defined criteria.

2.1 Obtaining the most representative set of descriptions in Linguistic Descriptions of Data (LDD)

Heuristics are usually problem-dependent search techniques which guarantee finding a good solution (usually not the optimal one) in a reasonable amount of time. Heuristic strategies were proposed, for instance, for time series description [36] or the meteorological prediction generation in natural language [182, 183]. On the other hand, meta-heuristic algorithms are high-level problem-independent solutions which optimize a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. In [37] a multi-objective evolutionary algorithm was proposed for the times series description with the aim of generating the shortest and most precise summary that covered the time dimension. Also, a genetic algorithm was proposed in [6] for generating linguistic summaries about operational

UGR UNIVERSIDAD DE GRANADA

and financial healthcare data. Another bio-inspired approach for linguistic summaries generation is the Ant Colony Optimization model proposed in [65]. More recently, in [2] a case study was conducted on linguistic summarization of cradle design using a genetic algorithm. In general, the problem of generating LDD in contexts where the search space and/or the data size become too large did not receive much attention in the literature. Most approaches (both heuristic and meta-heuristic) rely on a single algorithm, and no comparisons with other alternatives are made in terms of performance or relevance of the generated sentences.

2.2 Obtaining the truth degree of fuzzy quantified sentences using fuzzy quantified methods

In fuzzy quantification, Zadeh [254, 255, 256, 257] proposed an extension of the classical existential and universal quantifiers ("exist" and "for all"), as well as other crisply defined quantifiers (e.g. "over 40%") to imprecise (fuzzy) quantifiers with a higher degree of expressiveness, such as "a few" or "most of". Later on, following a different perspective research line for the proposal of imprecise quantification models, the Theory of Generalized Quantifiers (TGQ) was developed [13, 18, 124] independently. In [90], a generalization of the TGQ based on Quantifier Fuzzification Mechanisms (QFM) was proposed. This generalization allows to define a fuzzy quantifier based on a transformation from semi-fuzzy quantifiers, which are easier to design. This mechanism applies to a wide range of quantifiers (not just absolute and relative ones), such as comparative, exception, ternary or quaternary quantifiers.

Evaluating quantified sentences involves the use of a fuzzy quantification model, which calculates the truth degree of the sentence (a value in the range [0,1]). The truth degree in quantified sentences is a measure that combines the cardinality, i.e., how many elements in the referential match the property in the statement, and the compatibility between the cardinality and the quantifier. Several fuzzy quantification models have been proposed in the literature [20, 51, 54] and were later studied from a theoretical perspective in terms of the properties they fulfill [10, 51, 52, 53, 56, 60, 61, 62, 90, 226]. An extensive list of properties (including monotonicity, continuity, correct generalization, negation, antonymy and duality, among others) has been described, considering different aspects that help to characterize the behavior of the fuzzy quantification models. From this perspective, all the fuzzy quantification models exhibit different behavior, since all of them fulfill different properties. Also, some of them exhibit nonplausible behavior for some uses, since they cannot fulfill some relevant

USC UNIVERSIDADE

properties. But the behavior of the models has not been studied yet from a practical or pragmatical perspective, by analyzing the real quantitative differences existing among them. This experimental approach has been adopted in other research fields such as Machine Learning, for instance [76, 77].

2.2.1 Unary fuzzy quantified models

In the literature, many fuzzy quantification methods have been proposed to evaluate unary fuzzy quantified statements ("Q Y are S"). In this review of the state-of-the art we consider the most widely used quantification methods in the literature extracted from [51, 54].

Zadeh's method

This method [257] is based on the scalar cardinality "power" defined by Zadeh as $P(A) = \sum_{i=1}^{n} A(x_i)$.

The evaluation of unary quantified sentences for relative quantifiers is defined as:

$$Z_Q(A) = Q\left(\frac{P(A)}{|X|}\right) \tag{2.1}$$

For absolute quantifiers, the evaluation is as follows:

$$Z_Q(A) = Q(||P(A)||) \tag{2.2}$$

where $|P(A)|$ is the integer part of the real number P(A).

Yager's method based on OWA operators

This method is a special case of the Choquet integral [21, 22]. For unary quantified statements, the truth degree based on Choquet integral is defined as:

$$C_Q(A) = \sum_{i=1}^{n} b_i \times \left(Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right)\right) \tag{2.3}$$

where b_i is the i-th largest value belonging to the fuzzy set A.

Yager's method [245] can only be used with coherent[4] and relative quantifiers.

[4] A quantifier Q is coherent if $Q(x_i) \leq Q(x_{i+1}) \forall x_i < x_{i+1}$ and $Q(0) = 0$, $Q(1) = 1$.

Being $w_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right)$, $i \in \{1,...,n\}$ and $Q(0) = 0$, the evaluation is:

$$Y_Q(A) = \sum_{i=1}^{n} w_i b_i \tag{2.4}$$

where b_i is the i-th higher value of the fulfillment degree to the fuzzy set A.

Yager's non-OWA family of methods

Yager [244] proposed to perform the evaluation of unary sentences with relative quantifiers as follows:

$$Y'_Q(A) = \max_{C \subseteq X} \wedge_1 \left(Q\left(\frac{|C|}{n}\right), \wedge_2 \underset{x_i \in C}{A(x_i)} \right) \tag{2.5}$$

For absolute quantifiers, the evaluation is as follow:

$$Y'_Q(A) = \max_{C \subseteq X} \wedge_1 \left(Q(|C|), \wedge_2 \underset{x_i \in C}{A(x_i)} \right) \tag{2.6}$$

Methods based on the Choquet and the Sugeno integrals

The evaluation of unary quantified statements using the Choquet and the Sugeno integrals [21, 22] has been proposed, restricted to coherent quantifiers.

The evaluation of relative quantifiers based on the Choquet's integral is defined as:

$$C_Q(A) = \sum_{i=1}^{n} b_i \times \left(Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right) \right) \tag{2.7}$$

and the method based on Sugeno's integral is expressed as:

$$S_Q(A) = \max_{1 \leqslant i \leqslant n} \min \left(Q\left(\frac{i}{n}\right), b_i \right) \tag{2.8}$$

Delgado's GD method

The GD method [51, 54] is a quantification model of the so-called-G-family that belongs to a method family based on a fuzzy cardinality E, defined as follows:

$$GD_Q(A) = \underset{i \in \{0,...,n\}}{\oplus} (E(A,i) \otimes Q(i)) \tag{2.9}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Using the product as t-norm and the Lukasiewicz's t-conorm, the evaluation of a unary quantified statement with relative quantifiers is as follows:

$$GD_Q(A) = \sum_{i=0}^{n} ED(A,i) \times Q(n) \tag{2.10}$$

where $ED(A,k) = b_k - b_{k+1}$ with $b_0 = 1$ and $b_{n+1} = 0$ is the ED fuzzy cardinality [52], a particular case of the E cardinality, using the minimum t-norm, Lukasiewicz's t-norm, the maximum t-conorm and the standard negation.

Zadeh's fuzzy cardinality based method

This method [53] is based on Zadeh's fuzzy cardinality:

$$Z(A,k) = \begin{cases} 0 & \text{if } \nexists \alpha \mid |A_\alpha| = k \\ sup\{\alpha \mid |A_\alpha| = k\} & otherwise \end{cases} \tag{2.11}$$

The evaluation for unary quantified statements with absolute quantifiers is:

$$ZS_Q(A) = \max_{k \in \{0,\ldots,n\}} \min\left(Z(A,k), Q(k)\right) \tag{2.12}$$

And for relative quantifiers:

$$ZS_Q(A) = \max_{k \in \{0,\ldots,n\}} \min\left(Z(A,k), Q\left(\frac{k}{n}\right)\right) \tag{2.13}$$

It can be proved [51] that (2.13) is equivalent to:

$$ZS_Q(A) = \max_{\alpha \in M(A)} \min\left(\alpha, Q(|A_\alpha|)\right) \tag{2.14}$$

where $M(A) = \{\alpha \in (0,1] \mid \exists x_i \in X \; with \; A(x_i) = \alpha\} \cup \{1\}$
so the method evaluation can be performed without calculating the Z cardinality.

2.2.2 Binary fuzzy quantified models

The evaluation of binary fuzzy quantified statements, "*Q KY* are *S*", is slightly more complex than the evaluation of unary sentences. In the following, we consider the most well-known methods in the literature extracted from [51, 54].

Zadeh's method

This method [257] is based on the relative cardinality of A and D, so the case $D = \emptyset$ cannot be evaluated:

$$P(A/D) = \frac{P(A \cap D)}{P(D)} \tag{2.15}$$

The evaluation is as follows:

$$Z_Q(A/D) = Q\left(P(A/D)\right) = Q\left(\frac{P(A \cap D)}{P(D)}\right) \tag{2.16}$$

Yager's method based on OWA

Yager's model [245] can only be generalized to binary sentences for coherent and relative quantifiers. Its parameters are calculated as follows:

$$w_i = Q(S_i) - Q(S_{i-1})\ i \in \{1, ..., n\} \tag{2.17}$$

where

$$S_i = \frac{1}{d} \sum_{j=1}^{i} e_i,\ d = \sum_{k=1}^{n} e_k \tag{2.18}$$

being e_k the i-th low value of D set's truth degree and $S_0 = 0$. The evaluation is:

$$Y_Q(A/D) = \sum_{i=1}^{n} w_i c_i \tag{2.19}$$

where c_i is the i-th highest value of the set of truth degrees of $(\neg D \vee A)$.

Method of Vila, Cubero, Medina and Pons

This method [226] uses the "or" or "orness" degree defined by Yager [245] for coherent quantifiers, $orness(\exists) = 1$ and $orness(\forall) = 0$. The major advantage of this method is the efficiency $O(n)$ together with a non-strict evaluation. Every coherent quantifier Q between $\exists$ and $\forall$ has an orness degree in the [0, 1] interval:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$$o_Q = \sum_{i=1}^{n} \left(\frac{n-i}{n-1}\right) \times \left(Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right)\right) \tag{2.20}$$

Then, the evaluation for a binary quantified statement is:

$$\begin{aligned} V_Q(A/D) = o_Q \max_{x \in X} (D(x) \wedge A(x)) \\ +(1-o_Q) \min_{x \in X} (A(x) \vee (1-D(x))) \end{aligned} \tag{2.21}$$

Delgado's GD method

The generalization of the GD method [54] uses the fuzzy cardinality ER, which uses the product as t-norm and the Lukasiewicz's t-conorm, as follows:

$$GD_Q(A/D) = \sum_{c \in CR(A/D)} ER(A/D, c) \times Q(c) \tag{2.22}$$

where ER is the fuzzy cardinality used by this method for this evaluation:

$$ER(A/D, c) = \sum_{c = C(A/D, \alpha_i)} (\alpha_i - \alpha_{i+1}) \; \forall c \in CR(A/D) \tag{2.23}$$

where

$$C(A/D, \alpha_i) = \frac{|(A \cap D)_{\alpha_i}|}{|D_{\alpha_i}|} \tag{2.24}$$

and CR is the set of crisp representatives of the relative cardinality defined by $\alpha - cuts$:

$$CR(A/D) = \{ \frac{|(A \cap D)_{\alpha}|}{|D_{\alpha}|} \; with \; \alpha \; \in \; M(A/D) \} \tag{2.25}$$

where $M(A/D)$ is the set of representative $\alpha - cut$ levels of A with respect to D:

$$\begin{aligned} M(A/D) = M(A \cap D) \cup M(D), \; being \; D \; a \; normal \; fuzzy \; set \; and \\ M(A) = \{ \alpha \in]0,1[: \; \exists x_i \in X \; with \; A(x_i) = \alpha \} \end{aligned} \tag{2.26}$$

It can be proved [51] that the evaluation 2.22 is equivalent to:

$$GD_Q(A/D) = \sum_{\alpha_i \in M(A/D)} (\alpha_i - \alpha_{i+1}) \times Q\left(C(A/D, \alpha_i)\right) \tag{2.27}$$

if $M(A/D) = \{\alpha_1, ..., \alpha_m\}$ an α-cut set defined in 2.25 with $1 = \alpha_1 > ... > \alpha_m > \alpha_{m+1} = 0$, then:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$$C(A/D, \alpha_i) = \frac{|(A \cap D)_{\alpha_i}|}{|D_{\alpha_i}|} \tag{2.28}$$

Thus, the evaluation of a binary quantified statement can be performed without calculating the $ER(A/D)$ cardinality.

Zadeh's fuzzy cardinality based method

This method [53], which uses the fuzzy cardinality ES, consists in a max-min composition between that cardinality and the quantifier, being the evaluation as follows:

$$ZS_Q(A/D) = \max_{c \in CR(A/D)} \min\left(ES(A/D), c), Q(c)\right) \tag{2.29}$$

where *ES(A/D, c)* is the relative cardinality of A with respect to D, defined as:

$$ES(A/D, c) = max\{\alpha \in M(A/D) \mid c = \frac{|(A \cap D)_\alpha|}{|D_\alpha|}\} \; \forall c \in CR(A/D) \tag{2.30}$$

$$ZS_Q(A/D) = \max_{\alpha \in M(A/D)} \min(\alpha, Q\left(\frac{|(A \cap D)_\alpha|}{|D_\alpha|}\right) \tag{2.31}$$

Thus, once again, the evaluation of a binary quantified statement can be performed without calculating the $ES(A/D)$ cardinality independently.

2.3 Fuzzy linguistic models for the management of the uncertainty in time series

Related D2T works based on LDD, as discussed in Chapter 1, propose approaches based on basic sentence structures, mainly unary and binary. Although these statements are useful most times, their structure is too basic in terms of the information they may contain, which sometimes may be too scarce and uninformative. Moreover, certain types of information have very important dimensions to be considered where the data is processed, since they provide a context. For example, meteorological data contains a geographical dimension, which is key since the meteorology at a point is influenced by its location. Consequently, the description has to take into account its characteristics, for example, snow episodes in coastal or mountain areas. These data require a processing in which these dimensions are treated in a prioritized and cautious manner.

Time is also a very important dimension in many real-world domains where facts that occur and vary in time need to be managed, such as signal control and monitoring, medicine

USC UNIVERSIDADE

or meteorology. In these cases, the timing and order of events is essential since an error, for example, in the administration of treatment to a patient can have fatal consequences. The complexity of the temporal dimension requires finding expressions that contain temporal references and various types of relationships between them. For the handling of this dimension, the time axis is usually modeled as real numbers, so that, in the first instance, the time intervals are just intervals of real numbers. Nevertheless, in some cases, the information is expressed in an approximate manner, being available only information about relationships to other intervals and time points, so imprecise terms to describe them are required.

In the fuzzy logic field, there are several approaches to manage the uncertainty of this temporal knowledge. First of all, Vítek [227] tried to formalize the representation of fuzzy time by means of intervals associated with fuzzy sets. A different approach is Allen's logic of temporal relations [4], which express the knowledge by means of qualitative temporal relations between intervals. This approach was extended in [9] by labeling the interval relations with fuzzy values. Based on Zadeh's possibly theory [67, 253], Dubois and Prade [68] introduced dates (possibility distribution in a continuous linear time scale) and intervals (fuzzy set of time points between two dates) as primitive temporal elements and temporal relations between these temporal entities through fuzzy relations. Similarly, [131] proposes the usage of fuzzy numbers to represent interval endpoints. Also, [179], based on Zadeh's theory of possibility, introduced the concept of similarity index in reasoning over time. Moreover, they proposed a temporal proposition ("X is A; T is τ") with two components (non-temporal and temporal) separately treated. Qian [178] integrated the conventional fuzzy rules with imprecise temporal representation for problem solving in industrial dynamic systems. Barro et al. [11] presented fuzzy temporal references, including date, time extent, interval, and relations between some temporal entities. With regard to intervals, in [162] an extension of the interval relations to fuzzy intervals is defined. Also Ohlbach [171] proposes a three-level approach to fuzzy relations between fuzzy intervals. The "Dynamical fuzzy reasoning method" is proposed in [137], incorporating a vague time delay into fuzzy if-then rules. OWL-Time [99, 172] is the temporal ontology recommended by the W3C [228]. In [173] is proposed a simple time sub-ontology that provides basic temporal concepts and relations. Similarly, in [200] an ontology is proposed with mechanisms to represent temporal and fuzzy information. Another approach is [30], where a formal definition of a grammar that describes Fuzzy Temporal Propositions (FTPs) is introduced for representation of fuzzy temporal knowledge.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

In the NLG field, there are also many approaches to summarize time series linguistically.

Many approaches include methodological proposals analyzing this problem from a general point of view [24, 34, 36, 113, 119, 144, 168]. An example is [87], where Gatt and Portet present a temporal framework of uncertainty based on the Possibility Theory and propose a model which uses its outputs to select linguistic expressions. Likewise, in [145] a general architecture for Generation of Linguistic Descriptions of Time Series (GLiDTS) system is proposed. They present a general approach for generating linguistic descriptions of time series, and a global view of the concepts, elements, and processes involved. More recently, in [157] Moreno-García et al. present a technique for obtaining linguistic descriptions from time series using a representation called Fuzzy Piecewise Linear Segments, which is a set of segments that have been fuzzified from the segments of Piecewise Linear Segments (PLS), which consists of representing the series using a set of segments, each of which corresponds to a piece of the series.

Approaches to describe time series have been used in several domains, including process management [120], decision support [97], investments [110, 114], trend descriptions and change detection [37, 83, 201], and detection of period events [159]. It is worth noting that the health domain has received special attention in the literature, for instance, focusing on elderly care [146, 239], physical activity analysis [204], patient data analysis [5] or menstrual cycle analysis [209].

2.4 General remarks about the application of fuzzy techniques in the Natural Language Generation field

In this chapter, the state-of-the art in the Natural Language Generation (NLG) field has been revised. Specifically, we described the creation of Linguistic Descriptions of Data (LDD) using fuzzy techniques within this field. Both approaches focus on providing comprehensible information through natural language. Specifically, NLG focuses on the problem of converting different data sources into understandable text while LDD focuses on handling the imprecision of language defining linguistic concepts through fuzzy sets building linguistic descriptions.

LDD is a useful tool that can be identified as the content determination stage in the classical NLG pipeline [188, 192, 193]. Therefore, when designing a solution, the integration of these two fields is an interesting alternative since the use of fuzzy terms allows to generate texts with more human-like speaking style, taking into consideration the inherent imprecision

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

of language. This decision is conditioned by domain constraints and expert knowledge. In fact, as previously discussed, the application of LDD techniques within NLG systems has been very successful in solving problems in the literature.

In this context, researchers should include these techniques in the definition of their systems as they have great potential and can help them with different tasks, such as imprecision management or evaluation of the generated texts. In particular, great strides have been made in the literature in dealing with the temporal dimension found in many types of data and which can be critical in many fields, such as medicine or aeronautics. As mentioned in Section 2.1, obtaining the most representative set of descriptions is addressed as a search problem, where all possible descriptions are generated and sorted according to a set of criteria afterwards. However, this approach becomes more inefficient as the data scales up. In the literature, this problem has already been addressed in some papers by means of heuristic and meta-heuristic solutions, e.g., [6, 36]. An essential aspect when ranking the resulting descriptions is to obtain the truth degree of the sentences by means of fuzzy quantification methods. Among the many alternatives in the literature, the researcher/designer must select the most appropriate method for each case, considering the constraints of his/her model, e.g., the types of quantifiers defined, the structure of sentences to evaluate, etc.

This Ph.D. dissertation fully addresses some points discussed in this chapter in the context of applying LDD in NLG:

- In Chapter 3, two meta-heuristic solutions are provided to deal with the problem of linguistic descriptions extraction to get the set of the most representative statements regardless of the size of the search space.
- An empirical comparison between the set of the most widely used fuzzy quantification methods is performed in Chapter 4 to determine if its selection conditions the evaluation of fuzzy quantified judgments or, on the contrary, should be guided by other criteria, such as performance.
- A Data-To-Text pipeline is proposed in Chapter 5 focused on improving the description of time series. Moreover, a fuzzy temporal ontology is defined in order to address the problem of managing the imprecision of the temporal dimension of this type of data.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CHAPTER 3

USING META-HEURISTICS FOR SEARCHING LINGUISTIC DESCRIPTIONS OF DATA

In Chapter 1, we described how getting the set of the most descriptive linguistic descriptions of data can be considered as a search problem that involves generating and assessing linguistic statements on a data set. This problem is aggravated as the data sets scale, since their size and its features, such as number of variables and values, make unfeasible to explore all possible combinations in an affordable amount of time. Because of this situation, alternatives based on heuristic [36, 182, 183] or meta-heuristic [6, 38, 65] solutions been proposed in the related work to obtain the set of the most representative sentences according to pre-defined criteria in those cases where the number of possible descriptions is so large that the search space cannot be fully explored efficiently. Notwithstanding, this problem has not been addressed in depth in contexts where the search space and/or the size of the data set becomes too large.

In this chapter, we propose two solutions based on meta-heuristic algorithms for handling the problem of finding an appropriate LDD based on unary ("In some locations the sky is clear") and binary ("In many locations in the North the temperature is high") fuzzy quantified statements. On one hand, we designed a solution based on the general definition of Genetic Algorithm (GA) following [73] . On the other hand, after analyzing different meta-heuristic algorithms, we also designed a solution based on Simulated Annealing (SA) [224] as an alternative, since this meta-heuristic was reported to achieve better results than other algorithms for several tasks [136, 215]. To evaluate these proposed solutions, we generated fuzzy quantified statements using meteorological data provided by the Galician Meteorological Agency,

USC UNIVERSIDADE

MeteoGalicia [154], which describe the weather situation in Galicia (NW Spain). Although our solutions could be applied to any type of data, we have validated them using these data because they come from a real environment of great interest to society.

In Section 3.1, the context of this problem is detailed, introducing the involved data sets to perform the subsequent experimentation. In Section 3.2, we describe the modeling of the problem, introducing the linguistic variables designed through fuzzy sets from the data. Section 3.3 describes the meta-heuristic approaches designed to address the problem of extracting LDD from the data, the proposed metrics to assess the fuzzy quantified statements and the measures to evaluate the rankings resulting from the execution of the meta-heuristic algorithms. In Section 3.4, we describe the results of the performed experiments. Firstly, we describe a random baseline designed to prove if the application of meta-heuristic approaches is justified in this scenario. Secondly, we present the results of our meta-heuristic approaches with our two data sets for the extraction of unary and binary fuzzy quantified sentences, and finally we also show a validation performed by an expert meteorologist to assess the adequacy of the judgments generated in the field of meteorology.

The contents of this chapter are extracted from the following publication:

> A. Cascallar-Fuentes[a], A. Ramos-Soto[a], A. Bugarín[a], *Meta-heuristics for generation of linguistic descriptions of weather data: experimental comparison of two approaches*. Fuzzy Sets and Systems, 2022. ISSN: 0165-0114. The publication is available at `https://doi.org/10.1016/j.fss.2022.02.016`.
>
> [a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

3.1 Linguistic Descriptions of Data in weather observation and forecasting

In meteorology, the current weather situation recorded from weather stations consists of a large set of numbers. On the other hand, prediction models are used to calculate the expected behavior in the coming days. These situations are composed of complex numerical models and large amounts of data that are not easily understandable, especially for users. Even for experts, analyzing these data is a challenging task, considering the data features, such as the high number of variables and values usually described.

An example of this difficulty is shown in Figure 3.1, where the three maps represent real

meteorological information about three variables (sky state, wind and temperature) for the 313 municipalities of Galicia. Both the sky and wind icons are the standard ones used by the Galician Meteorology Agency [154] (MeteoGalicia) whereas in the temperature we mapped the possible values to the following labels: VL (very low), L (low), N (normal), H (high), VH (very high) printed in different colors from red, associated to high temperatures, to dark blue, associated to low temperature values. Even though one map per variable is generated for easier comprehension, extracting information from them is practically impossible because of the number of icons and the difficulty of interpreting some of them, especially those related to wind whose values graphic representation are not easily distinguishable.



(a) Sky state (b) Wind (c) Temperature

Figure 3.1: Example of maps with meteorological information showing the difficulty of interpreting its contents.

MeteoGalicia provides, among others, two different related types of data: real-time data from 313 Galician municipalities (Figure 3.2) and numerical forecast data from a 4 kilometers grid which covers the entire Galician territory and its adjacent regions, having a total of 3,363 locations (see Figure 3.3). These services have very different data update frequencies: the observation service offers data collected in real time with a very high update frequency, every 10 minutes, so the effectiveness of the search method is very important (i.e., it should produce very high-scored descriptions in near real time). On the other side, prediction data has a 12-hour update frequency. Therefore, in this case, the time requirements are not as demanding as in the observation case, and, therefore, the effectiveness of the search method is not so relevant. From the data point of view, the main difference between real-time observation data and numerical forecast data is the data sets size, since real-time observation data is available from 313 locations whereas the prediction data provides information from a total of 3,363

locations (one order of magnitude larger). This is a critical feature as the size of the data set affects the efficiency of the linguistic description search process since evaluating the truth degree of a description involves evaluating, for each location, the degree to which it meets a given condition i.e. the computational cost of evaluating a quantified sentence increases with the number of locations to be checked.

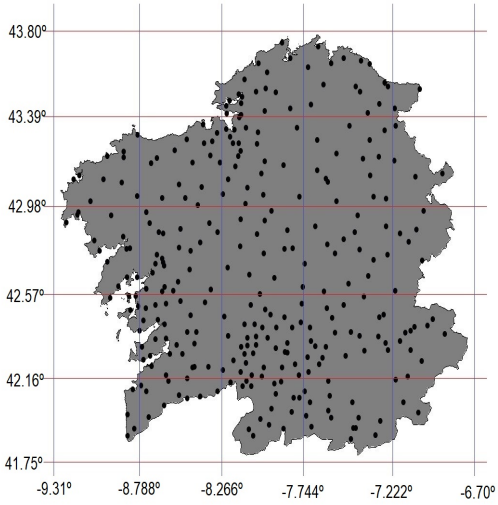


Figure 3.2: Described locations for the weather observation service provided by MeteoGalicia (one observation point per municipality).

Using a greedy approach, the amount of data to be processed in the generation consumes approximately 8 hours in the observation case and over 2 days in the prediction one. [5]. Since this approach is excessively time-consuming, we defined a time frame requisite for each case to ensure the time frame of our approaches is shorter than the data update frequency. Besides, each data source provides the information in a different format: while the observation data is provided as a JSON format, the prediction data is presented in a netCDF file so each data demands a specific processing in order to generate a common input to our approach.

[5] Tests executed on an Intel Core i7-6700HQ @2.60GHz 2.59GHz with 16 GB of RAM

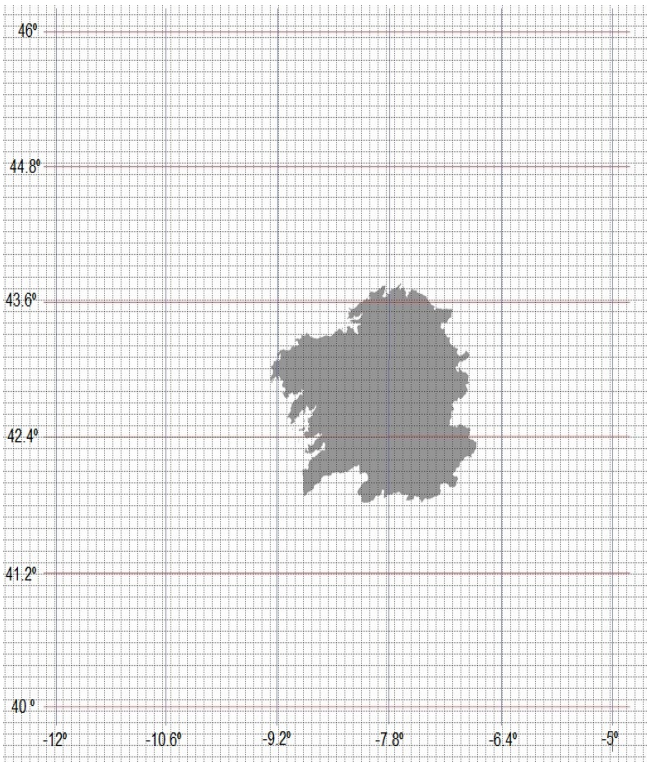


Figure 3.3: Described locations for the weather prediction service provided by MeteoGalicia. Each point is the intersection of the grid lines (being each cell 4km. x 4km. size)

These data sources also differ in the set of described variables and their data type. We select three present meteorological variables (sky state, wind and temperature) and two geographical variables (latitude and longitude) in both sources. Geographical information is defined commonly in both sources. Regarding to the meteorological variables, both sky and temperature do not need a transformation since the sky state is represented with the same 42 codes and the temperature is provided as a numerical value. However, the wind variable did not follow the same format. In the observation case, this variable is represented with 34 codes in the range [299, 322] with defined direction and speed. On the other side, in the prediction case, this variable is described by two complementary variables: *dir*, which represents the direction of the wind in grades and *mod*, which represents the wind speed in m/s. So a pre-processing step is performed in order to get the corresponding code.

From the available historical data, we have constructed a data set composed of 20 randomly collected records comprised between July 2017 and January 2019 for observation, whilst in prediction a data set composed of 12 elements, one for each month of 2018, was built. In both cases, our aim is to generate fuzzy quantified statements that cover the following meteorological variables: sky state, wind and temperature, and the geographical dimension. The temperature does not have a crisp definition that establishes precisely their limits according to how both users and experts use them. Therefore, this arises the necessity of processing it using techniques based on fuzzy logic, so we can generate imprecise expressions.

3.2 Modeling of solutions

To address the problem described above, where the uncertainty in the data needs to be handled, we model our solutions using LDD, namely in form of unary and binary fuzzy quantified sentences. Since unary quantified statements have the structure "*Q Y* are *S*" (e.g., "Almost all locations have high temperature") and binary quantified statements follow the "*Q KY* are *S*" structure (e.g., "Almost all location in the South have high temperature"), it is required to define their components as linguistic variables and their values. Specifically, this section details the definitions of the linguistic variables corresponding to the meteorological variables (state of the sky, wind and temperature), to the geographical descriptors (e.g., "North", "South",...), and the quantifiers (e.g., "Some", "Many",...).

3.2.1 Sky state

This meteorological variable describes the sky situation based on two dimensions: *i)* cloud coverage and *ii)* rainfall. In the official style guide of MeteoGalicia, meteorologists labeled the values of this variable with 42 integer codes (Table 3.1). Half of these values (21 integer numbers in the range [101, 121]) are used to describe the day situation (e.g. code 101 means "clear sky"), using terms such as, for example, "cloudy", "rainy" or "storm with few clouds". Also, the same 21 different situations are described for the night, which are coded with 21 integer numbers in the range [201, 221] are used (e.g. 211 means "night with clear sky"). From this variable, a crisp linguistic variable is defined using its numerical codes as values with a truth degree in the set {0, 1}.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.2.2 Wind

This is a variable that comprises two dimensions: *i)* the wind direction and *ii)* speed. It is labeled with 34 codes, which are integer numbers in the range [299, 332] (see Figure 3.2 to check the official identifiers of the style book of MeteoGalicia). Meteorologists consider eight wind directions (North (N), South (S), East (E), West (W), Northwest (NW), Northeast (NE), Southwest (SW), Southeast (SE)) combined with four wind speed values (weak, moderate, strong, very strong). Furthermore, two special cases are considered: *i)* calm and *ii)* variable direction and speed. For instance, 305 code means "South direction and weak speed". From this variable, a crisp linguistic variable is defined using its numerical codes as values with a truth degree in the set $\{0, 1\}$.

Code	Direction	Speed	Icon
299	Calm	Calm	
300	Variable	Variable	VAR
301	North	Low	
302	North-east	Low	
303	East	Low	
304	South-east	Low	
305	South	Low	
306	South-west	Low	
307	West	Low	
308	South-east	Low	
309	North	Moderate	
310	North-east	Moderate	
311	East	Moderate	
312	South-east	Moderate	
313	South	Moderate	
314	South-west	Moderate	
315	West	Moderate	
316	North-west	Moderate	

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

317	North	Strong	
318	North-east	Strong	
319	East	Strong	
320	South-east	Strong	
321	South	Strong	
322	South-west	Strong	
323	West	Strong	
324	North-west	Strong	
325	North	Very strong	
326	North-east	Very strong	
327	East	Very strong	
328	South-east	Very strong	
329	South	Very strong	
330	South-west	Very strong	
331	West	Very strong	
332	North-west	Very strong	

Table 3.2: List of wind codes with their respective icons defined in the MeteoGalicia style book.

3.2.3 Temperature

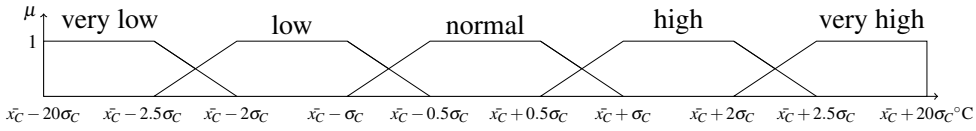
This is a numerical variable that represents the temperature in degrees Celsius. It is modeled as a fuzzy linguistic variable with five linguistic values defined as trapezoidal strong fuzzy partitions [202]: {“very low”, “low”, “normal”, “high”, “very high”}. According to the meteorologists criteria, information related to temperature is always provided taking a reference value for comparison (e.g., normal temperatures are those that are similar to the reference one), thus, to design the linguistic values of this variable, we need to define reference values.

USC UNIVERSIDADE

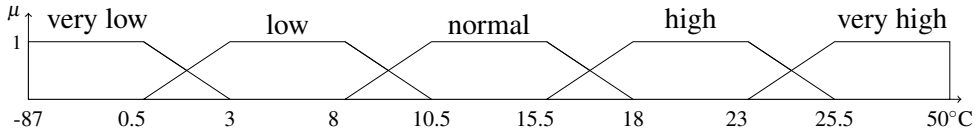
Label	Icon
Very low	VL
Low	L
Normal	N
High	H
Very high	VH

Table 3.3: Values of the fuzzy linguistic variable "temperature".

In this case, historical data is available both for observation and prediction, so we used this information as a reference. For the observation case, historical data is available for each municipality and each month of the year for the last twenty years (this is called "climatic temperature"). Therefore, the historical average temperature $\bar{x}_C$ and its standard deviation σ_C are taken as a reference to define the linguistic terms as stated in Figure 3.4a. For instance, the label "normal" is defined with the trapezoid with support $[\bar{x}_C - \sigma_C,\ \bar{x}_C + \sigma_C]$ and core $[\bar{x}_C - 0.5\sigma_C,\ \bar{x}_C + 0.5\sigma_C]$. So, for a specific point with, for example, $\bar{x}_C = 14.2$ and $\sigma_C = 4.8$, its "normal" label has support [9.4, 19] and core [11.8, 16.6]. For the prediction case, the historical daily data is accessible so, from this available data, a data set which covers all months of a year is generated, taking as a reference the average of the temperatures of this data set $\bar{x}_D$, and its standard deviation σ_D for defining the linguistic terms as stated in Figure 3.4b. For instance, the label "normal" is defined with the trapezoid with support [8, 18] and core [10.5, 15.5], since $\bar{x}_D = 13°C$ and $\sigma_D = 5°C$ for the data set available.



(a) Observation use case.



(b) Prediction use case ($\bar{x}_D = 13°C$ and $\sigma_D = 5°C$).

Figure 3.4: Definition of the linguistic terms of temperature.

3.2.4 Quantifiers

As described in Chapter 1, Linguistic Descriptions of Data (LDD) include quantified statements as "In some locations the wind has North direction and moderate speed", so quantifiers are necessary to count the amount of locations in the map (territorial coverage) that satisfy a given meteorological condition. Seven fuzzy quantifiers {"None", "A few", "Some", "About half", "Many", "Almost all", "All"} are defined as trapezoidal strong fuzzy partitions [202] in terms of the percentage of points that fulfill a condition as shown in Figure 3.5. Moreover, from their definition we defined the concept of "coverage" [63] which represents the percentage of locations it covers, having "None" the minimum because it covers 0% of the territory whereas the maximum is assigned to "All", which covers 100%.

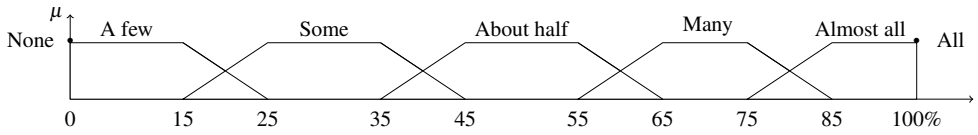


Figure 3.5: Observation and prediction quantifiers definition for the whole territorial coverage (percentage of locations in the map).

3.2.5 Geographical linguistic variables

In the meteorology field, the geographical dimension is very important due to the very nature of the data, i.e., nearby locations within a region with common meteorological conditions are often grouped together when referring to them. Thus, in this case, we construct binary fuzzy quantified statements, which consider the geographical dimension describing specific areas of the territory. These statements, as described in Chapter 1, have the structure "*Q KY* are *S*" where an additional qualifier *K* is included. In this scenario, a geographical descriptor is included as a qualifier, which allows us to describe smaller regions instead of the whole territory, as with the unary descriptions. For instance, "In some locations in the North the temperature is low".

Nine linguistic term descriptors were defined using longitude and latitude as the universe of discourse: "North", "South", "East", "West", "Center", "North-east", "North-west", "South-east" and "South-west". Figure 3.6 shows the location of the territory regarding its meridians and parallels. These geographical descriptors are defined as trapezoidal strong

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

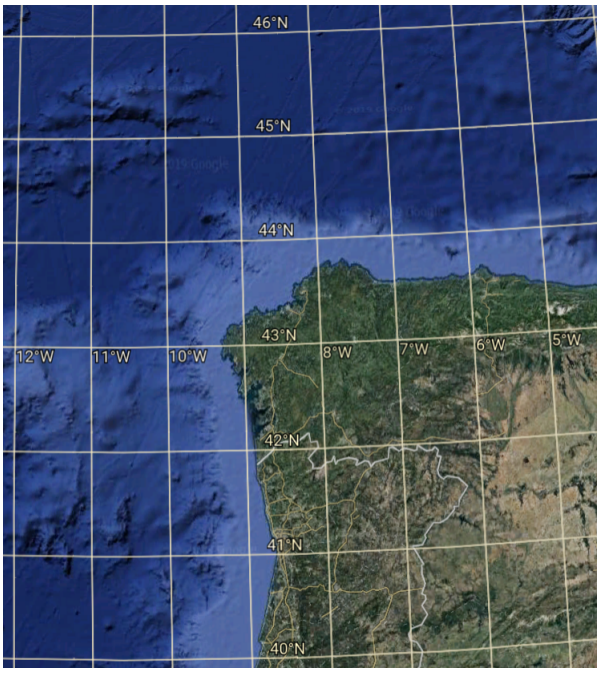
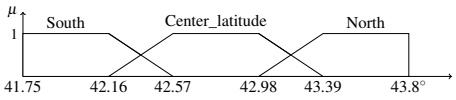


Figure 3.6: Reference coordinates from Google Earth [92].

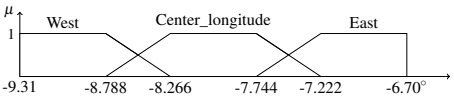
fuzzy partitions [202] with different values in observation (Figure 3.7) and in prediction (Figure 3.8). Figures 3.2 and 3.3 present the representation of the descriptors to show graphically the differences between the definitions. The reference values are adapted to cover only the Galician territory in the observation, whereas in prediction the descriptors cover larger pieces of territory, since forecast data cover a larger extension.

Geographical descriptors can be classified as simple or composite depending on the number of dimensions they are defined upon (one or two). Simple descriptors are defined through one geographical dimension, longitude or latitude. For instance, "North" is defined in terms of the latitude. On the other side, composite descriptors are defined as *AND* combinations of two one-dimensional descriptors for longitude and latitude. For instance, "Center" is defined by combining "Center_latitude" and "Center_longitude". When evaluating the position of a point regarding the "Center" descriptor, "Center_latitude" and "Center_longitude" are evaluated independently and their result is combined using the minimum operator. The rest of the composite geographical descriptors ("Northeast", "Northwest", "Southeast", "Southwest")

are defined in the same way.

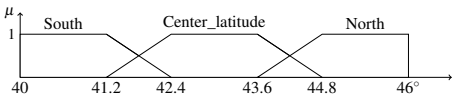


(a) Latitude definition.

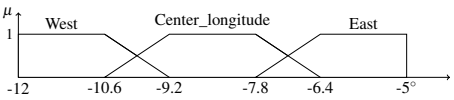


(b) Longitude definition.

Figure 3.7: Geographical descriptors in the observation scenario.



(a) Latitude definition.



(b) Longitude definition.

Figure 3.8: Geographical descriptors in the prediction scenario.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Code	Description	Icon
101	Clear sky	
102	High clouds	
103	Clouds and clear	
104	Very cloudy	
105	Covered sky	
106	Fog	
107	Rain shower	
108	Rain shower 75%	
109	Snow shower	
110	Dew	
111	Rain	
112	Snow	
113	Storm	
114	Mist	
115	Fog banks	
116	Mid-level clouds	
117	Light rain	
118	Light showers	
119	Storm with few clouds	
120	Sleet	
121	Hail	

(a) Codes for the day situation.

Code	Description	Icon
201	Night with clear sky	
202	Night with high clouds	
203	Night with clouds and clear	
204	Night with very cloudy sky	
205	Night with covered sky	
206	Night with fog	
207	Night with rain shower	
208	Night with rain shower 75%	
209	Night with snow shower	
210	Night with dew	
211	Night with rain	
212	Night with snow	
213	Night with storm	
214	Night with mist	
215	Night with fog banks	
216	Night with mid-level clouds	
217	Night with light rain	
218	Night with light showers	
219	Night with storm with few clouds	
220	Night with sleet	
221	Night with hail	

(b) Codes for the night situation

Table 3.1: Sky state codes with their respective icons defined in the MeteoGalicia's style book.

3.3 Proposal of meta-heuristic approaches for LDD extraction

Our approach is focused on the content determination and realization stages of the usual NLG pipeline [189]. For each meta-heuristic solution proposed, Figure 3.9 shows the experimentation stages with their inputs and outputs, described as follows:

- Description components: the input to this stage is the meteorological data with the format described in Section 3.1 and the reference values to define the linguistic variables. The output of this step are the values of the linguistic variables (Q, G, S, W, T).

- Description generation: in the generation process, a description is represented as a tuple of elements. For instance, "all North 101 303 normal" is the intermediate representation of the description "In all locations in the North the sky is clear, the wind has East direction and low speed and the temperature is normal". In this step, we generate new descriptions as candidate solutions where each of their components is selected randomly.

- Candidate evaluation: at this stage we evaluate the candidate description to accept, and add it to the solutions list if accepted. We store all the generated candidates in a set of candidates so, first, we check the first evaluation is checking if a new candidate has already been evaluated. If a candidate has been previously generated, it is discarded to avoid evaluating already assessed descriptions. The output of this stage is the set of accepted solutions verbalizing them with their corresponding unary or binary structure through the templates defined in Table 3.4.

Linguistic variable	**Template**
Sky state	the sky state is <value>
Wind	wind direction is <direction_value> and its speed is <speed_value>
Temperature	the temperature is <value>

Table 3.4: Template to describe each meteorological variable in the generated descriptions.

3.3.1 Meta-heuristic approaches used for candidate selection

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

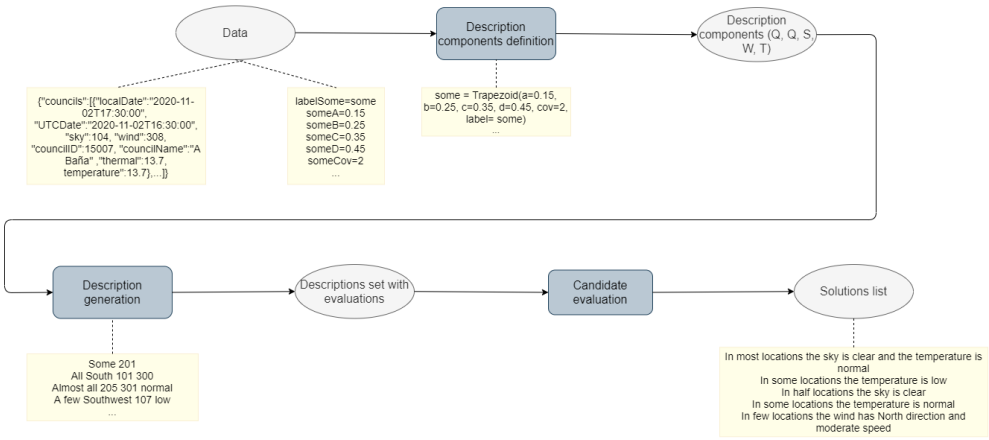


Figure 3.9: Description of the experimentation stages with their inputs and outputs with examples.

Simulated Annealing (SA)

SA [224] is a meta-heuristic search and optimization technique inspired by annealing in metallurgy. To avoid staying in local optimal, it accepts movements to worse solutions, decreasing the probability of accepting these movements as the search advances. This algorithm uses a variable called temperature, whose value determines to what extent candidate solutions worse than the current one can be accepted. This variable is initialized with a high value and is reduced in each iteration by a cooling mechanism which decreases progressively reducing the probability of acceptance.

In each iteration a number of candidate descriptions are generated, applying the acceptance criteria for each one to check if it replaces the current one. If the candidate description is better, it is automatically accepted, while if it is worse, there is still a probability that it will replace the current one. This probability depends on the cost difference between the current and the candidate description and on the temperature, so that the lower the cost difference and the higher the temperature, the higher the probability of acceptance. Once an iteration is finished, the temperature is cooled down and the next one is passed on until the stop criterion is reached.

The selection of this algorithm to develop this approach is its achievement of better results than other algorithms for several tasks [136, 215], although it was not previously applied to

USC UNIVERSIDADE

generate linguistic descriptions. We use the classic definition of the algorithm as a baseline. To apply this algorithm to our solution, we designed an experiment to define its design parameters with different values with the aim of finding a proper configuration. A preliminary phase of this experimentation was described in [33], which has evolved significantly during the development of this thesis.

- S_0 is the initial solution of the algorithm, designed in the same way for both observation and prediction. First, we aimed to generate a solution based on knowledge: we selected the most repeated value for each linguistic meteorological variable in the data and we generated all possible descriptions involving the entire set of defined quantifiers and geographical descriptors. These statements were sorted by coverage and truth value and, among them, we selected the best description that exceeds a pre-defined truth value threshold. If none of the descriptions satisfied these requirements, we generated a random initial solution.

- T_0: the initial temperature was also designed in the same way in both scenarios. To set its value high enough, we initialized it with a value inversely proportional to the maximum number of different unary and binary fuzzy quantified statements (the value of the variable *MAX_DESCRIPTIONS* is 63,203 for unary descriptions and 568,827 for binary sentences). The other two designed attributes are the probability ϕ that a new solution $\mu\%$, worse than the initial one S_0 is accepted. These values allow initializing T_0 as stated in Equation 3.1.

$$T_0 = -\frac{\mu}{\ln(\phi) * MAX_DESCRIPTIONS} \tag{3.1}$$

- New candidate solution: a new candidate is generated by applying some changes to the current solution. Any of the elements in the sentence (quantifier, geographical descriptor, weather variables) can be changed randomly. However, since the algorithm could stay at a local maximum due to generating too many repeated candidates, we limited the number of repeated candidates in one iteration. In the observation case, we limited it to 20 whereas in prediction to 10 because in this case SA tends to generate solutions that were previously generated, thus consuming too much time in this step. If this threshold is reached, a totally random candidate is generated.

- Cooling speed and strategy: we defined the limit of maximum candidates generated and the maximum number of accepted candidates so when one of these values is reached,

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

the temperature is decreased following a Cauchy cooling strategy. We set both limits to 350 and 35 respectively because, after experimenting with different values, we achieved an adequate cooling speed in the way worse solutions are not accepted when the search process is advanced.

- Stop condition: we designed a double stop condition composed by:
  - A maximum number of iterations: 1,000 in the unary generation and 3,000 in the binary generation for both observation and prediction. The execution ends when this threshold is reached.
  - A maximum number of candidates previously generated and evaluated in earlier iterations: 1,000 in the unary generation and 3,000 in the binary generation for both observation and prediction. All generated solutions are stored in a list of candidate solutions so, when we generated a new solution we first checked if it was already generated. The number of previously generated candidate solutions is stored in a counter that increased each time a new solution was previously generated and the execution ends if this threshold was achieved. This condition is designed in order to prevent the execution from getting stuck generating already evaluated descriptions.
- Acceptance criterion: this condition is the same in observation and prediction. SA has a default acceptance mechanism which allows the acceptance of a solution worse than the current one with a certain probability based on the cost difference between solutions (δ). Apart from this usual mechanism, in our design, solutions with a truth value higher than 0 are also accepted in case they also they also meet one of the following conditions:
  - Having an equal or higher truth value than the current solution.
  - Having a higher coverage than the current solution.

Genetic Algorithm (GA)

GA [73] is a population based evolutionary meta-heuristic algorithm based on natural selection. It starts with a random generated population that undergoes modifications along the execution. At each iteration, the Genetic Algorithm selects the best solutions, whose features will be transmitted to their offspring in the next iteration by relying on bio-inspired operators such as crossover, selection, mutation and replacement. To design the Genetic Algorithm is

UGC UNIVERSIDADE

necessary to define the representation of the population. The resulting ranking of descriptions is represented by a chromosome composed of a set of genes that encode a quantified statement.

As we mentioned in Chapter 1, unary fuzzy quantified statements have the structure "*Q Y* are *S*" where *Q* are the quantifiers defined in Section 3.2.4, *Y* is one or a combination of the linguistic variables defined in Sections 3.2.1, 3.2.2 and 3.2.3; and *S* a linguistic value of *Y*. Figure 3.10 shows the representation of the chromosome in the unary quantified statements generation (SS for state sky, W for wind and T for temperature). The information within each gene is the quantifier and the values for each of the meteorological variables.

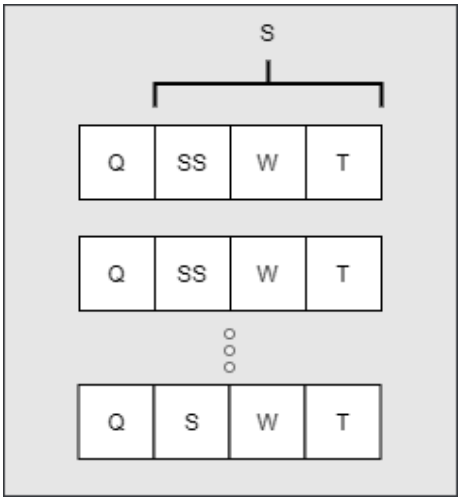


Figure 3.10: Structure of the unary chromosome in the Genetic Algorithm.

On the other side, binary fuzzy quantified descriptions have the structure "*Q KY* are *S*", where *K* are the geographical descriptors defined in Section 3.2.5. Figure 3.11 shows the representation of the binary chromosome, where each gene contains the values of the quantified, the geographical descriptor, and the meteorological variables. To avoid repetition of descriptions, each component of the gene has a fixed place, as represented in Figures 3.10 and 3.11.

For our design we follow a classical definition of a genetic algorithm, testing several options to set the most appropriate parameters values and methods:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Initial population: we decided to generate a fixed set of random descriptions as initial

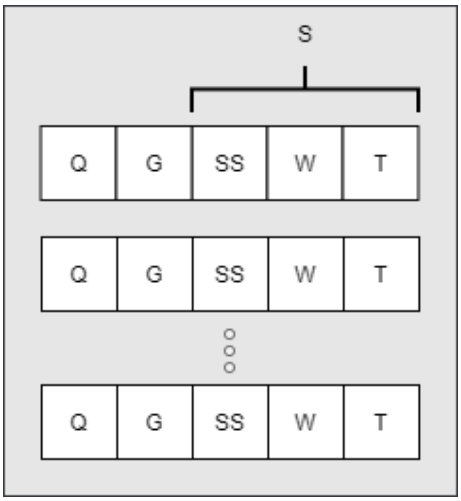


Figure 3.11: Structure of the binary chromosome in the Genetic Algorithm.

population. To define the size of the initial population, we generate all possible descriptions from several meteorological situations analyzing the number of descriptions that we consider of interest in each case in the ranking. We observed in the cases analyzed that the maximum number of representative descriptions was around 200, so we defined this value as the size of the initial solution.

- Tournament phase: regarding the tournament phase, the binary option was selected since is a widely used type and tests with this option reported satisfactory results.

- Crossover: to select the crossover type, we tested with one-point and two-point crossover with percentages between 80% and 98% in pairs analyzing the quality of the descriptions set. We select the two-point crossover 90% because we obtained satisfactory results in terms of the representativeness of the descriptions.

- Mutation: we experimented with percentages between 1% and 20% also in pairs, choosing a mutation of all elements with a probability of 16%.

- Replacement: for the replacement phase we tested with percentages between 5% and 50% in 5% increments, selecting a replacement of the 30%.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Stop condition: we define the stop condition as a fixed number of iterations. To select the number of iterations to finalize the execution, we performed tests with iterations from 500 to 4,000 in five hundred until achieving a balance between the quality of the solutions and the execution time. We set a maximum of 3,000 iterations with observation data and 500 with prediction data.

3.3.2 Fuzzy quantified statements assessment

Once a fuzzy quantified statement is generated, it is necessary to assess how descriptive it is in terms of the meteorological situation described by the data. As described in Chapter 2, the evaluation of LDD is a domain dependent complex task where the selection of the criteria is conditioned by the problem features. When assessing the representativeness of a description regarding a data set, we aim to check the following aspects: *i)* the description gives correct information, *ii)* it contains only the necessary information and no more, *iii)* it contains relevant information and *iv)* it is clear (see conversational Gricean Maxims [94]). Thus, for this purpose, we have considered a combination of three different criteria: *i)* the truth value, *ii)* the coverage, and *iii)* the length of the sentence, as described in what follows.

Truth value

To calculate the truth value ($\mu \in [0,1]$), which is a measure that represents the quantity of elements in a referential which fulfill a given condition, In Section 2.2, the most widely used quantification methods in the literature have been defined. Notwithstanding, as previously discussed, the selection of the most suitable fuzzy quantification method is not an easy task that so far has to be taken into account its theoretical properties, since they have not yet been analyzed from an empirical point of view.

The truth value has a high impact on the performance of meta-heuristic approaches, since this measure conditions their behavior when deciding whether a candidate solution is accepted or not. Therefore, we have considered three fuzzy quantification methods to analyze whether their selection conditions the ranking result generated by each of the meta-heuristic approaches. From the methods discussed in Section 2.2, we have selected *i)* Zadeh's scalar quantification method (Z_Q) [257], *ii)* Delgado's GD method (GD_Q) [54] and *iii)* Zadeh's fuzzy cardinality based method (ZS_Q) [53]. The selection of these methods instead of others was guided by their properties, since these methods are compatible with the defined elements in the design of the proposed fuzzy quantified statements (Section 3.2).

After obtaining the truth value with a fuzzy quantification method, we generated a measure called *fulfil_score*. The decision to define this measure was conditioned by the results observed in preliminary experiments, where we found that in some cases fuzzy quantified statements with a low truth value occupied high positions in the ranking. In order to avoid this undesirable effect, we designed the *fulfil_score* variable, which essentially comprises a mapping of the truth value, aiming to boost descriptions with high truth value, from 0.8 on, and penalize descriptions with low truth value, less or equal to 0.5. The mapping of the *fulfil_score* is the weight function represented in Figure 3.12 which expands the [0, 1] range to a broader one [-1,000, 1,000]. This expanded range was experimentally selected, analyzing the resulting rankings, starting from [-100, 100] until reaching a range where unsuitable descriptions did not occupy the top positions of the ranking.

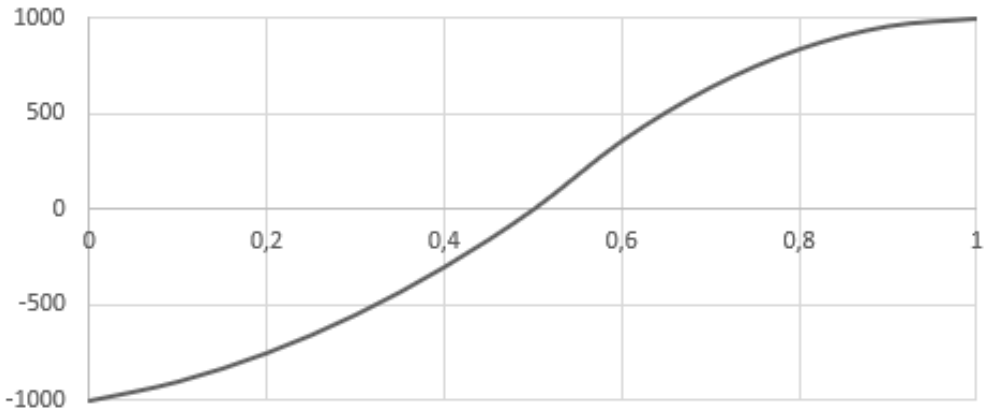


Figure 3.12: *fulfil_score* representation

Coverage

The coverage refers to the extent of land covered by the quantifier, i.e. the number of locations that satisfy the sentence. A fuzzy quantified statement is more suitable the larger the proportion of the land extent it covers as it is more representative of the general meteorological situation. Therefore, arises the need of defining a heuristic that boosts the most representative descriptions and, at the same time, penalizes those descriptions that are not (e.g., “In almost all locations the sky is covered” vs “In a few locations the sky is covered”).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

To this aim, we defined a heuristic mapping that associates a numerical weight to each quantifier, according to the proportion of terrain it covers, so that the higher the proportion covered by a quantifier, the greater the associated weight. We designed the measure *cov_score* which essentially comprises a mapping of the concept of coverage, defined as a value in the range [0, 1] which represents the percentage of locations covered by the quantifier. Also in this case, our purpose generating this measure was to extend the range of the values in order to stressing the differences between low coverage quantifiers and high coverage quantifiers, boosting descriptions involving quantifiers that describe a larger proportion of the territory while penalizing those involving quantifiers that cover less land area. To design this measure, we used as a reference the designed fuzzy set based on the percentage of locations in the map covered by each quantifier (Figure 3.5). We opted for a generic definition for these weights based on the quantifiers definition. Since all quantifiers are defined as trapezoidal fuzzy sets by four points *a, b, c, d*, the numerical weights were initially defined as the average of the points *a* and *d* which define its support. From these values, we performed an empirical study adjusting them until achieving a suitable ranking of quantified statements for a specific situation. In Table 3.5 the assigned coverage to each quantifier is presented.

Quantifier	**Coverage**
None	0
A few	0.17
Some	0.34
About half	0.5
Many	0.67
Almost all	0.84
All	1

Table 3.5: Coverage values for each defined quantifier.

So, with these coverage values, we calculate the *cov_score* measure. Also, in this case, we set negative values for low coverage values under 0.5 and, from the quantifier "Some" on-wards, the assigned coverage score is higher than 0. So, *cov_score* expands the coverage values from the range [0, 1] to [-1,000, 1,260] as represented in Figure 3.13 of the paper. As in the previous case, the selection of the range resulted from an empirical study, starting from [-100, 100] until reaching a range where unsuitable descriptions did not occupy the top positions of the ranking.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

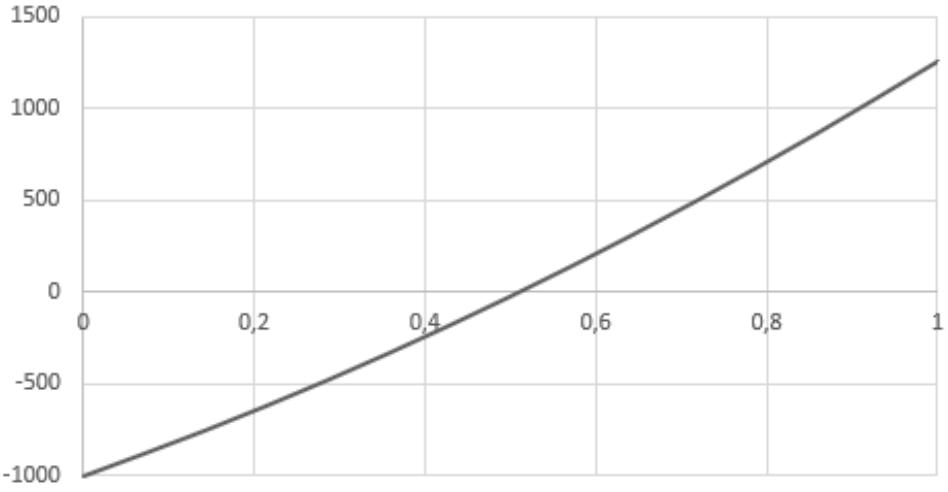


Figure 3.13: *cov_score* representation

Sentence length

The last fuzzy quantified descriptions feature is the sentence length, preferring longer sentences in terms of the number of meteorological variables contained in them (sky, wind and temperature). We normalized this value in the range [0, 1] called *length_score*, which is calculated by dividing the number of variables by the maximum number of meteorological variables a description can contain (3, since we are considering sky state, wind and temperature). Therefore, a full length score (1) will be obtained when a sentence provides a description that includes all variables.

Resulting score

With these three described measures, truth value, coverage and sentence length, we defined the composite measure *score* as follows:

$$score = fulfil_score + cov_score + length_score \tag{3.2}$$

In the related linguistic descriptions literature [26, 57, 182], it is shown that both the truth degree and the coverage are more relevant when it comes to ordering descriptions according

to how informative they are. Therefore, in our approach, we have decided to assign a greater weight to both variables. However, for those cases where two descriptions have equal or similar values for truth degree and coverage, the sentence length will break the tie between the scores however negligible it may seem beforehand.

3.3.3 Evaluation of the resulting rankings

Once we obtain the resulting ranking of a meta-heuristic approach composed of unary or binary fuzzy quantified statements, its adequacy must be evaluated. To test these results, we set as a baseline a greedy approach which generates all possible different fuzzy quantified statements from the combination of the elements that make them up (Q, G, S, W, T) after a full search throughout the search space, getting a total of 632,030 quantified statements (63,203 unary descriptions and 568,827 binary sentences). This set with all possible quantified statements, AS from now on and with cardinality $|AS|$, is used as a baseline to design the quality evaluation measures of the resulting raking of descriptions of each meta-heuristic approach MH (which represents Simulated Annealing or the Genetic Algorithm and with cardinality $|MH|$).

Accuracy

First of all, we defined the accuracy measure which calculates the percentage of sentences generated by the meta-heuristic approach that matches with the set formed by all possible descriptions above the threshold.

$$accuracy = \frac{|AS \cap MH|}{|AS|} \tag{3.3}$$

Jaccard index

This measure [186] calculates the percentage of the matches between the sets (sets intersection) with regard to all different sentences generated by joining them (sets union).

$$union\ intersection = \frac{|AS \cap MH|}{|AS \cup MH|} \tag{3.4}$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Quality of the solutions

Since meta-heuristic algorithms usually do not generate all the descriptions in AS, the number of descriptions that have not been generated and its position in the ranking provides a quality measure for the meta-heuristics, as defined as follows:

$$quality = 1 - \left(\frac{\sum_{i=0}^{n-1} score_i}{|AS|} \right) \tag{3.5}$$

being n the size of the set $AS \setminus MH$ and $score_i$ the score of each description of that set. This quality measure consists of calculating the relative cardinality (using Zadeh's cardinality operator [54]) of the number of descriptions generated by the meta-heuristics that are not included in the full set of descriptions ($AS \setminus MH$). This cardinality is applied to *score*, which is the normalized score described in Equation 3.2. The cardinality value is subtracted from 1 in order to define a quality measure (higher values mean better behavior).

Execution time

For each meta-heuristic algorithm, we collected its execution time in seconds [6]. For binary descriptions, we designed two versions of both approaches: *i)* a sequential and *ii)* a parallel version in order to analyze if improvements could be achieved. We found out that most of the computational resources were consumed by loops in the sequential version, so those parts were optimized in the parallel version in order to improve its performance.

Statistical significance tests

We performed statistical significance tests to check if there are significant differences between our two approaches when generating unary and binary quantified descriptions or not. For this purpose, we compared the accuracy obtained by each algorithm (Table 3.6 shows an example).

To compare them, we tested both approaches with 20 different observation data sets and 12 different prediction data sets. Since in our case the data sets we used include weather data in the same geographical area, we selected cases separated temporally by over 15 days, which we consider independent. The basis for this temporal separation is the average number of days that forecast services cover. For instance, MeteoGalicia only provides prediction data for 8 days. Other weather agencies show similar temporal windows, e.g., AccuWeather [1]

[6] Tests executed on an Intel Core i7-6700HQ @2.60GHz 2.59GHz with 16 GB of RAM

Data set	Algorithm	
	SA	GA
A	40	100
B	57	100
C	100	100
D	100	100
E	75	75

Table 3.6: Example of input data where the numerical data means the percentage of accuracy of our approaches when processing the available data sets.

shows five days, Windguru [242] eight days, and Meteosat [155] nine days. Thus, we assumed 15 days as a safe condition for ensuring data independence, so we have 20 independent observations in the observation use case, whereas we have 12 independent observations in the prediction scenario.

For each case, observation and prediction, we performed statistical tests twice: *i)* for unary and *ii)* for binary descriptions comparing if they achieve or not the same accuracy. We do not have previous information about the behavior of these algorithms in this case, but since both of them were used in similar use cases, we assumed there are no differences between their performances:

H_0: **There are not significant differences between the two compared meta-heuristic algorithms used for generating unary and binary fuzzy quantified statements from observation and prediction data sets**.

To determine which is the most adequate statistical test for this purpose, we needed first to analyze the features of our data, performing the three following tests: *i)* independence of the variables, *ii)* normal distribution and *iii)* heterocedasticity. First of all, regarding to the compared variables, SA and GA accuracy, it can be assured that they are independent since the execution of one algorithm does not depend on the execution of the other. To analyze the data distribution and test if they follow a normal distribution, we analyzed their skewness and kurtosis and we performed the Jarque-Beras test. To analyze these values, we previously defined the following thresholds:

- Since a total skewed data has a skewness of 0, we consider the data has a normal skewness if this value is in the range [-1.5, 1.5]. In particular, if the skewness is in the range [-0.5, 0.5] we consider the data is slightly skewed, and if the value is in the ranges [-1.5, -0.5] and [0.5, 1.5] we consider the data is moderately skewed.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- We accept as normal kurtosis a value in the range [2, 4] since a normal distributed data as a kurtosis of 3.
- To interpret the results of the Jarque-Bera test, we set an $\alpha = 0.05$. Since this test follow a chi-square distribution with 2 Degrees of Freedom (DF), we consider the data satisfies the null hypothesis (H_0: *the data is normally distributed*) if the statistic has a value lower than 5.991 and the p-value is higher than α.

Table 3.7 contains the results of the normality test. Only one data set, GA generating unary quantified descriptions with observation data with the ZS_Q method, does not follow a normal distribution. The results with this data set do not satisfy the mentioned conditions for none of the metrics. All other data sets have a skewness value in the pre-defined range. Regarding the kurtosis measures, some data sets have values lower than 2, however, they have their values of skweness and Jarque-Bera test satisfies the conditions, so we consider they follow a normal distribution. Except the mentioned data set above, all cases have an $\alpha > 0.05$ for the Jarque-Bera test.

Finally, we checked if this data satisfies the homocedasticity condition by applying the Breusch-Pagan test, which allows to detect heteroscedasticity in the data. The null hypothesis of this test is H_0*: homocedasticity exists*. To interpret the results of this test, we set an $\alpha = 0.05$. This test follows a chi-squared distribution with 1 degree of freedom, so we consider the data satisfies the H_0 if the statistic is lower than 3.841. In Table 3.8 we present the results of this test applied to the different analyzed cases. In this case, we can affirm our data satisfies the homocedasticity condition since the statistic value is lower than 3.841 and the p-value is greater than 0.05 in all cases.

After analyzing the results described above, the most adequate test for the data sets which satisfy all previous conditions is an unpaired two-sample t-test setting an $\alpha = 0.05$ [206]. For that data set which does not follow a normal distribution, the most appropriate test is the Mann-Whitney-Wilcoxon Test [141, 241] also with an $\alpha = 0.05$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

				Skewness	**Kurtosis**	**Jarque-Bera test**		**Normal**
						statistic	p-value	
Z_Q	Observation	Unary	SA	-0.74	2.32	2.20	0.33	yes
			GA	-1.33	3.07	5.90	0.052	yes
		Binary	SA	-0.13	2.18	0.62	0.73	yes
			GA	-1.10	2.69	4.14	0.13	yes
	Prediction	Unary	SA	-1.01	3.34	0.81	0.67	yes
			GA	-1.07	3.17	2.09	0.35	yes
		Binary	SA	-0.03	1.41	1.26	0.53	yes
			GA	-0.82	2.27	1.60	0.45	yes
GD_Q	Observation	Unary	SA	-0.26	1.56	1.86	0.40	yes
			GA	-1.36	3.24	5.92	0.0519	yes
		Binary	SA	-0.98	2.51	2.74	0.25	yes
			GA	1.30	-0.55	3.24	0.20	yes
	Prediction	Unary	SA	-0.55	2.49	0.73	0.69	yes
			GA	-1.17	2.60	2.82	0.24	yes
		Binary	SA	-0.06	1.64	1.01	0.60	yes
			GA	-1.19	3.47	3.17	0.21	yes
ZS_Q	Observation	Unary	SA	-0.05	1.29	2.33	0.31	yes
			GA	-1.70	4.34	10.57	0.0005	no
		Binary	SA	-0.53	1.81	2.01	0.37	yes
			GA	-0.06	2.17	0.55	0.76	yes
	Prediction	Unary	SA	-0.28	1.96	0.69	0.71	yes
			GA	-0.56	2.14	1.01	0.60	yes
		Binary	SA	-0.30	1.87	0.82	0.66	yes
			GA	-0.16	2.10	0.46	0.80	yes

Table 3.7: Results of skewness, kurtosis and Jarque-Bera test about the compared data.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

			Breusch-Pagan test		**Homocedasticity**
			statistic	p-value	
Z_Q	Observation	Unary	2.08	0.15	yes
		Binary	0.57	0.45	yes
	Prediction	Unary	0.14	0.71	yes
		Binary	1.42	0.23	yes
GD_Q	Observation	Unary	3.76	0.052	yes
		Binary	1.67	0.20	yes
	Prediction	Unary	0.11	0.74	yes
		Binary	1.01	0.31	yes
ZS_Q	Observation	Unary	0.05	0.84	yes
		Binary	0.01	0.92	yes
	Prediction	Unary	0.06	0.81	yes
		Binary	0.21	0.64	yes

Table 3.8: Results of the Breusch-Pagan test for heterocedasticity.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.4 Experimental results

In this section, we detail the experimentation carried out to evaluate the performance of the meta-heuristic solutions we have proposed applied to meteorological observation and forecasting data. We carried out a very rigorous validation, which included an expert-based assessment together with an automatic evaluation through the metrics we devised. This follows the good practices of evaluation within the NLG field, which are mainly focused on humans evaluation but the results of automatic validation are interesting to validate the conclusions obtained.

Specifically, regarding automatic validation, first, we conducted an experiment to determine whether the application of solutions based on meta-heuristic algorithms is justified, comparing our approximations with a random baseline in terms of accuracy and execution time. We also carried out statistical significance tests between this baseline and our approximations. Furthermore, this section details the performance analysis of the solutions we have proposed for both observation and forecasting data in obtaining rankings of unary and binary quantified sentences using the previously defined metrics. Again, we conducted statistical significance tests between the two meta-heuristic solutions. On the other hand, in order to evaluate the adequacy of the rankings obtained, we designed a validation by experts with the collaboration of a meteorologist at the Galician Meteorology Agency (MeteoGalicia) [154].

3.4.1 Baseline approach

In order to have a baseline to compare the proposed meta-heuristics with, a simple strategy (random search) was implemented. Confronting the meta-heuristics with the baseline will provide an experimental basis for showing that meta-heuristic algorithms are more suitable for this problem than simpler approaches. Considering that in this case we can generate 63,203 different unary statements, we executed four random baselines with 30,000, 50,000, 70,000 and 90,000 unary generated descriptions. In the case of binary statements, we can generate 568,203 different statements. Then, we also performed four baselines with respectively 100,000, 300,000, 500,000, and 700,000 generated descriptions.

Table 3.9 shows the summary of accuracy and execution time of the baseline for the unary scenario, both observation and prediction. This baseline obtained a low accuracy in both scenarios (lower than a 60 % on average regardless of the number of generated descriptions).

On the other side, Table 3.10 shows the summary of accuracy and execution time of the

#attempts	**Observation**				**Prediction**			
	Accuracy (%)		Time (s)		Accuracy (%)		Time (s)	
	Average	SD	Average	SD	Average	SD	Average	SD
30,000	36.64	8.59	28.64	5.99	32.68	7.60	282.08	76.31
50,000	38.02	4.76	45.34	1.75	41.10	4.52	490.58	132.45
70,000	45.40	4.65	72.39	15.86	49.44	5.05	663.14	104.68
90,000	50.90	4.13	95.44	17.81	55.30	3.93	876.95	283.85

Table 3.9: Summary of the results of the unary random baseline.

#attempts	**Observation**				**Prediction**			
	Accuracy (%)		Time (s)		Accuracy (%)		Time (s)	
	Average	SD	Average	SD	Average	SD	Average	SD
100,000	39.08	12.49	134.89	51.84	44.51	8.56	1,124.97	385.14
300,000	65.88	10.01	627.89	42.39	79.94	12.17	2,705.69	212.22
500,000	84.73	9.26	897.53	38.05	82.89	8.38	5,480.28	1,363
700,000	91.07	9.11	1,056.75	27.11	84.34	10.31	8,255.75	1,183.52

Table 3.10: Summary of the results of the binary random baseline.

baseline for the binary case, also for the observation and prediction scenarios. This baseline obtained a high average accuracy generating 700,000 attempts, with an accuracy of 91.07% for observation and 84.34% for prediction, on average. Nevertheless, its execution time is very large, with over 17 minutes in observation and over 2 hours in prediction, generating 700,000 attempts.

3.4.2 Observation

First, the results obtained with the real-time data are discussed, firstly for the generation of unary sentences and then binary sentences. For the three quantitative measures (accuracy, Jaccard index and time), their average and Standard Deviation (SD) are provided.

Unary fuzzy quantified statements generation

In Table 3.11 we present the summary of the involved measures for the unary fuzzy quantified statements generation test with observation data where we compare both approaches regarding four measures: *i)* the percentage of fuzzy quantified sentences generated by the approach that match with *AS*, *ii)* the percentage of matches between the descriptions generated by the meta-

heuristic approach and *AS* with regard to those sets union, *iii)* the quality of the solution identifying the position in the *AS* set of the non-generated descriptions by the approach; *iv)* and the time in seconds.

		Accuracy (%)		**Jaccard index (%)**		**Solution quality**	**Time (s)**	
		Average	SD	Average	SD		Average	SD
Z_Q	SA	84.56	2.85	80.56	2.83	0.55	11.43	2.40
	GA	96.11	7.11	96.11	7.11	0.65	1.00	0.00
GD_Q	SA	63.38	30.73	63.38	30.73	0.45	12.23	2.10
	GA	91.05	16.18	91.05	16.18	0.64	1.11	0.00
ZS_Q	SA	67.12	32.00	67.12	32.03	0.47	11.55	2.00
	GA	89.41	22.29	73.43	17.43	0.52	1.00	0.10

Table 3.11: Metrics results from the unary observation analysis.

Regarding the accuracy average, the GA achieved better results than its SA counterpart regardless of the quantification method used in the evaluation. Analyzing their SD, only in the case where GA evaluated the fuzzy quantified statements with Zadeh's quantification method, its results have a higher SD than SA; whereas in all other cases the results of SA present a higher deviation. It is also noticeable that the standard deviation increased considerably with the GD_Q and ZS_Q methods in both approaches.

Analyzing the results of the Jaccard index, GA obtained better results on average than SA with all considered fuzzy quantification methods. For instance, with Z_Q, SA got an 80.56%, which means about a 20% of generated sentences it generated differed from the expected whereas GA got a 96.11%. Both approaches had a low standard deviation, which means most of the results are concentrated on average. Moreover, in terms of quality, GA also achieved better results analyzing the unmatched positions, which means it avoids generating descriptions in lower positions in the ranking than the SA.

The results obtained from the comparison between the fuzzy quantification methods for each of the meta-heuristic solutions show differences in their behavior. Nevertheless, these differences found in the metrics cannot be justified solely by the selection of the method but also by the non-deterministic nature of the algorithms. In Table 3.12, a list of descriptions with its corresponding truth value calculated by each quantification method proves they have a similar behavior evaluating the sentences. Besides, in [32] we proved there are no significant differences between this set of quantification models in the evaluation of unary fuzzy quantified statements with data from 15 different data sets.

	Z_Q	GD_Q	ZS_Q
In almost all locations the sky is covered	1.00	1.00	1.00
In about half locations the sky is covered and the temperature is high	1.00	0.91	0.97
In a few locations the sky is rainy and the temperature is normal	1.00	1.00	1.00
In all locations in the Southeast the sky is covered	1.00	0.99	1.00
In almost all locations in the West the sky is covered and the temperature is high	1.00	0.90	1.00
In all locations in the Northeast the sky is covered	1.00	1.00	1.00
In many locations in the Center the sky is covered and the temperature is high	0.94	0.58	0.58
In almost all locations in the East the temperature is high	0.67	0.44	0.47
In almost all locations in the East the sky is covered	1.00	0.99	1.00
In some locations in the West the wind has North direction and moderate speed	0.94	0.67	0.87

Table 3.12: Example of the truth value obtained by the three quantification methods for a set of unary fuzzy quantified sentences.

Regarding the execution time, results show the selection of the fuzzy quantification model does not affect the performance of the approaches in terms of this feature. Nevertheless, the comparison between the meta-heuristic approaches proves that GA performs better than its counterpart also in terms of execution time. For example, with Z_Q, SA consumed 11.43 seconds on average while GA 1 second, so the genetic approach achieved the best result not only in terms of accuracy and Jaccard index but also in duration.

Results of the statistical tests are summarized in Table 3.14. As mentioned above, to interpret the test results, we set an $\alpha = 0.05$. In this case, we tested the data with a t-test for all cases except for the unary descriptions generation with ZS_Q, where we applied the Wilcoxon-Mann-Whitney test because it did not follow a normal distribution. In all cases, the p-value is lower than our α so we can conclude there exist significant differences between our two approaches when generating unary descriptions in terms of accuracy.

Binary fuzzy quantified statements generation

The results of the involved metrics for the generation of binary fuzzy quantified descriptions are presented in Table 3.13.

In this scenario, GA obtained better results of accuracy than SA on average, regardless the quantification model. Remarkably, the accuracy is very similar in both approaches evaluating the descriptions through the ZS_Q method. Nevertheless, as described above, these percentages are also influenced by the algorithms' features, therefore, we cannot conclude that the selection of the fuzzy quantification method has a high impact on the results.

USC UNIVERSIDAD

Analyzing the Jaccard index, it is noticeable that even when the performance seems very similar, GA generated better solutions, whereas SA generated a higher percentage of sentences different from the expected. For instance, with Zadeh's fuzzy quantification method, SA obtained a 72.81% whereas GA achieved a 85.21%. On the basis that this measure calculates the proportion of the matches, the union of the SA set and the *AS* set contains about a 30% of not matched sentences whereas in the Genetic Algorithm case is about a half of Simulated Annealing. Both approaches had a low standard deviation, so the results distribution is very close to the average. If we refer to the score obtained in the solution quality metric, on top of getting worse accuracy and Jaccard index, SA also achieved a lower score in this metric. Therefore, its quality can be considered worse in general.

Both SA and GA performed worse than in the unary fuzzy quantified descriptions. This can be caused by the larger amount of binary sentences that can be generated. In the binary scenario, it is easier to remain within a local maximum, generating many undesirable sentences and increasing the execution time as a result.

Comparing the execution time, it seems not to be influenced by the selection of the quantification model. Also, in this case, GA performs better in terms of execution time, consuming, on average, less than 30 seconds than SA. In this case, the execution duration is not excessive as the observation data set is not too large. The parallel version reduced the execution time considerably in SA but even so, GA achieved better results.

In this case, all binary data satisfied the conditions to apply a t-test. As in the unary case, the results of the t-test with 38 DF are presented in Table 3.14 where also significant

			Accuracy (%)		**Jaccard index (%)**		**Solution quality**	**Time (s)**	
			Average	SD	Average	SD		Average	SD
Z_Q	SA	sequential	72.81	3.29	72.81	3.29	0.54	39.80	3.05
		parallel						26.63	2.32
	GA	sequential	95.02	3.71	85.21	3.76	0.77	4.75	0.89
		parallel						4.71	0.72
GD_Q	SA	sequential	82.37	15.18	82.37	15.18	0.62	39.00	2.73
		parallel						27.03	2.23
	GA	sequential	96.32	4.96	87.81	8.52	0.89	5.82	1.20
		parallel						4.83	0.68
ZS_Q	SA	sequential	80.55	9.92	70.88	5.65	0.59	37.80	5.56
		parallel						28.20	3.14
	GA	sequential	80.63	7.37	80.55	9.92	0.60	4.95	1.73
		parallel						4.81	1.22

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Table 3.13: Metrics results from the binary observation analysis.

differences between the two compared algorithms are detected, since the p-value is lower than α.

		Test	Statistic	DF	p-value
Z_Q	Unary	t-test	-2.5106	38	0.0160
	Binary	t-test	-11.9910	38	$1.742e^{-14}$
GD_Q	Unary	t-test	-3.3659	36	0.0018
	Binary	t-test	-3.8068	36	0.0005
ZS_Q	Unary	Wilcoxon-Mann-Whitney	86	-	0.002964
	Binary	t-test	-0.027581	33	0.9781

Table 3.14: Summary of the results of the statistical tests applied to the observation case.

3.4.3 Prediction

In this section, results obtained generating fuzzy quantified statements from prediction data are discussed.

Unary fuzzy quantified statements generation

In Table 3.15 we present the summary of the involved measures for the unary fuzzy quantified statements generation test with prediction data where we compare both approaches regarding the four considered measures.

		Accuracy (%)		Jaccard index (%)		Solution quality	Time (s)	
		Average	SD	Average	SD		Average	SD
Z_Q	SA	73.26	22.97	73.26	22.97	0.60	126.71	2.43
	GA	74.58	29.15	74.58	29.15	0.66	36.34	3.35
GD_Q	SA	59.25	27.89	59.25	27.89	0.56	182.32	5.57
	GA	78.41	35.94	78.41	35.94	0.67	42.36	3.97
ZS_Q	SA	72.08	18.55	70.92	17.61	0.66	194.10	4.55
	GA	85.50	14.68	82.58	15.34	0.68	50.07	4.06

Table 3.15: Metrics results from the unary prediction analysis.

Both algorithms achieved similar results evaluating descriptions with Z_Q, with an accuracy average of 73.26% obtained by SA and a 74.58% obtained by GA. With GD_Q, SA experienced a drop in accuracy of up to 59.25% whereas GA had an accuracy of 78.41% on average,

increasing the difference between the two approaches in terms of accuracy. On the other side, with ZS_Q method, SA achieved a 72.08% whereas GA a 85.5%.

In this case, both approaches obtained the same average for the Jaccard index and accuracy metrics except for ZS_Q method. This can be possible if the set of results from each approach is contained in the *AS* set, but at the same time it does not contain all the most relevant sentences in the highest tier of the ranking. In terms of their standard deviations, SA results present a 22.97% whereas GA a 29.15% evaluating descriptions with Z_Q whilst it decreased with ZS_Q method, where SA obtained a 18.55% and GA 14.68%. Unlike in the observation case, data are not as concentrated on average. In the solution quality metric, our approaches also achieved similar results, but the Genetic Algorithm result is slightly higher, performing better than SA in terms of the solutions quality.

Regarding the results of the significance statistical test, with Z_Q and GD_Q method a t-test with 22 DF was performed for unary descriptions where no significant differences were detected by the statistical test between the two meta-heuristic algorithms. Also with GD_Q and ZS_Q methods, the statistical significance tests between our two approaches did not find significant differences These results are presented in Table 3.17.

Regarding the execution time, in the prediction scenario, our meta-heuristic approaches need more time to converge than in the observation case due to the amount of data they have to deal with. SA needs about 2 minutes, whereas GA needs about 30 seconds with Z_Q. Nevertheless, with the other methods, the necessary time to complete the execution increased by over 1 minute for SA whereas only a few seconds for GA. This means the prediction data type not only means a challenge in terms of time to our approaches but also in terms of accuracy. This is due to unary descriptions show a general idea of the status, but the covered area is larger than in the observation case. Therefore, low coverage quantifiers have a low score because of our penalization and high coverage values have a low truth value because of the diversity of the area. This means that the best descriptions subset is too small, so generating it is a hard task for our algorithms.

Binary fuzzy quantified statements generation

Results of these experiments, presented in Table 3.16, show that GA achieved better results than SA.

Also in this case, the accuracy and the Jaccard index have the same average, which means the approach set is contained in *AS* set. Regarding the quality of the solutions, both approaches

USC UNIVERSIDADE

			Accuracy (%)		Jaccard index (%)		Solution quality	Time (s)	
			Average	SD	Average	SD		Average	SD
Z_Q	SA	sequential	72.35	13.04	72.35	13.04	0.52	130.61	4.32
		parallel						43.34	3.56
	GA	sequential	75.02	8.71	75.21	8.76	0.56	45.33	3.24
		parallel						13.0	2.42
GD_Q	SA	sequential	70.81	12.59	70.70	12.60	0.56	719.63	2.32
		parallel						212.88	2.04
	GA	sequential	83.35	14.46	85.51	22.04	0.67	362.43	2.17
		parallel						171.96	1.88
ZS_Q	SA	sequential	77.73	6.78	71.90	12.55	0.42	790.23	7.78
		parallel						322.33	5.88
	GA	sequential	87.81	5.92	85.97	6.98	0.49	481.44	5.64
		parallel						227.60	3.57

Table 3.16: Metrics results from the binary prediction analysis.

achieved similar results although the Genetic Algorithm score is better than the Simulated Annealing also in this case. With regard to the t-test (Table 3.17), only with Z_Q no differences were detected between our two approaches. Nevertheless, both with GD_Q and ZS_Q methods, the t-test found significant differences between the two approaches generating binary descriptions.

		Test	Statistic	DF	p-value
Z_Q	Unary	t-test	-0.3311	22	0.7437
	Binary	t-test	-1.0244	22	0.3168
GD_Q	Unary	t-test	-1.4596	22	0.1585
	Binary	t-test	-2.3585	24	0.0268
ZS_Q	Unary	t-test	-1.9647	22	0.06221
	Binary	t-test	-3.8794	22	0.0008

Table 3.17: Summary of the results of the t-test applied to the prediction data.

Regarding the execution time, SA needed more time than GA both in their sequential and in their parallel version for the three quantification methods. Besides, the performance of the approaches seems to be conditioned by the quantification model, since the execution time increased with GD_Q and ZS_Q models.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

		Observation	Prediction
		p-value	p-value
30,000	SA	$7.88e^{-13}$	$5.65e^{-05}$
	GA	$< 2.2e^{-16}$	$< 2.2e^{-16}$
50,000	SA	$1.94e^{-12}$	0.001
	GA	$< 2.2e^{-16}$	$< 2.2e^{-16}$
70,000	SA	$2.24e^{-10}$	0.03
	GA	$< 2.2e^{-16}$	$< 2.2e^{-16}$
90,000	SA	$8.68e^{-09}$	0.017
	GA	$< 2.2e^{-16}$	$< 2.2e^{-16}$

Table 3.18: Summary of the results of the statistical tests between the baseline and the meta-heuristic approaches for the unary case.

3.4.4 Comparison between random baseline and meta-heuristic approaches

We performed statistical significance tests with an $\alpha = 0.05$ to compare the accuracy of this baseline with our two meta-heuristic approaches. Table 3.18 shows the results for the unary scenario, both for observation and prediction. Results show significant differences between the two approaches in terms of accuracy in all cases, so these results justify the proposal of this type of algorithms for this use case. The baseline obtained an average accuracy lower than 60% both for observation and prediction, whilst our meta-heuristic approaches achieved an average accuracy higher than 70%.

		Observation	Prediction
		p-value	p-value
100,000	SA	$1.4e^{-08}$	0.002
	GA	$< 2.2e^{-16}$	0.01
300,000	SA	0.60	0.001
	GA	$3.91e^{-08}$	0.41
500,000	SA	$1.05e^{-07}$	$9.28e^{-05}$
	GA	$1.47e^{-12}$	0.25
700,000	SA	$2.51e^{-10}$	$6.30e^{-05}$
	GA	$6.94e^{-08}$	0.19

Table 3.19: Summary of the results of the statistical tests between the baseline and the meta-heuristic approaches for the binary case.

On the other side, Table 3.19 shows the results of the statistical significance tests for the binary case also for both observation and prediction. In this case, no significant differences are detected between GA and the baseline in several cases in terms of accuracy since the baseline. However, if we analyze the execution time, the choice of GA over the random baseline is justified since its execution time is too long with more than 17 minutes in observation and over 2 hours in prediction generating 700,000 attempts whilst GA needs less than 5 seconds in the observation case and less than 44 seconds in the prediction scenario.

3.4.5 Human Expert validation

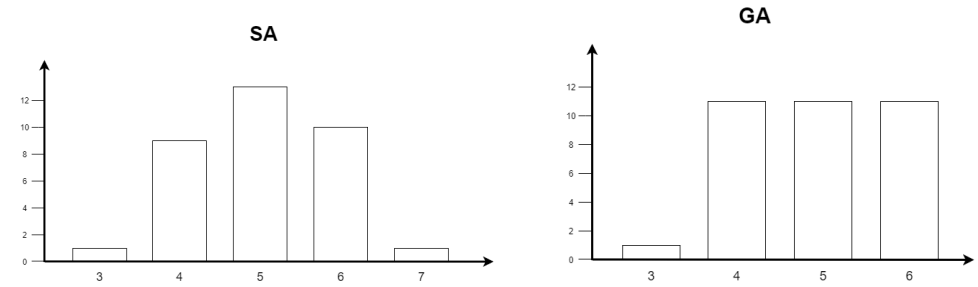
In the automatic validation described in previous sections, we evaluated the quality of the results of the meta-heuristic algorithms in terms of the descriptions position in the ranking. It was a metric-based assessment where the adequacy of these descriptions to the domain problem was not considered. Good practices in the validation of systems in the NLG field require both automated and manual validation by experts in the field, the results of which are complementary.

Therefore, we also asked an expert meteorologist who has extensive experience in the generation of meteorological forecasts from the Galician Meteorological Agency (Meteo-Galicia [154]) to assess the quality of the descriptions in this domain. We forwarded him a questionnaire with a variety of meteorological situations, asking him to evaluate the suitability of the unary and binary fuzzy quantified statements to describe the different situations. We have only counted on one meteorologist since we have collaborated with him on several occasions. In addition, it is important to note that in real environments it is not possible to have a large number of experts and their time availability is also limited.

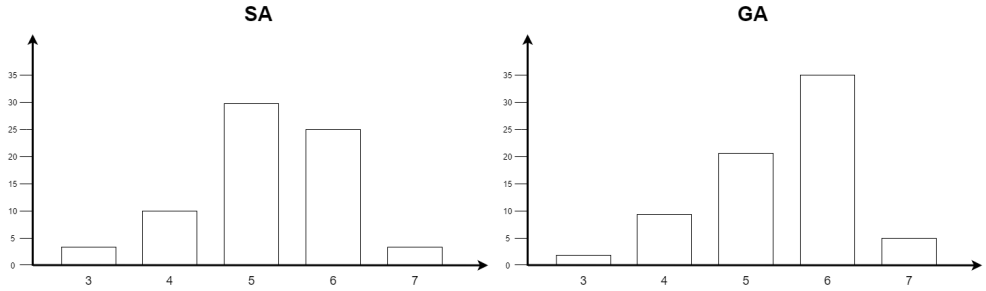
The expert rated a set of 30 different meteorological cases composed by a map described by eight unary and binary quantified descriptions which potentially described the case. A fully blind assessment was performed, since no details were provided about what algorithm generated the descriptions and how these were generated. In fact, the expert was not aware that the descriptions were automatically generated. For each case, the expert rated suitability of the descriptions using a 7-point Likert scale in the range [1, 7] where 1 means “the description is absolutely unsuitable” and 7 “the description is absolutely suitable”. According to the literature, using seven values in the assessment scale maximizes reliability, validity, and discriminative power [222].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

The scores, whose distribution is represented in Figure 3.14, show that no description was



(a) Unary scores distribution.



(b) Binary scores distribution.

Figure 3.14: Distribution of the expert scores separated by algorithm and type of quantified statement.

rated with the values 1 or 2 of the scale. The median of the scores of the entire set of quantified statements is 5, whereas the interquartile range is 1 (average of 5.22 and standard deviation of 0.91). Furthermore, it was also never the case that all statements of a case were rated with a score of 4 or less. Therefore, these results prove that these sentence structures are suitable for describing meteorological situations according to the expert.

Regarding unary statements, 32.35% of the descriptions were rated with a score of 4 or lower, whereas 17.61% of the binary statements were rated with these scores. Besides, also 32.35% of the unary descriptions were rated with a score of 6 or more while 47.89% of the binary sentences obtained these scores. It is also noticeable the expert did not rate the entire set of descriptions associated with a specific case with 4 or less. Analyzing these results, we can conclude binary descriptions are more suitable than unary statements to describe a meteorological situation since the unary have a median of 5 with an interquartile range of 2

USC UNIVERSIDADE

(score average of 4.99 and standard deviation of 0.89) whereas the median of binary is also 5 with an interquartile range of 1 (average of 5.33 and standard deviation of 0.90).

Analyzing the data by approach, the median of the Simulated Annealing score is 5 with an interquartile range of 2 (average of 5.03 and standard deviation of 0.9) in the unary scenario and it has a median of 5 and an interquartile range of 1 (average of 5.21 and standard deviation of 0.90) in the binary case. On the other hand, Genetic Algorithm have a median of 5 with an interquartile range of 1 (average of 4.94 and standard deviation of 0.89) in the unary case and a median of 6 with an interquartile range of 1 in binary scenario (average of 5.45 and standard deviation of 0.90). These values indicate the descriptions quality is similar for both approaches.

Unary score distributions are very similar in both of our two approaches. Only 1.47% of the statements for the two approaches were rated with 3. It is noticeable 1.47% of the SA statements were rated with 7, whereas none of the descriptions generated by GA reached that score. The remaining descriptions for GA were uniformly distributed in the range [4, 6] whereas the SA are more concentrated in 5.

Binary scores are also similar between the two meta-heuristic algorithms. Here, 2.11% of the statements were rated with 3 for SA, whereas 1.41% of the GA descriptions obtained this score. Besides, also 2.11% of the SA sentences were rated with 7 while 3.52% of the GA statements achieved this score. The 20.42% of the SA statements are mainly rated with 5 whereas the GA descriptions were mostly rated as 6, with a 24.65%.

The results of the expert evaluation do not reveal a difference between our two approaches in terms of quality in the unary scenario, with a similar distribution of their scores. Nevertheless, in the binary case, although the rating distribution is also rather similar, the GA percentage of descriptions rated with 6 and 7 is higher than for SA, so we can conclude the set of GA binary descriptions is more suitable to describe a meteorological situation.

3.5 Discussion and final remarks

In this chapter, we assessed the quality of the linguistic descriptions of data automatically generated for fuzzy quantified unary and binary protoforms using two meta-heuristic algorithms for three fuzzy quantification models. In an experimental setting made up by real meteorological observation and prediction data corresponding to several situations for the sky state, wind and temperature variables, we evaluated the statistical significance of the differences between

UCC UNIVERSIDADE

the adequacy of the linguistic descriptions generated by the meta-heuristics. These sources allowed us to test both meta-heuristics and models in different ways: with real-time observation data, we tested the performance of our algorithms in terms of execution time, since in this case this was the critical dimension; on the other side, with prediction data, we tested their behavior with big amounts of data. The meta-heuristics considered were *i)* a solution based on the Simulated Annealing (SA), which was not previously applied to linguistic descriptions generation and could be used as a lightweight alternative; *ii)* a Genetic Algorithm (GA), already used in the literature for linguistic descriptions generation in other realms. The fuzzy quantification models considered were Zadeh's scalar quantification model [257], Delgado's GD quantification method [54] and Zadeh's fuzzy cardinality method (ZS_Q) [53].

Empirical validation was conducted, indicating that, overall, the Genetic Algorithm performs better than Simulated Annealing in the involved quality metrics and in terms of execution time in the performed experiments obtaining the most representative set of unary and binary fuzzy quantified descriptions. The statistical significance of this out-performance depends on the following design factors: the type of meteorological data, the quantified statements structure and the quantification model selected.

First of all, regarding the type of meteorological data, results show that for smaller data set sizes GA performs significantly better than SA in terms of quality of the descriptions, thus being GA clearly the best choice. For larger data sizes the differences in performance are not significant is many cases. Therefore, the search algorithm selection should be guided by other metrics, such as execution time, where GA is also better than SA.

Regarding the structure of the fuzzy quantified descriptions seems, for the unary scenario, it does not to have an impact in the performance of the meta-heuristic algorithms. Therefore, the meta-heuristic selection should be guided by other dimensions such as data size. For the binary case, we found significant differences between GA and SA in many cases with GA performing significantly better than SA. These results indicate that in the generation of binary fuzzy quantified sentences, the fuzzy quantification method conditions the performance of meta-heuristic algorithms.

In the comparison of the quantification methods, results show that in unary the selection of this method does not seem to have an impact in the performance of the meta-heuristic approaches since, as can be seen in Table 3.20, GA performs significantly better than SA with observation data regardless of the quantification method and no significant differences between the two algorithms were detected for either algorithm in prediction data. Neverthe-

UDC UNIVERSIDADE

less, in the binary scenario the selection of the quantification method has an impact on the performance of the meta-heuristic algorithms. With observation data significant differences are detected between them for the Zadeh and GD methods while for prediction differences are detected with the GD and ZS methods. Thus, selection of the fuzzy quantification method should be made considering this.

		Observation	**Prediction**
Unary	Zadeh	✓	✗
	GD	✓	✗
	ZS	✓	✗
Binary	Zadeh	✓	✗
	GD	✓	✓
	ZS	✗	✓

Table 3.20: Summary of the significance difference detection between the GA and SA meta-heuristic where ✓ means "significant difference between GA and SA" and ✗ means "no significant differences detected between GA and SA". Differences are always in favor of GA.

Also human expert assessment on the adequacy of the descriptions was conducted, showing that both meta-heuristics behave similarly for unary descriptions. For binary linguistic descriptions, human expert evaluation indicates that the adequacy of the descriptions generated by the Genetic Algorithm are more suitable than the ones generated by Simulated Annealing.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CHAPTER 4

EMPIRICAL STUDY OF THE BEHAVIOR OF FUZZY QUANTIFICATION METHODS FOR LINGUISTIC DESCRIPTIONS OF DATA

In Chapter 3, we proposed two meta-heuristic approaches to extract Linguistic Descriptions of Data (LDD) from different types of data sets. To assess the generated fuzzy quantified descriptions, we proposed the use of three different fuzzy quantification methods to analyze the impact of its selection on the meta-heuristic algorithms' performance. Notwithstanding, the obtained results did not provide a direct conclusion on the impact of the selection of the fuzzy quantification method. Therefore, in this chapter we detail our proposal for empirically comparing a set of the most widely used fuzzy quantification models in the literature with the aim of experimentally test whether there are significant differences among them in the evaluation of unary and binary fuzzy quantified sentences.

In Section 4.1, we describe the context of these experiments, detailing, first of all, the structure of the involved fuzzy quantified statements. Moreover, we included an analysis of the considered fuzzy quantification methods with regard to the properties they meet, or not, for the unary and binary evaluation. In Section 4.2, we introduce the preliminary statistical correlation study performed between the considered methods, including the details of the used data sets and the study modeling. In Section 4.2.2, we describe the experimental execution of these experiments and its results are discussed. Section 4.3 describes the significance study between the fuzzy quantification methods, which is an extension of the previous study,

USC UNIVERSIDADE DE SANTIAGO

including the features of the considered data sets and the study modeling, where we stated the null hypothesis of this experimental study. In Sections 4.3.1 and 4.3.2, we describe how the linguistic components are designed and the pipeline of the experimental tests with their inputs and outputs, whose results are discussed in Section 4.3.3 for unary and binary quantified descriptions. Finally, in Section 4.4, this chapter concludes with a discussion of the objectives set and the results obtained in the empirical studies carried out.

The contents of this chapter are extracted from the following publications:

©2021 IEEE. Reprinted, with permission, from C. Heble-Lahera[a], A. Cascallar-Fuentes[b], A. Ramos-Soto[b], A. Bugarín[b], *Empirical Study of Fuzzy Quantification Models for Linguistic Descriptions of Meteorological Data*. In proceedings of the 2020 IEEE International Conference of Fuzzy Systems (FUZZ-IEEE), pp. 1-7, Glasgow (United Kingdom), 2020. ISSN: 1558-4739. The publication is available at https://doi.org/10.1109/FUZZ48607.2020.9177716.

[a]Máster Universitario en Investigación en Inteligencia Artificial, Universidad Internacional Menéndez Pelayo, Spain.
[b]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

Reprinted from Publication ECAI 2020: 24th European Conference on Artificial Intelligence, A. Cascallar-Fuentes[a], A. Ramos-Soto[a], A. Bugarín[a], *An experimental study on the use of fuzzy quantification models for linguistic descriptions of data*. In proceedings of the 24th European Conference of Artificial Intelligence (ECAI), pp. 267-274, Santiago de Compostela (Spain), ©(2020), with permission from IOS Press. ISBN: 978-1-64368-100-9 (print), 978-1-64368-101-6 (online). The publication is available at IOS Press through https://dx.doi.org/10.3233/FAIA325.

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

4.1 Analysis of the considered fuzzy quantification methods

As described in Section 2.2, the evaluation of a quantified statement involves computing its truth degree. In that evaluation, two elements must be considered: *i)* the cardinality, i.e., how many elements in the referential fulfill the (fuzzy) linguistic values stated for the variables in the statement; *ii)* the compatibility between the cardinality and the quantifier. In this section, we present the set of fuzzy quantification methods empirically compared in this study. The methods we analyzed were selected based on their presence in the related literature and

their cardinalities. In fact, the cardinality measure that each method uses directly impacts its performance.

Several studies [10, 51, 53, 60, 62, 226] theoretically analyzed the fuzzy quantification methods by checking the properties they fulfill. This analysis show that, in general, quantifiers should behave (very) differently, although, for unary descriptions and coherent quantifiers, it was theoretically proved [54] that Delgado's GD method [54] is a generalization of Yager's method [245] and Zadeh's fuzzy cardinality based method (ZS) [53] is a generalization of the Sugeno integral base method [22]. A quantifier Q is called coherent (or monotonically non-decreasing quantifiers [245]) if $Q(x_i) \leq Q(x_{i+1}) \forall x_i < x_{i+1}$, $Q(0) = 0$, $Q(1) = 1$ [54]. Nevertheless, unlike other research fields as Machine Learning [76, 77], the behavior of the models has not been studied yet from an experimental point of view, by analyzing the real quantitative differences existing among them.

In the experiments performed in this chapter, we assess the behavior of the considered fuzzy quantification methods evaluating unary and binary fuzzy quantified statements with the structure defined in Chapter 1. Our unary quantified statements follow the standard form previously described (*"Q Y are S"*), being Q is the set of equidistant or random quantifiers and S is the set of linguistic terms for each of the linguistic variables. Likewise, we generated binary quantified statements following the previously described structure *"Q KY are S"*. Once again, Q and S maintain their respective roles from unary quantified statements, and K is another term in the set of linguistic terms of the corresponding linguistic variables.

4.1.1 Properties

A fuzzy quantification method is proposed according to the semantics of a specific approach. In the following, we present a set of properties for the unary and binary quantification methods, which have been proposed to analyze the behavior of these methods. In order to prove that the results of a method are coherent, several authors have proposed some desirable properties it should fulfill.

Properties for unary descriptions

Being an unary fuzzy quantified statement "*Q Y* are *S*", fuzzy quantification methods should fulfill these properties in their evaluation (these rules are extracted from [51]):

I.1. If S is crisp, then if Q is relative, the evaluation must be $Q(|S|/|X|)$, and $Q(|S|)$ if Q is an absolute quantifier.

I.2. Evaluation must be coherent with fuzzy logic for the quantifiers *exists* and *all*, i.e. for evaluating a unary sentence:

$$\exists(S) = \bigvee_{y_i \in Y} Y(y_i) \; and \; \forall(S) = \bigwedge_{y_i \in Y} Y(y_i) \tag{4.1}$$

where $\vee$ and $\wedge$ are a t-conorm and a t-norm respectively.

I.3. Evaluation must be coherent with quantifiers inclusion, i.e. if $Q \subseteq Q'$ (Q is more restrictive than Q') then $Q(S) \leqslant Q'(S)$.

I.4. Evaluation must be time-efficient, that is, the efficiency should be between $O(n)$ and $O(n \; log \; n)$ where $y = |Y|$.

I.5. Evaluation must be not too "strict", i.e., given Q defined over $H = \{p/q : p \in \{0,...,n\}, q \in \{1,...,n\}\}$ with $Q \neq$ and $Q \neq H$ there should exist a fuzzy set S such that the evaluation of the sentence is not in {0, 1}.

I.6. The evaluation method must allow to use any quantifier (not only coherent quantifiers).

Properties for binary descriptions

Being a binary fuzzy quantified statement with the structure "$Q \; KY$ are S" and $Eval_Q(S/K)$ the evaluation of a binary description, the fuzzy quantification method should fulfill the following properties (these rules are extracted from [51]):

II.1. If S and K are crisp, then the evaluation must be $Q(|S \cap K|/|K|$ where Q is a relative quantifier.

II.2. The evaluation method must be valid for unary descriptions by changing K for Y.

II.3. Evaluation must be time-efficient (between $O(n)$ and $O(n \; log \; n)$).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

II.4. If $K \subseteq S$ and K is a normal fuzzy set, then the evaluation method must return the value 1 because the percentage of K that are S is 100%.

II.5. If $K \cap S =$ then the evaluation method must return the value 0.

II.6. Evaluation must be coherent with fuzzy logic for the quantifiers *exists* and *all* giving:

$$\bigvee_{y_i \in Y} (S(y_i) \wedge K(y_i)) \; and \; \bigwedge_{y_i \in Y} (K(y_i) \rightarrow S(y_i)) \tag{4.2}$$

respectively where $\rightarrow$ is a fuzzy implication and $\vee$, $\wedge$ are a t-conorm and a t-norm.

II.7. Evaluation must allow to use any quantifier.

II.8. Evaluation must be not too "strict", i.e., given Q defined over $H = \{p/q : p \in \{0,...,n\}, q \in \{1,...,n\}\}$ with $Q \neq$ and $Q \neq H$ there should exist fuzzy sets S, K such that the evaluation of the sentence is not in $\{0, 1\}$.

II.9. If $Q \subseteq Q'$ (Q is more restrictive than Q') then $Eval_Q(S/K) \leqslant Eval_{Q'}(S/K)$.

II.10. The evaluation of "Q K are S" must return the same value that the evaluation of "Q *of* K *are* $S \cap K$", i.e., $Eval_Q(A/D) = Eval_Q((A \cap D)/D)$.

II.11. Let Q be a quantifier and let *ant* $Q(y) = Q(1-y)$ be its antonym. Let $\neg S(y) = 1 - S(y)$. Then $Eval_Q(S/K) = Eval_{ant\ Q}(\neg S/K)$, i.e., the evaluation of "Q of K are S" must return the same value than the evaluation of "*ant* Q of K are $\neg S$".

II.12. $Eval_Q(S/K) = 1 - Eval_{\neg Q}(S/K)$ where $\neg Q(y) = 1 - Q(y)$, i.e., the evaluation of "Q of K are S" must return the negation of the evaluation of "$\neg Q$ of K are S".

II.13. De Morgan's laws must be satisfied:

- $Eval_\forall(S/K) = Eval_\exists(\neg S/K)$.
- $Eval_\exists(S/K) = Eval_\forall(\neg S/K)$.

An evaluation method verifying properties **II.11** and **II.12** also verifies this property.

USC UNIVERSIDADE

4.1.2 Unary fuzzy quantification methods

In this section, we describe the unary fuzzy quantification methods selected for this comparison in terms of the properties they fulfill, since the formulas for evaluating the fuzzy descriptions have been detailed in Section 2.2.1. Table 4.1 extracted from [51] summarizes the fulfillment of unary properties by these quantification methods:

- Zadeh's method: this method [257] does not satisfy properties **I.2** and **I.5**.
- Yager's method based on OWA operators: the unary quantified sentences with the quantifiers "exists" and "all" are evaluated by t-norm and t-conorm (generally min and max operators). The sentences with other coherent quantifiers are evaluated by OWA operator. The OWA operator provides a result between min and max. This method [245] does not satisfy the property **I.6** since only relative and coherent family of quantifiers are considered.
- Delgado's GD method: the method GD_Q [51, 54] is a general case of Yager's method based on OWA operators and Choquet integral based method. It allows to use any quantifier regardless of whether or not it is coherent. Moreover, the method GD_Q satisfies all the properties from **I.1** to **I.6**.
- Sugeno integral based method: this method [21, 22] satisfies the properties from **I.2** to *I.5*. The property **I.1** is only satisfied when the quantifier is coherent.
- Zadeh's fuzzy cardinality based method: this method [53] satisfies all the properties from **I.1** to **I.6**.

4.1.3 Binary fuzzy quantification methods

In this section, we detail the binary fuzzy quantification methods selected for this comparison in terms of the properties they fulfill. As in the previous case, the formulas for evaluating the fuzzy descriptions have been detailed in Section 2.2.2. Table 4.2, extracted from [51], summarizes the fulfillment of unary properties by these quantification methods:

- Zadeh's method: the method [257] is a general case of the formula used for unary quantified statements. Therefore, the method does not satisfy the property **II.6** and the property **II.8**.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Yager's method based on OWA operators: the method [245] is only valid for the coherent family of quantifiers. This method does not fulfill the properties **II.4**, **II.5** and **II.6**.
- Delgado's GD method: it has been proved that the GD_Q method [54] satisfies all the properties except the property **II.6** for the quantifier "all".
- Vila, Cubero, Medina and Pons' method: this method [226] only satisfies properties **II.3**, **II.6** and **II.7**.
- Zadeh's fuzzy cardinality based method: the method ZS_Q satisfies all the properties from **II.1** to **II.8**.

	Z_Q	S_Q	Y_Q	GD_Q	ZS_Q
I.1	✓	✓*	✓	✓	✓
I.2	X	✓	✓	✓	✓
I.3	✓	✓	✓	✓	✓
I.4	$O(n)$	$O(n\ log\ n)$	$O(n\ log\ n)$	$O(n\ log\ n)$	$O(n\ log\ n)$
I.5	X	✓	✓	✓	✓
I.6	✓	✓**	X	✓	✓

* Property satisfied only if Q is a coherent quantifier
** In conflict with property **I.1**, although the method allows any type of quantifier

Table 4.1: Recompilation of fulfillment of the unary properties from [51] by the involved fuzzy quantification methods.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

	Z_Q	Y_Q	V_Q	GD_Q	ZS_Q
II.1	✓	✓		✓	✓
II.2				✓	✓
II.3	$O(n)$	$O(n\ log\ n)$	$O(n)$	$O(n\ log\ n)$	$O(n\ log\ n)$
II.4	✓			✓	✓
II.5	✓	X		✓	✓
II.6 ($\exists$)		X	✓	✓	✓
II.6 ($\forall$)		✓	✓	?	✓
II.7	✓		✓	✓	✓
II.8		✓		✓	✓
II.9				✓	
II.10				✓	
II.11				✓*	
II.12				✓	
II.13				✓*	

* Means that property is satisfied when K is crisp

Table 4.2: Recompilation of the fulfillmenting of the binary properties from [51] by the involved fuzzy quantification methods (blank spaces mean no information in the paper).

4.2 Study of the correlation between fuzzy quantification methods

So far, the fuzzy quantification methods considered have been analyzed taking into account only the theoretical properties that they fulfill, or not. In this chapter, we propose to analyze these methods empirically to check if there are differences in their behavior or not. First of all, in this section, we introduce the statistical correlation study between the selected fuzzy quantification methods in the evaluation of unary and binary statements.

4.2.1 Design of linguistic descriptions of data

This statistical correlation study was performed using real data from the real-time observation service provided by the Galician Meteorology Agency (MeteoGalicia) previously described in Section 3.1.

From these data, also as in Chapter 3, we described the three following meteorological variables: *i)* sky, *ii)* wind and *iii)* temperature, defining the corresponding linguistic variables also as described in that previous chapter:

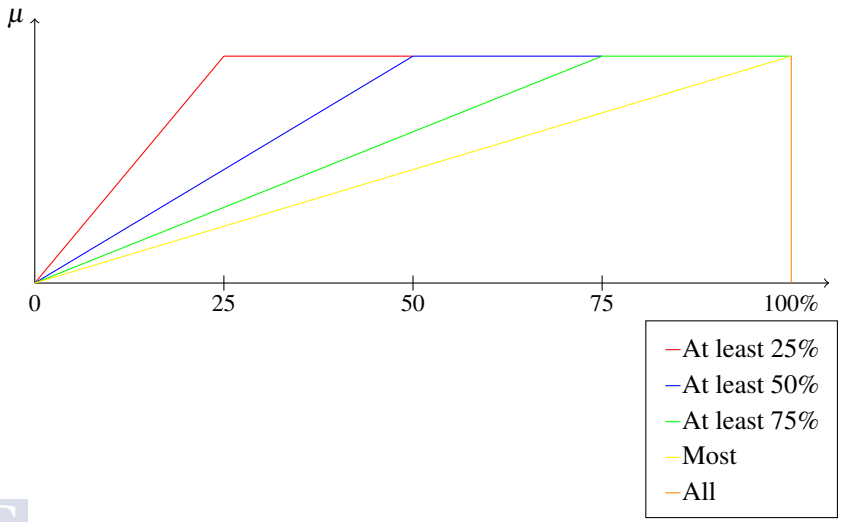
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Sky is treated as a crisp variable, therefore the values of the meteorological variable are the values of the resulting linguistic variable. Each value in the meteorological variable is defined as a singleton having an integer code as a label and a degree of truth in the set {0, 1}.

- Wind is also a crisp variable, so the resulting linguistic variable has values in the range [299, 332] defined as singletons. Likewise, each value has a truth degree in the set {0, 1}.

- Temperature is a numerical variable represents the temperature in degrees Celsius. We modeled the linguistic variable as a fuzzy variable with the following five labels: {“very low", “low", “normal", “high", “very high"}, which are defined as fuzzy sets (Figure 3.4a).

Second, to build the fuzzy quantified statements, we defined the five quantifiers (“at least 25%", “at least 50%", “at least 75%", “most", “all"), represented in Figure 4.1. All of them are modeled as coherent since some quantification methods (Yager’s method, Sugeno integral based method, and Vila et al. method) only support quantifiers that fulfill this property.



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Figure 4.1: Quantifiers’ definition of the percentage of locations.

Finally, in the generation of binary quantified descriptions, we defined the nine linguistic geographical descriptors previously described in Section 3.2.5 for the observation case (Figure 3.7). This allows us to describe smaller regions instead of the whole territory.

4.2.2 Statistical correlation experiments

The experimentation we performed in this study consisted mainly of two steps: *i)* generating the linguistic descriptions and *ii)* analyzing the quantification models behavior to evaluate their similarity. Figure 4.2 describes the performed stages in this experimentation and their corresponding inputs and outputs.

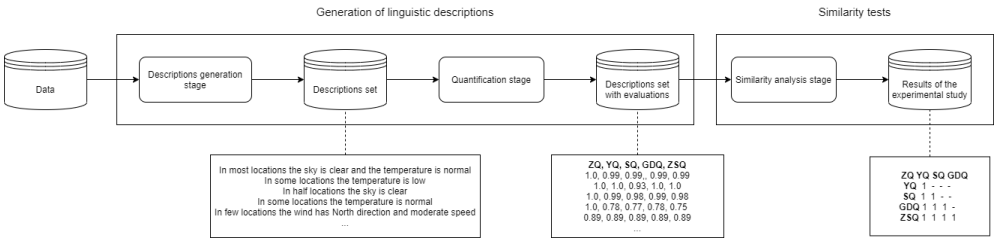


Figure 4.2: Description of the experimentation stages with their inputs and outputs.

Linguistic descriptions generation

In the first stage, we generated all unary and binary descriptions from several real meteorological situations. Firstly, we collected data from 15 different days and times of the day from 30th July 2019 to 30th August 2019. In this study, the data set was narrow because it was a preliminary experiment, which was later extended. We generated all possible descriptions from this resulting data set, obtaining for each meteorological situation 45,145 unary descriptions and 406,305 binary descriptions. For each generated sentence, we assessed how descriptive it was in terms of the meteorological situation described by the data. Thus, at this stage, we evaluated the descriptions with the selected quantification methods.

Quantification methods correlation comparison

We performed correlation tests in order to determine whether the quantification models have similar behaviors when evaluating quantified descriptions. Among the different correlation

tests, we selected the Pearson test, which is applied to the truth degree, giving information about the similarity of quantification methods results, instead of ranking the descriptions based on the associated truth value. We performed this test in two stages: *i)* the test was performed over the data set from 30th July 2019 and *ii)* test was performed for each description over the average result of the evaluation of 15 data sets.

Besides, in this second stage we performed two separate tests: *i)* applying the correlation coefficient test only to those quantified sentences which described the temperature and *ii)* applying test to quantified sentences which describe the three meteorological variables (temperature, sky state and wind). Sky state and wind values were filtered in order to only generate descriptions that described actual situations in the data sets. So descriptions were generated with 5 different sky state values (101, 103, 104, 105 and 111) and with 18 different wind state values (from 300 to 316 and 318).

Experimental results

In this section, we present the obtained results by applying the Pearson correlation test to the different sets of descriptions for unary and binary descriptions. As we mention before, for unary descriptions and coherent quantifiers, it was theoretically proved [54] that Delgado's GD method [54] is a generalization of Yager's method [245] and Zadeh's fuzzy cardinality based method (ZS) [53] is a generalization of the Sugeno integral base method [22]. Therefore, in order to keep consistence with the theoretically proved result, we state the null hypothesis H_0 as follows:

H_0: **There are not significant differences between the two compared fuzzy quantification models when used for evaluating the truth degree of fuzzy quantified statements.**

With regard to binary quantified statements, there are no studies that compare theoretically or experimentally the behavior between a pair of fuzzy quantification models. In the absence of previous studies, we kept the same null hypothesis H_0 defined in the unary quantified statements scenario.

Unary fuzzy quantified descriptions

First of all, we performed an experiment with data from the 30th July 2019, the results of which are summarized in Table 4.3. These results confirm there is a very high correlation between the quantification methods, with correlations between 0.9925 and 1. A total correlation

is obtained between the pairs Y_Q - GD_Q and S_Q - ZS_Q, which is consistent with [54], where it is stated that these methods have the same behavior for coherent quantifiers.

	Z_Q	Y_Q	S_Q	GD_Q	ZS_Q
Z_Q	1	0.9995	0.9925	0.9995	0.9925
Y_Q		1	0.9945	1	0.9945
S_Q			1	0.9945	1
GD_Q				1	0.9945
ZS_Q					1

Table 4.3: Pearson correlation coefficient for quantification methods evaluating unary descriptions from 30 July 2019.

To perform the next experiments, we generated a data set from the average of data sets corresponding to 15 different dates from 30th July 2019 to 30th August 2019, averaging the compliance values associated with the same quantification method for each associated with the same quantification method for each of the sentences. With this resulting data set, we first performed an experiment describing only the temperature variable, since it is the only fuzzy linguistic variable defined from the meteorological variables. The reason for this experiment is to avoid any possible bias caused by the inclusion of crisp variables. In Table 4.4, we present the results of the Pearson test with this resulting average data set describing only the temperature situation. Also in this case, the correlation coefficient between all quantification methods pairs is very high and the total correlation between the pairs Y_Q - GD_Q and S_Q - ZS_Q was also obtained.

	Z_Q	Y_Q	S_Q	GD_Q	ZS_Q
Z_Q	1	0.9983	0.9987	0.9998	0.9987
Y_Q		1	0.9992	1	0.9992
S_Q			1	0.9992	1
GD_Q				1	0.9992
ZS_Q					1

Table 4.4: Pearson correlation coefficient for quantification methods evaluating unary descriptions from the average data set describing temperature.

Finally, with the resulting average data set, we performed an experiment including the entire set of linguistic variables to check if any bias is detected because of the crisp variables. Table 4.5 shows the correlations in this experiment, including the entire set of described me-

teorological variables. This experiment also confirms the conclusions of the previous tests, therefore, there are no significant differences between the quantification methods when evaluating unary quantified descriptions. Moreover, no biases were detected when including both crisp and fuzzy variables with respect to analyzing only the correlation with fuzzy variables.

	Z_Q	Y_Q	V_Q	GD_Q	ZS_Q
Z_Q	1	0.9999	0.9976	0.9999	0.9976
Y_Q		1	0.9979	1	0.9979
S_Q			1	0.9979	1
GD_Q				1	0.9979
ZS_Q					1

Table 4.5: Pearson correlation coefficient for quantification methods evaluating unary descriptions from the average data set including all meteorological variables.

Analyzing the results of these three performed tests, we can conclude when evaluating unary statements, the compared methods do not show differences between their behaviors since their correlation coefficient is higher than 0.99 between all models pairs. This is consistent with the analysis of properties described in [51] for the unary evaluation. We gather in Table 4.1 a summary of the most relevant properties, where it can be seen that all methods fulfill the six considered properties, with the only exceptions of Z_Q (which does not fulfill properties **I.2** and **I.5**) and Y_Q, which does not fulfill property **I.6**. Finally, the method pairs GD_Q - Y_Q and ZS_Q - S_Q had a correlation coefficient of 1, which is consistent with the theoretical previous work [54] when it is proved that the behavior of these pair of methods are equivalent evaluating coherent quantifiers. This verifies that the tests performed are correct.

Binary fuzzy quantified descriptions

With binary fuzzy quantified statements, we also performed three correlation tests: *i)* with the 30th July 2019 data set, *ii)* with the average data set for the temperature variable and *iii)* with the entire set of variables of the average data set. First of all, results of the Pearson correlation coefficient applied to the 30th July 2019 data set (Table 4.6) of the compared quantification models when evaluating binary quantified descriptions show a high correlation between the methods Z_Q, GD_Q and ZS_Q. However, the correlation between V_Q with the others is lower than 0.85. Also, Y_Q has low correlation coefficient when comparing its behavior with the other four methods.

	Z_Q	Y_Q	V_Q	GD_Q	ZS_Q
Z_Q	1	0.3899	0.7610	0.9862	0.9396
Y_Q		1	0.3330	0.3811	0.3825
V_Q			1	0.7727	0.8338
GD_Q				1	0.9595
ZS_Q					1

Table 4.6: Pearson correlation coefficient for quantification methods evaluating binary descriptions from the 30th July 2019 data set describing temperature.

With regard to the average data set considering only the temperature variable, results presented in Table 4.7, show high correlation between Z_Q, GD_Q, V_Q and ZS_Q. Also in this case, Y_Q has lower correlation with the other methods than in the unary evaluation.

	Z_Q	Y_Q	V_Q	GD_Q	ZS_Q
Z_Q	1	0.7716	0.9694	0.9870	0.9806
Y_Q		1	0.7711	0.7508	0.7514
V_Q			1	0.9834	0.9850
GD_Q				1	0.9935
ZS_Q					1

Table 4.7: Pearson correlation coefficient for quantification methods evaluating binary descriptions from the average data set describing temperature.

Finally, the results of the experiments performed using the entire set of meteorological variables, as performed with the 30th July 2019 data set, show a high correlation between Z_Q, GD_Q and ZS_Q, whereas V_Q and Y_Q have lower correlation with the others. Results are presented in Table 4.8.

	Z_Q	Y_Q	V_Q	GD_Q	ZS_Q
Z_Q	1	0.2686	0.8168	0.9926	0.9759
Y_Q		1	0.2875	0.2547	0.2754
V_Q			1	0.8256	0.8703
GD_Q				1	0.9831
ZS_Q					1

Table 4.8: Pearson correlation coefficient for quantification methods evaluating binary descriptions from the average data set including all meteorological variables.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Moreover, the results of the three different performed tests set show lower correlation be-

tween the analyzed models. V_Q and Y_Q have low correlations with the other methods, whereas Z_Q, GD_Q and ZS_Q maintain a similar behavior. One plausible explanation for the correlation between Z_Q and GD_Q is their type, since both models are sum-based methods. On the other hand, the correlated behavior between GD_Q and ZS_Q can be explained by their cardinality type, since both of them are based on fuzzy cardinalities. Regarding the binary properties described in Table 4.2, the three fuzzy quantification models share properties **II.1**, **II.4**, **II.5** and **II.7**. Besides, GD_Q and ZS_Q also fulfill properties **II.2**, **II.6** ($\exists$) (ZS_Q also fulfills properties **II.6** ($\forall$)), and **II.8**. Regarding Y_Q, from the provided information in [51] this model is the only one which does not fulfill properties **II.5** and **II.6** ($\exists$). This non-compliance could explain the low correlation between this method and the other four quantification methods. On the other hand, V_Q fulfills properties **II.6** $\exists$ and $\forall$; and **II.7**. Therefore, the correlation between this method and Z_Q, GD_Q and ZS_Q could be explained by these shared properties.

4.3 Statistical significance study between fuzzy quantification methods

In the obtained results of the preliminary study described in the previous section, we observed low correlations between some of the quantification methods involved in the evaluation of binary statements. Therefore, our aim is to extend the empirical results, including an experimental significance study to test whether there are significant differences between these fuzzy quantification methods in the evaluation of unary and binary fuzzy quantified sentences. Thus, this study attempts to extend the previously mentioned theoretical studies and the correlation test to determine if significant differences exist among the methods and assess whether the presence or absence of differences justifies the application of one method over the others.

4.3.1 Design of linguistic descriptions of data

Data sets

We have used in the experiments fifteen data sets which have been used for different Artificial Intelligence-related tasks, such as classification, regression or others, and are available in the UCI machine learning repository [66]. The most relevant quantitative features of the data sets we considered are described in Table 4.9. The data sets meet the following conditions:

UFG UNIVERSIDADE

- They have at least two attributes to be able to generate binary descriptions, for which a variable is required for K and for S.
- They have at least one numerical attribute to have at least one fuzzy linguistic variable.
- No large-scale data set were considered.

The number of instances in the collection of 15 data sets included in the experiment is ample, with ranges from 100 to 50,000. The selection of this range of sizes has been made in order to have a sufficient variety of cases so that the results are as representative as possible. It is relevant to note that we are considering larger data sets than the ones reported in the related literature of linguistic descriptions of data using fuzzy quantification (for instance, 282 instances in [5], 513 in [37] or 1,268 in [233, 235]). A pre-processing step was performed for each considered data set, removing those instances which contained missing values.

Data set	**#instances**	**#attributes**			**Associated task**	**Missing values**
		#total	#numerical	#categorical		
Abalone [66]	4,177	8	7	1	Classification	No
Acute inflammations [47]	120	8	1	7	Classification	No
Liver disorders [66]	345	7	6	1	N/A	No
Balance [66]	625	5	4	1	Classification	No
Iris [66]	150	5	4	1	Classification	No
Blood transfusion service center [250]	748	5	4	1	Classification	No
Credit approval [66]	690	16	15	1	Classification	Yes
Wine [66]	178	14	13	1	Classification	No
Breast cancer Wisconsin [17]	699	11	10	1	Classification	Yes
Bank marketing [158]	45,211	17	7	10	Classification	No
Air quality [50]	9,358	14	14	0	Regression	Yes
Airfoil self-noise [66]	1,503	6	6	0	Regression	No
Glass identification [66]	214	11	10	1	Classification	No
Energy efficiency [218]	768	10	10	0	Classification, regression	No
Concrete compressive strength [249]	1,030	9	9	0	Regression	No

Table 4.9: Some relevant features of the data sets used in the experiments

Linguistic variables

The data sets in our study contain both categorical and numerical variables. Consequently, we created linguistic variables for both variable types, so that fuzzy quantified statements containing both types of variables (crisp and fuzzy) could be computed. In the case of categorical variables, the different values or classes are directly taken as the linguistic terms of the corresponding linguistic variable, which were modeled as crisp sets (or singletons). For instance, the attribute *"class"* of the *"Iris"* data set has the following values: {"Iris Setosa", "Iris Versi-

color", “Iris Virginica"}. From this kind of variables, crisp categorical variables were created using their values as linguistic terms.

On the other hand, for numerical variables, linguistic terms were modeled as trapezoidal fuzzy sets. For each numerical variable in a data set, we generated four different linguistic variables, each one containing a different number of linguistic terms. Three of these linguistic variables correspond to a fuzzy partition with equidistant fuzzy sets including, respectively 3, 5 and 7 linguistic terms. The remaining linguistic variable involved one partition with 5 randomly defined fuzzy sets. As in most approaches in LDD, the fuzzy partitions for every linguistic variable were defined as trapezoidal strong fuzzy partitions [202], i.e., for each point the sum of the truth degree is 1. Figure 4.3 shows an example of an equidistant partition, where a single parameter α models the distance between each pair of contiguous fuzzy sets. This parameter is calculated for each numeric variable in a data set as $\alpha = (\max - \min)/(2n-1)$ being $\min, \max$ respectively the minimum and maximum values of the corresponding numerical domain and n the number of terms.

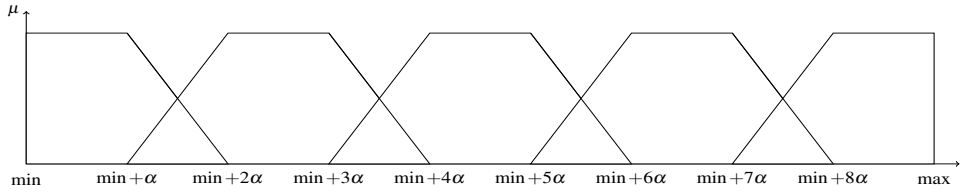


Figure 4.3: Linguistic variable definition from numerical variable with 5 equidistant linguistic terms.

Quantifiers

We selected seven quantifiers {“At least one", “A few", “Some", “About half", “Most", “Almost all", “All"}. As in the previous study, all of them were defined as coherent fuzzy quantifiers, as this was a necessary condition by Yager’s method, Sugeno integral based method, and Vila et al. method. We designed two different fuzzy partitions for these quantifiers: equidistant and random. Figure 4.4 shows the equidistant coherent definition for these quantifiers, since they are defined as monotonically increasing while fulfilling that $Q(0) = 0$ and $Q(1) = 1$.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

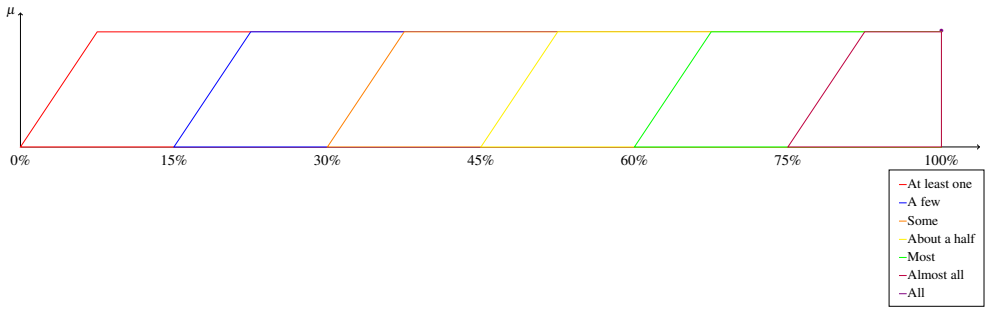


Figure 4.4: Representation of the equidistant definition of the quantifiers.

4.3.2 Experiments

Our objective of this study was to perform a comparison among the selected fuzzy quantification models with the aim of detecting significant differences, if any. This study consisted in two stages (Figure 4.2): *i)* generation of linguistic descriptions, which involves two substages: *i-a)* "descriptions set", where we generated all possible unary and binary fuzzy quantified statements for every fuzzy quantification model and data set and *i-b)* "quantification stage", where we evaluated them with the six fuzzy quantification models; *ii)* based on the lists of resulting fuzzy quantified statements, we run tests of statistical significance to detect significant differences between pairs of fuzzy quantification models.

Linguistic descriptions generation

At this stage, we generated all possible unary and binary fuzzy quantified statements for each considered data set. First, we generated linguistic variables from all attributes in a data set, creating both crisp categorical and fuzzy numerical linguistic variables based on the nature of the source variables. In the generation of binary statements, all variables of a data set have the role of qualifiers K and S with the restriction of $K \neq S$ in a sentence.

The next step of this stage is the evaluation of the statements. For each data set and fuzzy quantification model, we performed two different studies: *i)* with a complete set composed of both categorical and fuzzy linguistic variables; and *ii)* considering only the subset of fuzzy linguistic variables generated from numerical attributes. The reason for these two separate studies lies in the convergent behavior of fuzzy methods when dealing with crisp sets. Thus, we chose to additionally study the same data sets considering only numerical variables (and

consequently only fuzzy linguistic variables) to avoid any possible bias caused by including categorical variables. Table 4.10 shows a summary of the number of experiments we undertook for each empirical study. For binary quantified statements, the study based only on the fuzzy linguistic variables was performed with 14 data sets, since the *Acute inflammations* data set [47] had to be discarded as it only contains one numerical attribute. This means only one fuzzy linguistic variable can be generated and therefore can not be used to produce binary quantified statements, since these require at least two: one for K and another one for S.

Quantifier	**# Linguistic values**	**# Experiments**
7 equidistant	3 equidistant	1
	5 equidistant	1
	7 equidistant	1
	3 random	5
	5 random	5
	7 random	5
7 random	3 equidistant	5
	5 equidistant	5
	7 equidistant	5
Total		33

Table 4.10: Number of experiments performed for each scenario.

The result of this stage for each data set is a pair of lists of all possible statements (one for unary quantified and another one for binary quantified sentences) with the results from the evaluation with the compared fuzzy quantification models. These lists are ordered by their associated truth degree (in descending order). Thus, for a given experiment, the test to determine the possible difference between a pair of fuzzy quantification models consists of a comparison of their corresponding sentence rankings ("all vs. all" comparison, as we will describe in what follows).

Significance tests

We performed a non-parametric test since our data do not meet the following conditions: normality, i.e., the distribution of the data follows a Gaussian distribution and homocedasticity, i.e., the distributions of different groups are equal. Considering the number of the compared fuzzy quantification models, 5 for each description type, and the data sets size, we performed the Friedman test [80] and the post-hoc Nemenyi test [164] since we made an "all vs all"

comparison. According to the Nemenyi post-hoc test, two methods differ significantly when $p < 0.05$ and differ highly significantly when $p < 0.01$. The tests were performed separately for unary and binary quantified statements. Our aim is to empirically extend the results of the theoretical studies and the significance test performed above. Therefore, in order to keep consistence with the theoretically proved result, we state the null hypothesis H_0 as follows:

H_0: **There are not significant differences between the two compared fuzzy quantification models when used for evaluating the truth degree of fuzzy quantified statements.**

With regard to binary quantified statements, there are no studies that compare theoretically or experimentally the behavior between a pair of fuzzy quantification models. In the absence of previous studies and that the results of the correlation study need to be extended to obtain a conclusion, we kept the same null hypothesis H_0 defined in the unary quantified statements scenario.

4.3.3 Experimental results

We have conducted a total number of 957 experiments (495 with the entire linguistic variables set and 462 with the fuzzy linguistic variables set).

Unary fuzzy quantified descriptions

A summary of the results of the tests for unary models is presented in Table 4.11, with all linguistic variables, and in Table 4.12, only for fuzzy variables. These tables show the percentage of experiments where no significant differences were detected between the corresponding pair of fuzzy quantification models. Analyzing the results in Table 4.11, we can conclude they support the null hypothesis and the theoretical affirmations presented in [54] because in almost all experiments, there were no differences between the behavior of the fuzzy quantification models. However, three pairs of methods, Z_Q with S_Q, Y_Q with S_Q, S_Q with GD_Q, showed differences between their behavior in two of five executions of the *"Wine"* data set with random partitions quantifiers. Thus, in general, only in 0.21% experiments significant differences were detected for each pair of methods, which in our opinion is not representative enough.

Analyzing the properties of these pairs of methods in [51] where significant differences were detected, there exists at least one property that is fulfilled by one method of the pair but not by the other: induced operators for the Z_Q - S_Q pair, monotonicity in the quantifiers for S_Q - Y_Q and duality for S_Q - GD_Q. This could be a motivation for the significance of the differences we observed, although this hypothesis needs further study and testing. On the

	Z_Q	Y_Q	S_Q	GD_Q
Y_Q	100			
S_Q	99.57	99.57		
GD_Q	100	100	99.57	
ZS_Q	100	100	100	100

Table 4.11: Percentages of cases where unary quantified models show similar behavior for all the variables (crisp and fuzzy).

	Z_Q	Y_Q	S_Q	GD_Q
Y_Q	100			
S_Q	99.57	99.57		
GD_Q	100	100	99.57	
ZS_Q	100	100	100	99.78

Table 4.12: Percentages of cases where unary quantified models show similar behavior for the subset of fuzzy variables.

other hand, for the pairs of methods where no significant differences were detected, we found that at least one property exists which is fulfilled by one model of each pair, whereas not by the other. Therefore, our study shows these pairs of methods behave similarly when evaluating unary quantified statements, although they should behave differently from the point of view of the properties they fulfill.

Results based only on fuzzy linguistic variables (Table 4.12) show also the detected differences as in the previous case and also between GD_Q and ZS_Q only in one of the 462 fuzzy tests.

Comparing the results of these two experiments, there is not a high dependency between the linguistic variable type and the results, since their results differ in 0.52% of the experiments. Despite these differences, the null hypothesis H_0 is accepted in this case since detecting significant differences in 2 out of 195 tests between Z_Q - S_Q, Y_Q - S_Q and S_Q - GD_Q and 1 difference between GD_Q - ZS_Q in only one data set is not representative enough of their behavior. Finally, the obtained results are consistent with the theoretical results that state that GD_Q [54] and Y_Q [245] methods, on one hand, and ZS_Q [53] and S_Q [22] methods, on the other, coincide for coherent quantifiers. This verifies that the tests performed are correct.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Binary fuzzy quantified descriptions

Results from binary quantified statements tests with the entire set of variables are presented in Table 4.13. In this scenario, the percentage of cases with non-significant differences is lower than 50% of the experiments in almost all pair-wise comparisons, except between Z_Q and GD_Q (72.17% of cases with non-significant differences). Besides, two pairs of methods, Z_Q - V_Q and GD_Q - V_Q, have the lowest percentage of similarity (5.80%), showing therefore the most different behavior in the evaluation of binary quantified statements. This result supports the similarity of Z_Q and GD_Q with binary quantified statements because these two methods, which have a similar behavior evaluating binary quantified statements, have the same behavior with respect to V_Q. In contrast with the previous case, here the null hypothesis H_0 is rejected because in almost all method pairs, significant differences were detected in more than 50% of performed tests, except in the Z_Q - GD_Q comparison.

	Z_Q	Y_Q	V_Q	GD_Q
Y_Q	19.42			
V_Q	5.80	12.75		
GD_Q	72.17	21.16	5.80	
ZS_Q	29.28	41.74	9.28	32.17

Table 4.13: Percentages of cases where binary quantified models show similar behavior for all the variables (crisp and fuzzy).

In Table 4.14, the results of the experiments with the subset of fuzzy variables are presented. As with the total set of variables, the results reflect behavior as described above, hence, the null hypothesis is rejected.

	Z_Q	Y_Q	V_Q	GD_Q
Y_Q	5.38			
V_Q	0.77	11.85		
GD_Q	71.08	5.38	0.77	
ZS_Q	23.08	36.92	3.08	20.00

Table 4.14: Percentages of cases where binary quantified models show similar behavior for the subset of fuzzy variables.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

4.4 Discussion and final remarks

In this chapter, we presented an experimental study to compare the behavior of six well-known quantification methods in the evaluation of unary and binary quantified descriptions. This study is composed of two empirical tests: *i)* a preliminary statistical correlation comparison between the quantification methods applied to meteorological data sets and *ii)* a statistical significance study between the quantification methods, if any. First of all, in the correlation study, we performed our experiments using meteorological data provided by the Galician Meteorology Agency: *i)* data from 30th July 2019 and *ii)* a data set composed by the meteorological data from 15 days between 30th July 2019 and 30th August 2019 (performing experiments only with the temperature descriptions and with the entire set of descriptions). Test results show a high correlation between the behavior of the quantification models in the evaluation of unary descriptions. On the other side, in the binary scenario, correlation between the quantification models is lower than for unary descriptions, but methods can be clustered into three categories according to their correlation: *i)* Z_Q, GD_Q and ZS_Q, *ii)* V_Q and *iii)* Y_Q.

In the significance study, we presented an experimental comparison between the methods' behavior on fifteen data sets from the UCI machine learning repository [66]. We have analyzed experimentally whether there exist significant differences between them when applied to the calculation of the truth degree in fuzzy quantified statements. We designed a test set composed by (equidistant and random partitions) seven quantifiers and both crisp and fuzzy linguistic variables, with three, five and seven (equidistant and random) linguistic terms. Tests results were evaluated with a pair-wise comparison performing statistical significance tests with a null hypothesis H_0 for unary and binary quantified statements which state there are not significant differences between a pair of fuzzy quantification models when calculating the fulfillment degree of the fuzzy quantified statements. H_0 is inspired on and extends the previous theoretical result [54] which show that two of the fuzzy quantification models we have studied ([53] and [54]) are respectively generalizations of [245] and [22] under certain conditions.

The experimentation for unary quantified statements only showed significant differences between three pairs of fuzzy quantification models (Z_Q with S_Q, Y_Q with S_Q and GD_Q with ZS_Q) in 4 of 26 experiments with random partitions quantifiers of one specific data set. Thus, this study confirmed the null hypothesis for 7 pairs of fuzzy unary quantification models in the entire set of experiments. In only 4 of a total 377 experiments for three pairs of fuzzy quantification models, significant differences were actually observed. Therefore, these results

point out that the selection of a fuzzy quantification model for a specific case should follow another criteria than their experimental behavior (e.g., its theoretical properties or computational cost), since from a quantitative point of view, the differences in their empirical behavior are, in general, not significant.

On the contrary, the experiments for binary quantified statements showed significant differences between two fuzzy quantification models at least in a 50% of experiments, except between Z_Q and GD_Q, where only in a 25.13% of experiments significant differences were detected. Therefore, since in 9 pair comparisons the methods behavior is significantly different in more than a 50% of the cases, the null hypothesis H_0 was rejected. Therefore, these results point out that the selection of a fuzzy quantification model for a specific case should be very careful, since from a quantitative point of view, their behavior is significantly different. Other criteria, such as the theoretical properties these models fulfill (or fail to) become more relevant for selecting the most appropriate model for a given use or application.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER 5

FUZZY MODEL FOR THE MANAGEMENT OF THE IMPRECISION IN TEMPORAL KNOWLEDGE

In previous sections, we described the importance of choosing a proper algorithm to explore the space of possible linguistic descriptions and how a specific quantification method may alter the truth degree of the descriptions where we summarize data linguistically. As described earlier in Chapter 1, the work that has been proposed in the literature proposes overly simple solutions that may not be sufficient to meet the requirements of many domains since they are mainly based on unary and binary quantified statements [117, 122, 145, 208, 268]. Furthermore, in many areas, the data have properties that are very important and must be carefully analyzed. This is the case of the temporal dimension in many fields, where it is necessary to contextualize the time series data according to the time at which it is produced and to know the order of the events. For instance, process management [120], decision support [97], investments [110, 114] or health [5, 146, 204, 209]

In order to provide support in describing time series and to address the problem of handling imprecise time information, in this chapter, we propose a Data-To-Text model that relies on fuzzy Logic and Natural Language Generation. Our model includes a fuzzy temporal ontology which provides mechanisms to represent and manage the imprecision of temporal knowledge, which can be used as a baseline to represent temporal and fuzzy information. We applied this model in two different real application areas within transfer projects, which cor-

respond to highly demanding environments: *i)* meteorology, namely Air Quality Index (AQI) time series and *ii)* medicine, in particular, cardiac rehabilitation.

In Section 5.1, we present the Data-To-Text (D2T) model designed for the linguistic description of time series, whose architecture stages are detailed. This model is composed of a temporal model, including the general guidelines that this model follows and the definition of the temporary concepts. Section 5.2 includes the application of this D2T model to AQI time series, including the details of the system modeling to generate the linguistic descriptions. This section also includes the carried out validation, designed in collaboration with expert meteorologists, with the aim of transferring the described system to society. Section 5.3 includes the specification of the application of the D2T model in the medical area within a transfer project, specifically in the area of cardiac rehabilitation, where we described 19 medical indicators of monitoring within this rehabilitation program. This section also includes the details of the validation of the system in collaboration with experts and examples of the resulting linguistic descriptions. Finally, this chapter concludes with a discussion of the results of the application of this model in two real application areas in Section 5.4.

The contents of this chapter are extracted from the following publication:

> A. Cascallar-Fuentes[a], J. Gallego-Fernández[a], A. Ramos-Soto[a], A. Saunders[b], A. Bugarín[a], *Automatic generation of textual descriptions in data-to-text systems using a fuzzy temporal ontology: application in air quality index data series*. Applied Soft Computing, 119, 2022. ISSN: 1568-4946. The publication is available at `https://doi.org/10.1016/j.asoc.2022.108612`.
>
> [a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.
> [b]MeteoGalicia, Xunta de Galicia, Spain.

5.1 Data-To-Text model for linguistically summarizing temporal data

We propose a D2T model using the standard classic pipeline as baseline [188, 192, 193], taking into account that the resulting model should be a sufficiently general tool but contain the necessary guidelines for the construction of a D2T system in any application area.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.1.1 Temporal model

Within this Data-To-Text (D2T) model, we designed a temporal model for addressing the problem of managing the imprecision of the temporal dimension in time series. The aim of this temporal model is grouping the data by the most general and representative temporal expression possible, for example at the level of month or week, and highlight exceptional cases such as a low temperature period, for instance, "During the month of September the temperature was normal except the second week of this month where it was low". First, we formally define the time concepts that make up this temporal model in terms of their properties and duration through an ontology. To define this temporal ontology, dates are denoted by lowercase (*a, b, c...*) whereas dates distribution is denoted by uppercase (*A, B, C...*).

We consider time as projected onto a discrete and sorted axis $\tau = \{t_0, t_1, ..., t_n\}$ where t_0 represents the time origin and each t_k represents a precise instant of time. We consider a total ordered relationship between the precise instants, so, for each t_k and t_{k-1} the difference between them, $\delta = t_k - t_{k-1}$, is constant. Therefore, considering an event as an occurrence with an instantaneous time duration, if two events happen in instants with a temporal distance lower than δ, we consider them simultaneously.

Furthermore, we assume μ_a unimodal, i.e. $\forall t, t', t'' \in \tau, t < t' < t'', \mu_a(t') \geqslant min(\mu_a(t), \mu_a(t_{\prime\prime}))$. This property expresses the convexity of the fuzzy set A. It guarantees that the possible values of a are grouped together. The time concepts (Figure 5.1) that make up the defined ontology are described as follows:

- Instant: A is defined as an instant if and only if exists a single time point $t_0 \mid \mu_A(t_0) = 1$ *and* $\forall\, t_i \neq t_0 : \mu_A(t_i) = 0$. A is therefore defined as a singleton on t_0. Example: t_0="at 1pm".

- Instantaneous event: an occurrence that has happened in an instant. Example: "Temperature was high at 1pm".

- Imprecise interval: the value of a is in the range $[t_0, t_1]$ with a possibility distribution A with equal core and support in this range where $\forall t_i \in [t_0, t_1]\ \mu_a(t_i) = 1$. Example: "Between 1pm and 5pm". In a special type of interval, one of its extremes tends to infinite. For instance, "period of time before Date/Event" or "period of time after Date/Event".

- Imprecise event: event with a date defined as Imprecise interval. Example: "Temperature was high between 1pm and 5pm".

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Fuzzy interval: the value of a is in the range $[t_0, t_k]$ with fuzzy probability distribution A in the range $[0,1]$. Values belonging to the core of A, $core(A) = [t_i, t_j]$, are considered the most possible values. Nevertheless, values belonging to the support of A, $supp(A) = [t_0, t_k]$, but not to its core are also considered ($t_i \leqslant t_j \implies \exists t \in \tau \mu_a(t) = 1$ *and* $core(A) \subseteq supp(A)$). Example: "Afternoon". Figure 5.1 shows an example with a trapezoidal definition.

- Fuzzy event: event with a date defined as Fuzzy interval. Example: "Temperature was high in the afternoon".

- Episode: occurrence that happens in an imprecise or fuzzy time interval, persistent and with a duration d. Example: "Temperature was high during the afternoon".

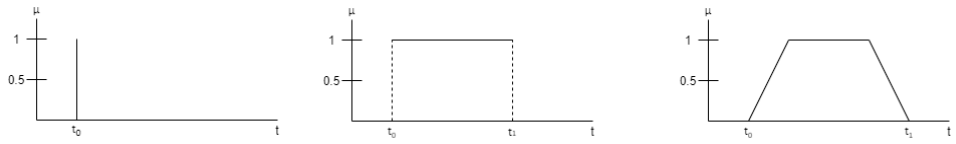


Figure 5.1: Graphical representation of Instant (1), Imprecise interval (2) and Fuzzy interval (3).

As mentioned above, this temporal model has been designed with the objective of grouping the data, if possible, in the most general time reference possible, such as "During the month of September" or "Second week of September". In this way, we seek to make the discourse more easily readable and understandable, even if some precision or accuracy is lost in the descriptions. This temporal model comprises references that group data at month level, for example "During the month of September" up to hour level, so that the most appropriate reference is automatically searched following the hierarchy of references in a descending way. Moreover, the concepts defined in the ontology do not have a strict meaning to improve the readability of the resulting texts. For this purpose, we have defined the following rules (Figure 5.2):

- An ascending hierarchy module has been defined for the evaluation of periods so that the ranges of time references do not follow their strict definition. For instance, in the example at the beginning of this section, the reference "the second week of this month"

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

can include, in addition to the days between Monday and Sunday of the second week of September, Sunday of the first week and Monday of the third week of September.

- A top-down hierarchy module of time references has been defined so that, when evaluating a time reference in order to select the most appropriate one for a particular episode, a reference can be selected even if its fulfillment is not 100%. In the example we say "the second week of this month where it was low" (i.e., the temperature was low during the second week of September). This time reference is accepted even if it is not true that the temperature is low for all the days of the second week of September.

- In general, the hierarchy gives preference to linguistic time references (such as "month" or "second week") rather than numerical ones (e.g., "between day 3 and 8"), if there is no excessive loss of precision in the descriptions. In the example we chose the reference "second week of this month" instead of "from September 6 to 12".

5.1.2 Data-to-text model pipeline

From among the three predominant architectures in NLG (see Chapter 2), our model is based on the classic modular pipeline designed in [188] and the model designed in [193], adapting it to the needs of our temporal model.

The pipeline we follow can be described as follows:

- Content determination: at this stage we defined the temporal components of the model through a temporal ontology and the information to be communicated, defining this model using a grammar.

- Discourse planning: at this stage, an order and structure is provided to the messages in order to generate a well-structured text.

- Microplanning: at this phase we defined the set of rules and restrictions to be followed when generating the components of the linguistic description.

- Structure realization: at this phase we defined the set of rules and requirements about the surface realization of the resulting linguistic description.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

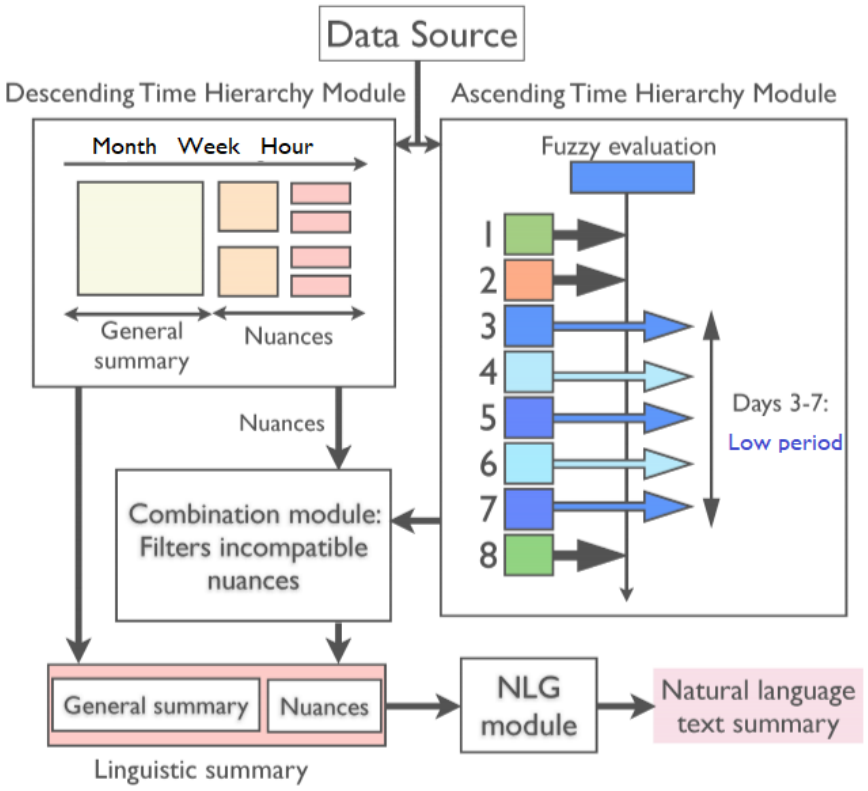


Figure 5.2: Top-down hierarchy module from "month" to "hour" and a bottom-up hierarchy module of period evaluation representation of the temporal model.

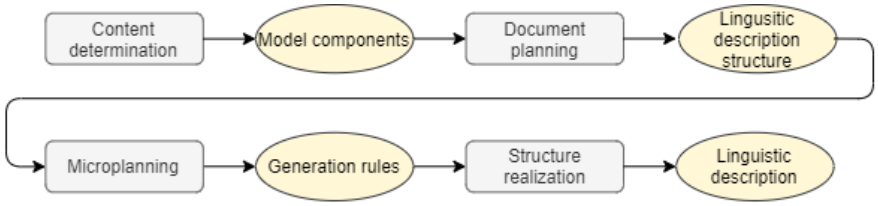


Figure 5.3: Pipeline of the proposed data-to-text model. Rectangles show the stages, while ellipses show the outputs.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Content determination

This first stage of the pipeline (Figure 5.3) is focused on deciding what information communicate in the resulting linguistic description. We defined the D2T model through a free-context grammar, which allowed us to formally define the parts of the linguistic description and their content. To generate the grammar, we used the BNF metalanguage [106] to define the model components. The terminal elements are represented with capital letters, and the metasymbols used are:

- "::=" rewriting metasymbol
- "[]" optional content
- "|" split the mutually exclusive options delimited by "( )"
- "<>" non-terminal element

Table 5.1 shows the resulting grammar where rules from (R1) to (R20) define the structure of the linguistic description, detailing the content of its sections; rules from (R21) to (R36) include in the grammar the temporal concepts of Section 5.1.1 to generate the temporal references in the description generation. Rules from (R37) to (R39) include the definition of quantifier, both absolute and relative, which are the most widely used in the literature. We needed them for modeling the following imprecise terms in the linguistic descriptions: *i)* temporal references and relations since temporal information can be both precise and imprecise and we need to combine them, and *ii)* knowledge about the quantity of elements that fulfill one or more properties.

Code	Structure
(R1)	<Linguistic description> ::= <General summary> [<Intensification>] [<Exception>]
(R2)	<General summary> ::= <General description> [<General trend>]
(R3)	<General description> ::= <Temporal reference> <Description general value> <Value> <Detail values> [<Cause>]
(R4)	<Description general value> ::= <Subject> <Verb>
(R5)	<Subject> ::= [<Quantifier>] SUBJECT | ...
(R6)	<Verb> ::= HAS BEEN | ...

USC DE COMPOSTELA

(R7) <Value> ::= HIGH | LOW | ...

(R8) <Detail values> ::= if <Value> == VARIABLE then WITH VALUES 1* <Value>

(R9) <Cause> ::= <Cause description> <Clause value>

(R10) <Cause description> ::= CAUSED BY THE CAUSE | ...

(R11) <Cause value> ::= [<Quantifier>] CAUSE | ...

(R12) <General trend> ::= WITH A[N] <Trend type> TREND FROM <Trend value> TO <Trend value>

(R13) <Trend type> ::= INCREASING | DECREASING | ...

(R14) <Trend value> ::= HIGH | LOW | ...

(R15) <Intensification> ::= <Intensification beginning> <Intensification description> <Value> [<Cause>] <Temporal reference> [<Intensification-exception worsening>]

(R16) <Intensification beginning> ::= IT SHOULD BE NOTED THAT | ...

(R17) <Intensification-exception description> IT HAS REACHED VALUES | ...

(R18) <Intensification-exception worsening> ::= WITH VALUES <Value> [<Cause>] <Temporal reference>

(R19) <Exception> ::= <Exception beginning> <Temporal reference> <Intensification-exception description> <Value> [<Cause>] [<Intensification-exception worsening>]

(R20) <Exception beginning> ::= HOWEVER | ON THE OTHER HAND | ...

(R21) <Temporal reference> ::= (<Instant> | <Instantaneous event> | <Interval> | <Episode> | <Not grouped moments>)

(R22) <Instant> ::= $t \in \tau$ | NOW | ...

(R23) <Instant event> ::= INSTANT EVENT | ...

(R24) <Interval> ::= (<Imprecise interval> | <Imprecise event> | <Fuzzy interval> | <Fuzzy event>)

(R25) <Imprecise interval> ::= (<Beginning-end interval> | <Beginning interval> | <End interval>)

(R26) <Beginning-end interval> ::= FROM (<Instant | <Instant event>) TO (<Instant | <Instant event>)

(R27) <Beginning interval> ::= FROM (<Instant | <Instant event>)

(R28) <End interval> ::= TO (<Instant | <Instant event>)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

(R29)	<Imprecise event> ::= (<Beginning-end interval> | <Beginning interval> | <End interval>)
(R30)	<Fuzzy interval> ::= (<Beginning-end interval> | <Fuzzy temporal interval>)
(R32)	<Fuzzy event> ::= (<Beginning-end interval> | <Fuzzy temporal interval>)
(R33)	<Modifier> ::= FIRST | LAST | ...
(R34)	<Time unit> ::= SECOND | HOUR | WEEK | ...
(R35)	<Episode> ::= <Quantifier> <Time unit> (<Beginning-end interval> | <Beginning interval> | <End interval>)
(R36)	<Not grouped moments> ::= SEVERAL TIMES OF <Time unit>
(R37)	<Quantifier> ::= (<Absolute quantifier> | <Relative quantifier>)
(R38)	<Absolute quantifier> ::= BETWEEN 5 AND 10 | MORE THAN 5 | ...
(R39)	<Relative quantifier> ::= HALF | MOST | ...

Table 5.1: Description of the rules which compose the grammar of the designed linguistic summarization model.

Document planning

Linguistic description content is composed of three parts, described in (R1) of the grammar:

- General summary (rules from (R2) to (R14)), which provides an overview of the situation (e.g. “The temperature has been high in the afternoon”). It also includes a trend description, if applicable (e.g., “with an increasing trend from low to normal”). It is important to notice that (R8) is a precondition, so it is only enabled if the situation is variable. This conditional clause format is based on [149].

- Intensification (if applicable) (rules from (R15) to (R18)) which highlights the events/episodes in which the general summary is fulfilled to a greater extent sorted in ascending order by value or date (e.g. “It should be noted that the temperature has reached extremely high values in the early hours of the afternoon”).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Exception (if applicable) (rules from (R19) to (R20)) which highlights the events/episodes in which the general summary goes against sorted in ascending order by value or date (e.g. "Nevertheless, the temperature was low in the last hours of the afternoon").

Also, at this stage, we defined the tools for the realization task based on the needs of each application case, e.g., the language of the resulting linguistic description. In the real use cases considered, we have produced the descriptions in both languages using SimpleNLG-ES [185] and SimpleNLG-GL [31], which are extended versions of SimpleNLG [88] for Spanish and Galician languages, respectively.

Microplanning

In this section, we stated rules for each of the three parts that make up the linguistic description that establish the limitations to be taken into consideration when generating the text. These microplanning rules are based on Gricean maxims [94], widely used in the Natural Language Generation field [49, 212]. Microplanning rules can be interpreted as follows:

- Quality: rules must be accurately described.
- Quantity: rules must contain enough information to be understandable without providing more information than necessary.
- Relevance: the pertinent rules for this model should be included.
- Manner: rules must be clearly defined, avoiding ambiguous descriptions and they must be sorted in terms of the linguistic description structure.

General summary

- The value of the general summary can be obtained in two different ways. First, we can calculate it as a weighted average where the most recent value is the most relevant to describe the general situation through a temporal reference, for instance *"In the last months"* or *"In the last weeks"*. On the other side, in other variables, we selected the last available record as the most representative.
- If a negative case is described, its cause must be included, otherwise this part is omitted.
- When selecting the time reference, the most general reference possible should be chosen whenever its truth degree is greater than or equal to a certain threshold.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Trend description is only included if a clear ascending or descending trend is detected in the data.

Intensification and exception

The following microplanning rules were applicable both for intensification and exception:

- If the cause of a negative situation is described in the general summary, it is omitted in the intensification or exception, even if applicable.
- When selecting the time reference, the most general possible should be chosen whenever its truth degree is greater than or equal to a certain threshold.
- Periods of time with the same value of intensification or exception will be grouped by that value.

Structure realization

It is clear that the content of a linguistic description is important, that is, it contains information in a brief and precise manner, but the way in which this content is presented is also very important. Therefore, with the aim of facilitating the generated texts reading, the following layout rules were included:

- Variable values and trends are highlighted in bold type.
- Both in intensification and exception, if the number of highlighted case is higher than a determined threshold, they will be arranged as an itemization. Nevertheless, when the number of elements is equal to or less than this threshold, they will be both included as plain text.

The D2T model described in the previous sections was intended to support the development of D2T systems to describe time series. For this purpose, within this model, we have designed a temporal model that includes the definition of temporal concepts that allow the design of the necessary references.

The characteristics of this model allow it to be applied to any use case because the generality of these linguistic descriptions makes it possible to adapt its structure to any need that may arise. In this case, we have applied this model to two very different application areas

that have been selected for their great interest in society because they both have a great impact on people's health: *i)* meteorology field, in particular air quality data; and *ii)* field of meteorology, namely cardiac rehabilitation.

5.2 ICA2TEXT: Linguistic descriptions of Air Quality Index (AQI) time series

The presence of pollutants in the air and, therefore, the deterioration of air quality can have harmful effects on people's health. The Air Quality Index (AQI) is a widely used indicator worldwide to measure the presence and density of several types of pollutants in the atmosphere. In the literature exist several approaches over the years to generate summaries from air quality data. A prototype of the TEMSIS systems is described in [27], focused on describing whether the (daily, monthly...) early warning thresholds for each pollutant were exceeded or not. Also, the MARQUIS system [229] of numerical AQI time series includes simple time references (specific hours or intervals) for describing the last AQI value and pollutants (concentration, archive information and forecast). In [181] a solution focused on generating AQI descriptions for a temporal window involving three daily values is proposed. In general, the linguistic descriptions in all these systems have: i) a fixed structure that allows limited variability, and ii) very specific temporal references which focus solely on the specific needs of the data they describe.

Due to the impact that air quality has on the health of the population, meteorological agencies around the world [75] provide this information in an understandable way to citizens, so far in graphic format. Specifically, the Galician Meteorological Agency (MeteoGalicia) [154] works with air quality index data for the community of Galicia (NW Spain), where air quality meteorologists identified the need to complement the graphical information provided to the public with linguistic descriptions to improve their understanding. Therefore, we developed ICA2TEXT, a D2T system to generate linguistic descriptions from air quality in collaboration with experts from MeteoGalicia. Having domain experts involved when building an NLG system helps to design higher quality systems [138]. In particular, we worked with hourly time series which describe the evolution and the prediction of the AQI in each meteorological station. Figure 5.5 represents an example of an hourly temporal series of 31 hours where all the performed measures obtained a "good" AQI label.

This service has recently been updated because of a process of unification of the AQI by

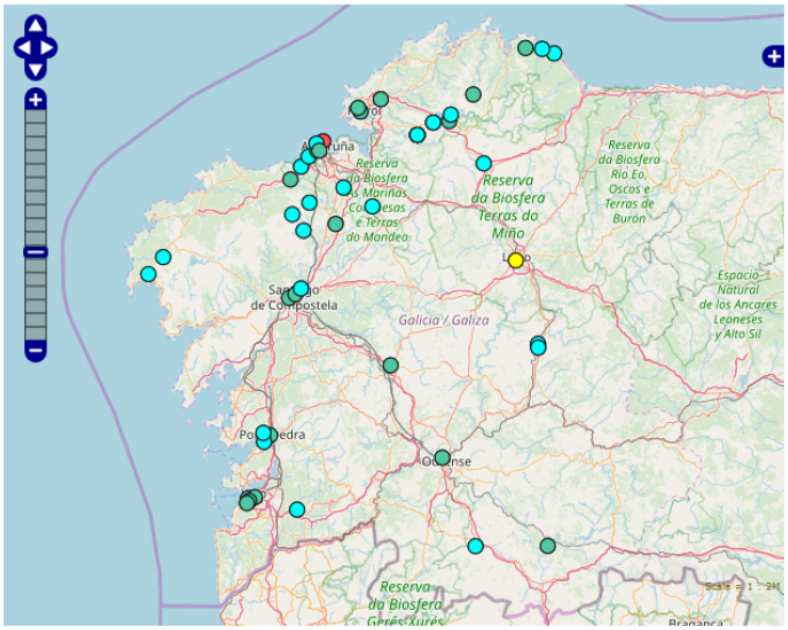


Figure 5.4: Map of the meteorological stations network of the Galician Meteorological Agency.

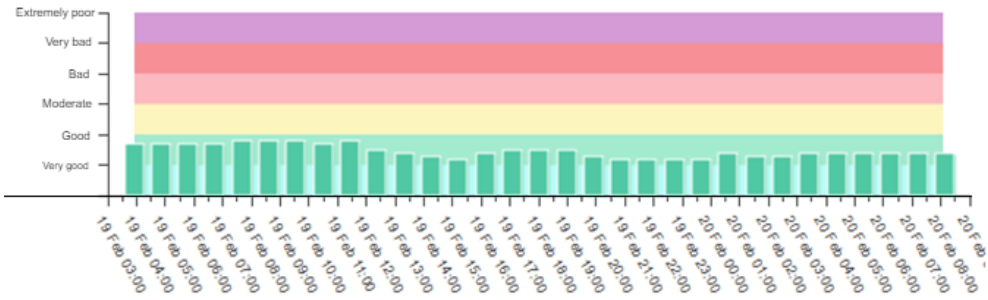


Figure 5.5: 31-hour AQI evolution time series where the station registered “good” values.

different meteorological agencies, based on the European Environment Agency criteria [75]. Currently, it has six labels with a positive, neutral or negative perception (Table 5.2). These labels are represented by a color code where, for instance, the color purple means “extremely

poor" whilst yellow means "good". To determine the suitable label for a situation, this service measures five different pollutants: sulfur dioxide (SO_2), nitrogen dioxide (NO_2), fine particulate matter ($PM25$ and $PM10$) and ozone (O_3).

Perception	Positive		Neutral	Negative		
Label	Very good	Good	Moderate	Bad	Very bad	Extremely poor
Index	0	1	2	3	4	5

Table 5.2: Air Quality Index (AQI) labels with their perception and numerical index.

The characteristics of this system and the requirements raised by the meteorologists make our Data-To-Text model described in previous sections very suitable for the development of ICA2TEXT, since in meteorological data the temporal component is very important and the experts stressed the importance of a discourse that includes both general information and exceptional cases. The following sections will describe the performed tasks during the development of this system, following the phases of the pipeline defined in our model.

5.2.1 Stages of the system development

In this section, we present the development process carried out to generate the AQI series linguistic description.

Content determination

In this phase, focused on deciding which messages are to be communicated, we defined the necessary components to generate the linguistic labels.

Calculation of general summary labels

First of all, we calculated the label (Table 5.2) that best represents the overall time series to include it in the general description. This label was obtained as a weighted average where the most recent value is the most relevant to describe the general situation. Figure 5.6 shows an hourly temporal series with the fuzzy set used to calculate the general label. In this example, the current time point (now) is 20 February, 9:00. The temporal reference "In the last hours" is defined accordingly as a fuzzy set with membership 0 for all the time points before 19 February, 3:00, membership 1 for all the time points in the last twelve hours (i.e., from February 19, 21:00 to February 20, 09:00) and membership in the range (0, 1) to all the time

points from February 19, 04:00 to February 19, 08:00. The general label is calculated through a weighted average of the AQI values of each time point, which considers the corresponding membership value to “In the last hours”.

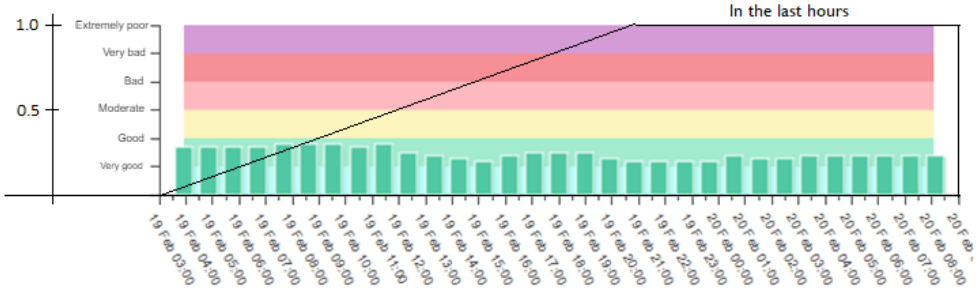


Figure 5.6: General linguistic description label calculation.

In the description trend, we also calculated the beginning and end trend with a weighted average. Figure 5.7 represents the fuzzy sets used to calculate those labels. In this example, the beginning label includes data values from February 19, 03:00 to February 19, 11:00 (with membership/weight 1), whereas values from February 19, 12:00 to February 19, 14:00 have a membership in the range (0, 1) and values from February 19, 15:00 on are not relevant to this label, so they have a weight of 0. On the other side, to calculate the ending label, values from February 20, 07:00 to February 20, 09:00 have a weight of 1 whereas the value on February 20, 06:00 has a weight in the range (0, 1). Values before and on February 20, 05:00 are not relevant, so they have a weight of 0.

Temporal references

In the official style guide of MeteoGalicia, they define the different parts of the day, {morning, afternoon, night}, through a range of hours for summer and winter (Table 5.3).

	Morning	**Afternoon**	**Night**
Summer	06:00 - 14:00	14:00 - 21:00	21:00 - 06:00
Winter	07:00 - 14:00	14:00 - 20:00	20:00 - 07:00

Table 5.3: Range of hours for the moments of the day {morning, afternoon, night} in summer or winter from the style guide of MeteoGalicia.

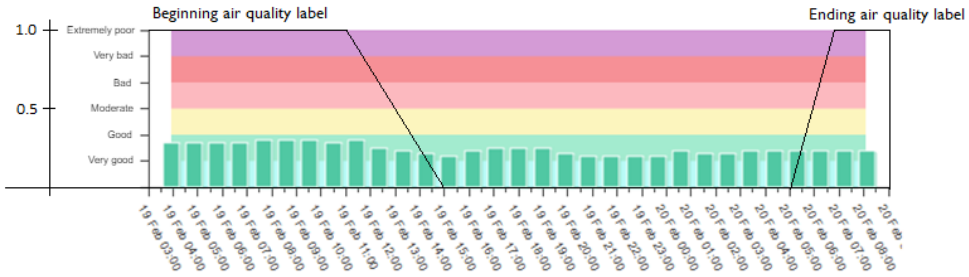


Figure 5.7: Definition of fuzzy intervals (R30) to represent the beginning and ending of the trend description.

Although the ranges that define these moments of the day are declared statically (as well as the definition of a full day from 00:00:00 to 23:59:59), their usage when speaking is conditioned by the imprecision in the language. We defined in a fuzzy way the following temporal references represented by trapezoidal fuzzy sets in Figures 5.8 and 5.9:

- Full day: instead of a strict definition from the 00:00:00 to the 23:59:59, we grouped as a day also the two hours before and after with a weight in the range [0, 1].
- Morning, afternoon, night: as we mention above, these temporal references are defined in the style guide of MeteoGalicia. Using that definition as baseline, we defined them as a trapezoidal fuzzy set where the two hours before and after the limits are considered with a weight in the range [0, 1].
- First, central, last hours of the {morning, afternoon, night}: we defined these three temporal references to describe more specific situations. These labels are also defined as trapezoidal fuzzy sets.

5.2.2 Document planning

At this stage, the general document planning rules of the model, focused on deciding the order and structure of the messages, were adapted to this use case. First, it is important to note that a linguistic description is focused of describing the evolution of the air quality index for a unique station. Thus, for each station, the three sections that make up the linguistic description include the following information:

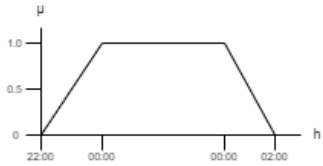


Figure 5.8: Fuzzy definition of a full day.

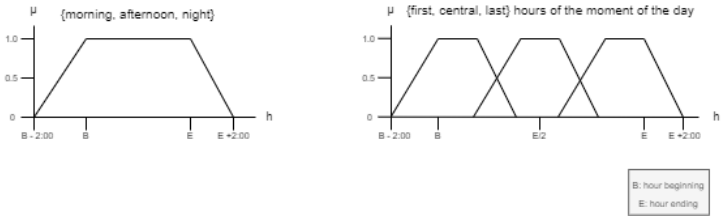


Figure 5.9: Fuzzy definition of the three moments of the day (morning, afternoon, night) (2) and of the {first, central, last} hours of a moment of the day.

- The general summary includes the most representative label of the air quality index in the last hours. In addition, the trend of the evolution is included if it is sufficiently significant. Example: “In the last hours, the air quality index of Campelo has been generally good. The trend has been worsening, starting with a very good AQI and ending with a good one.”.
- The intensification highlights the periods in which air quality values were presented that comply to a greater extent with the same perception as the one highlighted in the general summary. This part is only included if there are time periods that meet this condition. Example: “It should be noted that between 06:00 and 10:00 yesterday and at 12:00 yesterday, the air quality index was very good.”.
- The exception highlights time slots where the air quality index has a perception contrary to that of the general description. This part is only fulfilled if there is an outstanding period. Example: “On the contrary, at 11:00 a.m. yesterday and at 2:00 p.m. yesterday, the air quality index was bad, due to the pollutant PM10.”.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.2.3 Microplanning

In this section, we qualified the microplanning rules defined in the general model in order to adapt them to this system. Regarding the general summary:

- General value is calculated through a weighted average following the fuzzy set represented in Figure 5.6 using the temporal reference "In the last hours" as it is the most suitable for these time series, where the data usually do not exceed 31 hours.
- Causative pollutant is included if the perception of the AQI label is negative.
- A trend is considered to exist if the initial and final AQI of the time series are different.

On the other side, microplanning rules for the intensification and exception sections are the following:

- If the cause of an AQI period with negative perception is mentioned in the general description, it is omitted in this part. Example: "In the last hours, the air quality index of Cee has been generally bad due to NO2 and SO2 pollutants. It should be noted that at 13:00 yesterday and between 04:00 and 13:00 today, the air quality index has been extremely poor.".
- To select the most appropriate time reference, a threshold of 0.9 is set for the truth degree, so that the most general one that exceeds this value is chosen.
- Periods with the same AQI label are grouped together in the text. Example: "It should be noted that at 13:00 yesterday and between 04:00 and 13:00 today, the air quality index has been extremely poor.".

5.2.4 Structure realization

In order to ensure that the linguistic descriptions generated are informative, making sure that the most relevant information is included, and at the same time understandable, avoiding lengthy linguistic descriptions, a maximum of two cases of intensification and two cases of exception are included to facilitate their reading. Since the MeteoGalicia website offers information in both Spanish and Galician, we have generated the descriptions in both languages using SimpleNLG-ES [185] and SimpleNLG-GL [31], which are extended versions of SimpleNLG [88] for Spanish and Galician, respectively. A linguistic description resulting from the development of the ICA2TEXT system is shown below:

USC UNIVERSIDADE

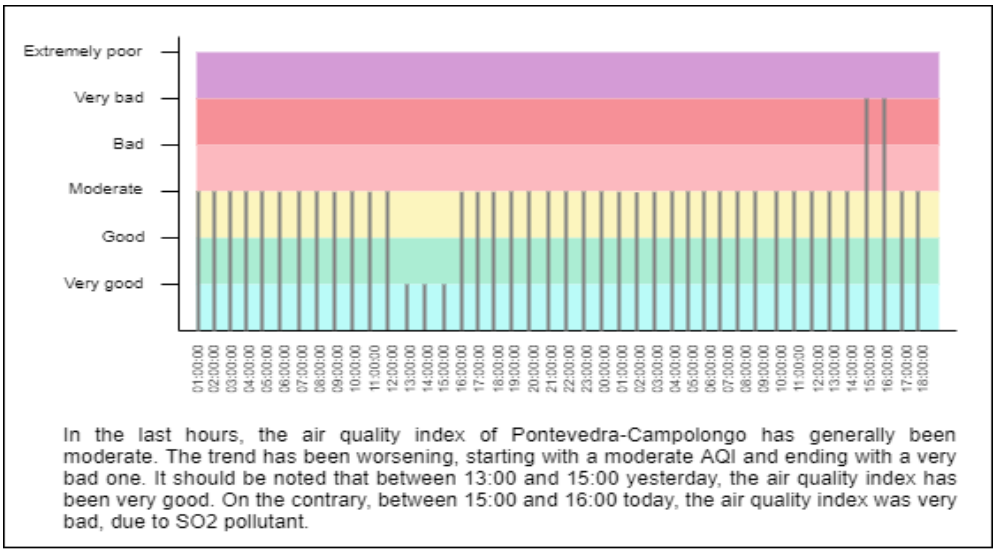


Figure 5.10: Example from the AQI questionnaire designed for the expert validation.

5.2.5 Human expert validation

In the field of Natural Language Generation, validation is currently a topic of great debate, where a standard methodology has been defined considering several dimensions [12]. As previously described in Chapter 2, regarding the object of validation, we can distinguish between: *i)* intrinsic validation, which is focused on measuring the performance of a system in terms of its effectiveness with respect to users; and *ii)* extrinsic validation, focused on measuring the effectiveness of a system in achieving a given goal or objective. Regarding the way validation is performed, we can talk about: i) manual validation, performed by human evaluators, who can be either experts or non-experts, depending on the final objective of the system, and ii) automatic validation, using performance metrics, which are unanimously used in deep learning systems.

Taking into consideration the characteristics and the requirements within the NLG validation methodology and the characteristics of this system, we followed a manual intrinsic validation by experts [102]. We asked three expert meteorologists from the Air Quality Network of the Galician Meteorological Agency (MeteoGalicia) [154] to assess the quality of the linguistic descriptions in this domain and its adequacy both for the description of the AQI

time series, filling out the questionnaire jointly agreeing on scoring several real situations. Experts rated the suitability of the linguistic description generated in the Spanish and the Galician languages (in this chapter we translated it for easy reading) using the appropriate SimpleNLG [88] version. They answered a questionnaire composed of five items (Table 5.4) using a 5-point Likert [222] scale in the range [1, 5] where 1 means "the expert is absolutely disagree" and 5 "the expert is absolutely agree". Items can be divided into two categories: content of the linguistic description (Q1, Q2) and layout (Q3, Q4, Q5).

Code	Item
Q1	Indicate the degree of agreement between the linguistic description provided and the data represented in the figure
Q2	Indicate the degree of agreement between the linguistic description provided and a how you would describe the data
Q3	Indicate the degree of agreement that the vocabulary is used correctly
Q4	Indicate the degree of agreement with the linguistic description organization to facilitate its understanding
Q5	Indicate the degree of agreement with spelling, punctuation and layout

Table 5.4: Questions of the air quality index expert validation test.

None of these three experts had taken part in the definition of any part of the model, since all the requisites were defined with the help of another different expert, also from the Air Quality Network of MeteoGalicia. Furthermore, in this validation, we emphasize the degree of conformity of the human experts with the generated texts, who performed a fully blind assessment. No details were provided about how the linguistic descriptions were generated.

The questionnaire was composed of 30 different meteorological situations with a variable range of hours from 20 different meteorological stations. The stations and the temporal series included in the questionnaire represented a variety of situations, including high variations in the data and different weather conditions. These cases included a graphical representation of the temporal series and the generated linguistic description which described the case, asking them to evaluate the suitability of the linguistic descriptions to describe the different situations. Figure 5.11 shows an example extracted from the questionnaire.

In Table 5.5, we present a summary of the experts' scores for each of the questions, the content (Q1, Q2) and layout (Q3, Q4, Q5) dimensions and the overall result. In general, the results show the experts agree with the presented linguistic descriptions since the scores average overall average is 4.67, with a standard deviation of 0.75. Besides, mode shows the

USC UNIVERSIDADE

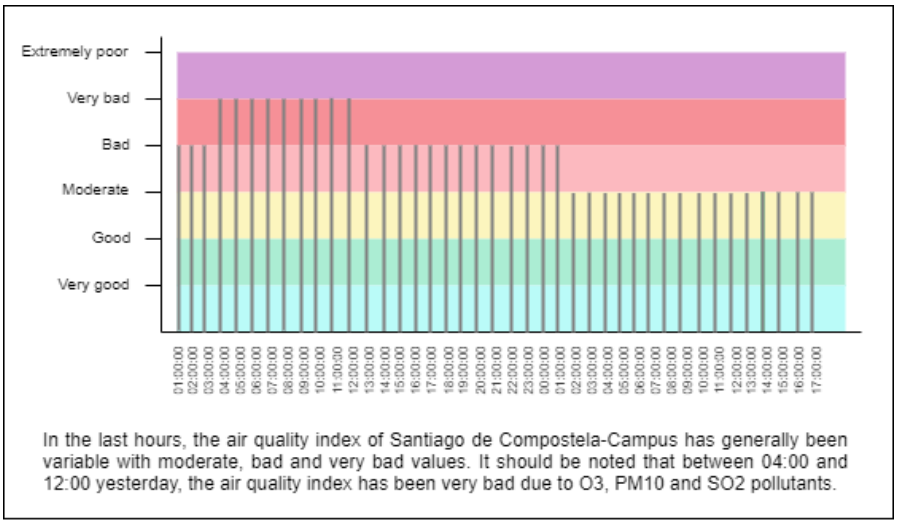


Figure 5.11: Example from the AQI questionnaire designed for the expert validation where a bad situation is described.

	Average	**SD**	**Mode**	**Median**	**IQR**
Q1	4.58	0.87	5	5	1
Q2	4.15	1.01	5	4	1
Q3	4.75	0.70	5	5	0
Q4	4.92	0.28	5	5	0
Q5	4.97	0.18	5	5	0
Content	4.37	0.96	5	5	1
Layout	4.88	0.46	5	5	0
Overall	4.67	0.75	5	5	0

Table 5.5: Expert validation results for AQI descriptions.

greatest used value is 5, i.e. the maximum score, and the median is also 5, with an interquartile range of 0. It is noticeable experts are more in agreement with the questions regarding the linguistic description layout with a score of 4.88 on average, whilst results show a score of 4.34 on average regarding the linguistic description content. Overall, we can conclude that these generated linguistic descriptions are very suitable both in content and form to describe air quality index temporal series, since the independent experts who assessed them indicated a very high level of satisfaction with the automatically generated texts and their adequacy

to the temporal series. Nevertheless, a deeper analysis of the experts' answers suggests that while the structure of the linguistic description is quite adequate for this case, a more thorough analysis of the weaknesses in its content should be made with the aim of improving it.

5.2.6 Transfer to society

The system described throughout this section is of great interest, not only from the research point of view as an application with real data of the temporal fuzzy model described above, but also from the point of view of transfer to society. As mentioned above, the idea of developing this system was a consequence of the need expressed by MeteoGalicia experts to provide understandable information to citizens on air quality data. After completion of system development, including expert validation, ICA2TEXT was deployed on a MeteoGalicia test server. Currently, it is in the last phase of testing, so it will be deployed very soon on the production server to be offered as a service on the official MeteoGalicia website.

5.3 SMART CARDIA: Intelligent management of the cardiac rehabilitation process

The clinical discharge report is the document issued by the doctor responsible for a patient at the end of the care process in a hospital center. The report is a tool for communication and quality control of the work performed in a program, such as cardiac rehabilitation. As in any communication process, there is a sender, the clinician who writes the report, and one or more receivers: the patient, the patient's relatives, the primary care doctor, the same doctor who wrote the report or others in the same department, other specialists and emergency doctors. Current medical practice is characterized by an increase in information needs and in the amount of health data being managed, both for professionals and for patients, who have more and more chronic diseases, for a longer period of time, with the need for a greater number of complementary test results and more complex care. The digitization of healthcare partly changes the concept of the discharge report. In the analog era, the report was also the information collector, however, nowadays medical records do this function and this should reorient and redefine the discharge report.

Within the field of cardiology, secondary prevention programs and Cardiac Rehabilitation Units (CRU) have been developed in several countries [98, 139] being the most effective tool for improving prognosis. Cardiac Rehabilitation (CR) is a specialized area focused on

the complete rehabilitation of patients who have recently suffered from heart disease. It is composed of a set of activities to understand the causes of the heart disease and to ensure the best physical, psychological and social conditions for patients who suffer a temporary disability due to illness.

In this field arises the necessity of a report with the patients' evolution data within the program, which assists the decision making of health professionals and helps patients to control their disease. Moreover, the clinical information has a very important temporal component that must be taken into consideration. Due to the importance of this feature, in the literature, several purposes proposed approaches focused on managing temporal reasoning in the medical domain [70, 126, 267]. This temporal information may be precise, for example, the results of a medical test obtained at a certain time; or imprecise, for instance, the patient's description of how the symptoms manifested. Both types of information are of interest and should be included in the reports generated.

Therefore, in this section, we propose a real application of our D2T model for automatically explaining data series in collaboration with the cardiology area of the University Clinical Hospital of Santiago de Compostela (Galicia, NW Spain) within the project SMART CARDIA. SMART CARDIA is a project led by the company Balidea aimed at developing a software platform for the intelligent management of the cardiac rehabilitation process. To this end, a technological solution has been implemented that makes use of dynamic workflows to implement the control model of the cardiac rehabilitation process, based on a service-oriented architecture which externalizes its functionalities, and which has a graphic interface specifically designed for the needs of healthcare professionals.

5.3.1 Stages of the system development

Content determination

In this phase, focused on deciding which messages are to be communicated, we defined the necessary components to generate the linguistic labels. A very large number of variables of interest are tracked in this program, so this system has been developed to assist experts in these tasks. Among the variables contemplated, cardiologists need to build natural language reports adapted to the profile of each user, so, within this project, we have developed a text

USC UNIVERSIDADE

generation module based on the model previously described, generating linguistic descriptions from the variables of interest in these cases, therefore, we were focused on five health categories, describing information about the nineteen variables of interest (Table 5.6).

Lipids	Smoking	Diabetes	Diet	Blood pressure
Alanine Transaminase (ALT)	Cooximetry	Blood sugar	Body Mass Index (BMI)	Home diastolic pressure
Aspartate Aminotransferase (AST)	Daily cigarettes number	Glycated Hemoglobin (Hba1C)	Weight	Hospital diastolic tension
HDL cholesterol			Body fat percentage	Home systolic pressure
LDL cholesterol				Hospital systolic pressure
Total cholesterol				
Triglycerides				
Alkaline Phosphatase (ALP)				
Gamma-glutamyl Transpeptidase (GGT)				

Table 5.6: Monitoring variables during the cardiac rehabilitation grouped by category.

All health-related applications must be supported by explicit medical knowledge and scientific evidence, which has been provided to us by professionals through guidelines. The references for each of the variables considered are described below. With regard to the description trend, we also calculated the beginning and the end trend through a weighted average through fuzzy intervals (Figure 5.12).

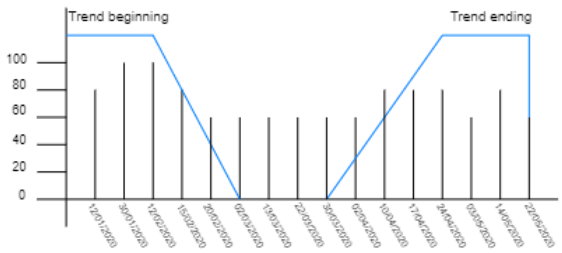


Figure 5.12: Definition of fuzzy intervals (R30) to represent the beginning and ending of the trend description.

Lipids

The variables in this category focus on measuring the presence of lipids in the blood.

HDL cholesterol HDL cholesterol is the well-behaved "good cholesterol". It cruises the bloodstream as it removes harmful bad cholesterol from where it doesn't belong. High HDL levels reduce the risk for heart disease, but low levels increase the risk. To generate the

linguistic description of this variable, we needed *i)* the sex of the patient and *ii)* the values history. Besides, the most representative value included in the general summary of this variable was the last available record, including descriptive label according to the reference values of Table 5.7.

	Men	**Women**
Low	40 mg/dL	50 mg/dL
Good	40 - 59 mg/dL	
Optimal	$\geq$ 60 mg/dL	

Table 5.7: Reference values of HDL cholesterol.

LDL cholesterol LDL cholesterol is often called the "bad" cholesterol because it collects in the walls of blood vessels, raising the chances of health problems like a heart attack or stroke. To generate the linguistic description of this variable, we needed the information about *i)* if the patient has a diagnosed cardiac disease, *ii)* if he/she has diagnosed diabetes and *iii)* the values history. Besides, the most representative value included in the general summary of this variable was the last available record.

	Cardiac disease	**Not cardiac disease**
Optimal	< 100 mg/dL	100 - 129 mg/dL
High limit	130 - 159 mg/dL	
High	160 - 189 mg/dL	
Very high	$\geq$ 190 mg/dL	

Table 5.8: Reference values of LDL cholesterol.

Total cholesterol Total blood cholesterol is a measure of LDL cholesterol, HDL cholesterol, and other lipid components. To generate the linguistic description of this variable, we only needed the values history, focusing on the last available value.

Triglycerides Triglycerides are a combination of three fatty acids or fats (i.e. saturated fat, unsaturated fat, or both) combined with glycerol, a form of glucose. To generate the linguistic description of this variable, we only needed the values history. To describe the general situation through a representative label, we calculated the weighted average of all available values, as represented in Figure 5.13.

	Reference values
Optimal	< 200 mg/dL
High limit	200 - 239 mg/dL
High	$\geq$ 240 mg/dL

Table 5.9: Reference values of total cholesterol.

	Reference values
Normal	< 150 mg/dL
High limit	150 - 199 mg/dL
High	200 - 499 mg/dL
Very high	$\geq$ 500 mg/dL

Table 5.10: Reference values of triglycerides.

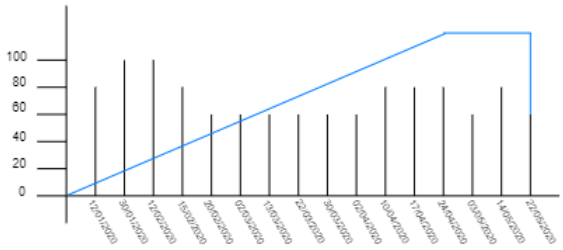


Figure 5.13: Calculation of the representative value in the general summary.

Alanine Transaminase (ALT) Alanine transaminase (ALT) is the most widely used clinical biomarker of hepatic health. ALT is one of the enzymes that helps the liver convert food into energy. A high concentration of this enzyme may be a sign that the liver is injured or irritated and that its enzymes are overflowing from the liver cells. To generate the linguistic description of this variable, we only needed the values history. To describe the general situation through a representative label, we calculated the weighted average of all available values, as represented in Figure 5.13.

Aspartate Aminotransferase (AST) AST is an enzyme found mainly in the liver, but also in the muscles. When the liver is damaged, it releases AST into the bloodstream. Also, in this case, we needed the values history and we calculated the most representative general

	Reference values
Low	< 4 U/L
Normal	4 - 36 U/L
High	> 36 U/L

Table 5.11: Reference values of alanine transaminase (ALT).

label through a weighted average of all records.

	Reference values
Low	< 8 U/L
Normal	8 - 48 U/L
High	> 48 U/L

Table 5.12: Reference values of aspartate aminotransferase (AST).

Alkaline Phosphatase (ALP) ALP is an enzyme found in several tissues throughout the body. Elevated levels of ALP in the blood are most commonly caused by liver disease, bile duct obstruction, gallbladder disease, or bone disorders. Also, in this case, we needed the values history and we calculated the most representative general label through a weighted average of all records.

	Reference values
Low	< 44 U/L
Normal	44 - 147 U/L
High	> 147 U/L

Table 5.13: Reference values of alkaline phosphatase (ALP).

Gamma-glutamyl Transpeptidase (GGT) GGT is an enzyme that is found in many organs throughout the body, where the highest concentrations are found in the liver. GGT is elevated in the blood in most diseases that cause damage to the liver or bile ducts. Also, in this case, we needed the values history and we calculated the most representative general label through a weighted average of all records.

USC UNIVERSIDADE

	Reference values
Low	< 7 U/L
Normal	7 - 50 U/L
High	> 50 U/L

Table 5.14: Reference values of gamma-glutamyl transpeptidase (GGT).

Blood pressure

These variables measure the types of blood pressure.

Diastolic blood pressure Diastolic blood pressure indicates how much blood pressure is exerting against artery walls while the heart is resting between beats. Patients can measure this variable from home or at the hospital. To generate the linguistic descriptions, we needed information about *i)* the age of the patient, *ii)* if the patient has been diagnosed arterial hypertension and *iii)* the values history. The general label was calculated through a weighted average.

	Hypertensive patient				**Non-hypertensive patient**	
	Home		Hospital		Home	Hospital
	> 60 years	< 60 years	> 60 years	< 60 years		
Optimal	-	-	-	-	< 76 mmHg	< 80 mmHg
Normal	-	< 80 mmHg	-	< 80 mmHg	76 - 84 mmHg	80 - 84 mmHg
Normal-high	-	-	-	-	85 - 89 mmHg	85 - 89 mmHg
Arterial hypertension	-	-	-	-	$\geq$ 90 mmHg	$\geq$ 90 mmHg

Table 5.15: Reference values of diastolic blood pressure.

Systolic blood pressure Systolic blood pressure indicates how much pressure blood is exerting against artery walls when the heart beats. As in the description of diastolic blood pressure, to generate the linguistic descriptions, we needed information about *i)* the age of the patient, *ii)* if the patient has been diagnosed arterial hypertension and *iii)* the values history. The general label was calculated through a weighted average.

Smoking

If a patient in the program is a smoker, this dimension is monitored in order to help him or her to quit.

USC UNIVERSIDADE

	Hypertensive patient				Non-hypertensive patient	
	Home		Hospital		Home	Hospital
	> 60 years	< 60 years	> 60 years	< 60 years		
Optimal	-	-	-	-	< 125 mmHg	< 120 mmHg
Normal	< 150 mmHg	< 140 mmHg	< 150 mmHg	< 140 mmHg	125 - 134 mmHg	120 - 129 mmHg
Normal-high	-	-	-	-	135 - 139 mmHg	130 - 139 mmHg
Arterial hypertension	-	-	-	-	$\geq$ 140 mmHg	$\geq$ 140 mmHg

Table 5.16: Reference values of systolic blood pressure.

Daily cigarettes The patient can record the number of cigarettes consumed daily. To generate the linguistic description of this variable, only the history of values was needed, including the result of calculating the weighted average of the available values in the resulting text.

Cooximetry Cooximetry is a technique by which carbon monoxide (CO) in exhaled air is determined. To generate the linguistic description of this variable, only the history of values was needed, calculating the general label through a weighted average.

	Reference values
No intoxication	0 - 5 ppm
Light intoxication	6 - 10 ppm
Moderate intoxication	> 11 - 30 ppm
Severe intoxication	> 30 ppm

Table 5.17: Reference values of cooximetry.

Diabetes

Patients, both with previously diagnosed and undiagnosed diabetes, can track diabetes-related variables.

Blood sugar Blood sugar, also called "glucose" is the main sugar found in the blood. It comes from food and is our main source of energy. To generate the linguistic description of this variable, we needed information about *i)* if the patient has been previously diagnosed with diabetes and *ii)* the history of values. In the general summary, we included the last available measure and its corresponding label.

USC UNIVERSIDADE

	Diabetic patient	Non-diabetic patient
Low	< 80 mg/dL	< 102 mg/dL
Normal	80 - 130 mg/dL	102 - 125 mg/dL
Diabetes compatible values	> 130 mg/dL	> 125 mg/dL

Table 5.18: Reference values of blood sugar.

Glycated Hemoglobin (Hba1C) The glycated hemoglobin (HbA1c) test is a blood test for type 2 diabetes and prediabetes. It measures the average blood glucose or blood sugar level over the past three months. Also in this case, to generate the linguistic description of this variable, we needed information about *i)* if the patient has been previously diagnosed with diabetes and *ii)* the history of values. In the general summary, we included the last available measure and its corresponding label.

	Diabetic patient	Non-diabetic patient
Normal	< 7%	< 5.7%
High	$\geq$ 7%	5.7 - 6.4%
Diabetes compatible values	-	> 7%

Table 5.19: Reference values of glycated hemoglobin.

Diet

In the cardiac rehabilitation program, it is also necessary to monitor the patients' eating habits, in addition to controlling the variables related to this dimension.

Body Mass Index (BMI) BMI is a value derived from the mass (weight) and height of a person. The BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m^2, resulting from mass in kilograms and height in metres. To generate the linguistic description of this variable, we only needed *i)* the patient's height and *ii)* the weight records. To describe the general situation through a representative label, we only considered the last available record to obtain the most representative BMI label.

Weight The weight is also tracked by constructing a linguistic description that includes the weighted average of the available values. We also needed the height of the patient to

	Reference values
Underweight	< 18.5
Normal weight	18.5 - 24.9
Overweight grade I	> 25 - 26.9
Overweight grade II	> 27 - 29.9
Obesity grade I	> 30 - 34.9
Obesity grade II	> 35 - 39.9
Obesity grade III	> 40 - 49.9
Obesity grade IV	> ≥ 50

Table 5.20: Reference values of body mass index (BMI).

generate this linguistic description, including a general label that represents the overall result based on the patient's BMI.

	Reference values
Underweight	< 18.5
Normal weight	18.5 - 24.9
High	≥ 25

Table 5.21: Reference values of weight.

Body fat percentage In addition to the previously described variables, it is necessary to control the patient's body fat percentage to contextualize the above variables. To describe this variable, we needed the following patient's information: *i)* age, *ii)* sex and *iii)* the history of values. To generate the general summary, we calculated the weighted average of the available values.

	20 - 29		**30 - 39**		**40 - 49**		**≥ 50**	
	Man	Woman	Man	Woman	Man	Woman	Man	Woman
Low	< 11	< 16	< 12	< 17	< 14	< 18	< 15	< 19
Good	11 - 13	16 - 19	12 - 14	17 - 20	14 - 16	18 - 21	15 - 17	19 - 22
Normal	14 - 20	20 - 28	15 - 21	21 - 29	17 - 23	22 - 30	18 - 24	23 - 31
High	21 - 23	29 - 31	22 - 24	30 - 32	24 - 26	31 - 33	25 - 27	32 - 34
Very high	> 23	> 31	> 24	> 32	> 26	> 33	> 27	> 34

Table 5.22: Reference values of body fat percentage.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Document planning

At this stage, the general document planning rules of the model, focused on deciding the order and structure of the messages, were adapted to this use case. First of all, it is important to note that in a linguistic description, only one variable is considered, thus, for one patient, this system generates one linguistic description per medical variable. The resulting linguistic description is composed of three sections, including the following information:

- The general summary includes the most representative label using the most appropriate temporal reference. Moreover, the trend of the evolution is included if applicable. Example: "In recent weeks, weight has been normal with an average of 61.2 kgs. The trend has been of increase of approximately 3% from 60 kgs to 61.5 kgs.".
- The intensification highlights periods in which the patient records that comply to a greater extent with the same perception of the general summary. This part is only included if there are time periods that meet this condition. Example: "It should be noted that particularly low values have been reached on 09/01/2020, reaching 60 kgs.".
- The exception section highlights time slots where the patient records have a perception contrary to that of the general description. This part is only included if there are time periods that meet this condition. Example: "On the contrary, high values were reached on 09/04/2020, reaching 61.5 kgs.".

Microplanning

In this section, we defined the microplanning rules from our D2T model, including the specific details of this system. First of all, regarding the general summary:

- The general value is calculated in two ways, depending on the variable described, according to the interest of the doctors for each of the indicators considered: *i)* trough a weighted average using the most appropriate temporal reference or *ii)* including only the last recorded value in the patient's history.
- Together with the general label, we included the numerical value of either the weighted average or the last record, including the units of measurement.
- The trend includes information about the percentage, indicating also the initial and the last available record.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

On the other side, microplanning rules for the intensification and exception sections are the following:

- The numerical value of either the weighted average or the last record, including the units of measurement, is included.
- To select the most appropriate time reference, we stated a threshold of 0.9 for the truth degree, so that the most general one that exceeds this value is chosen.
- Periods with the same record of a medical variable are grouped together in the text.

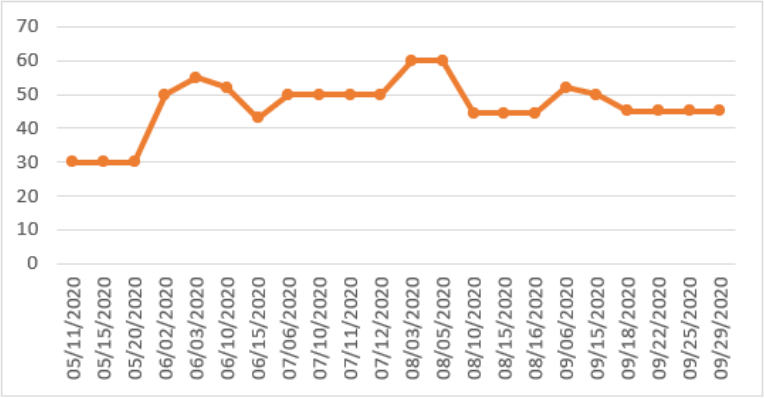
Structure realization

In order to ensure that the linguistic descriptions generated are informative and useful for doctors, making sure that the most relevant information is included and at the same time understandable, they required a maximum of one case of intensification and one case of exception to facilitate their reading. At this moment, in SMART CARDIA we generate linguistic descriptions in Spanish using SimpleNLG-ES [185]. However, in this chapter, we show all examples translated into English for easy reading. A linguistic description resulting from the development of this system is shown below:

5.3.2 Human expert validation

This system is not only an application of our model with real data, but has also been developed in collaboration as part of the industrial R+D project SMART CARDIA. The characteristics of this project, in particular the design of the linguistic module in terms of content determination specified in detail by the medical professionals based on clinical guidelines, scientific evidence and project reports; ensure its adequacy by design. Taking into consideration these methodologies and the characteristics of this system, we performed a manual intrinsic evaluation by experts. In this evaluation, we asked two independent external expert physicians to fill out assess a questionnaire composed by 21 situations, in which they had to evaluate the quality and adequacy of the linguistic descriptions generated by our system. Figure 5.15 shows an example of case extracted from the questionnaire.

Experts rated the suitability of the generated linguistic descriptions generated in the Spanish language (in this chapter we translated it for easy reading) using SimpleNLG-ES [185]. For each case of the questionnaire, they had to assess the five statement items described in



HDL cholesterol was low at **45 mg/dL**. In addition, **an increasing trend of approximately 50%** has been detected with an initial value of 30 mg/dL until reaching 45 mg/dL. In particular, it should be noted that particularly low values have been reached during the month of May with **30 mg/dL**, on 06/15/2020 with **43 mg/dL** and between 08/10/2020 and 08/16/2020 with **44.5 mg/dL**. On the contrary, particularly high values were reached during the month of July with **50 mg/dL**, during the first week of September with **52 mg/dL** and on 03/06/2020 with **55 mg/dL**.

Figure 5.14: Example of the generated linguistic description from data of the HDL cholesterol.

(Table 5.23) using a 5-point Likert [222] scale in the range [1, 5], where 1 means "absolutely disagree" and 5 "absolutely agree" with the corresponding statement. Items Q1 and Q2 refer to the linguistic description content, whilst Q3-Q5 refer to manner or form. In addition, this questionnaire has a free-writing item where the professional was allowed to provide extra feedback, comments or indications for each case.

None of these experts participated in the definition of any part of the Data-To-Text system, since all the requisites were defined with the help of another different expert. Furthermore, in this evaluation, we emphasize the degree of conformity of the human experts with the generated texts, who performed a fully blind assessment. No details were provided about how the linguistic descriptions were generated.

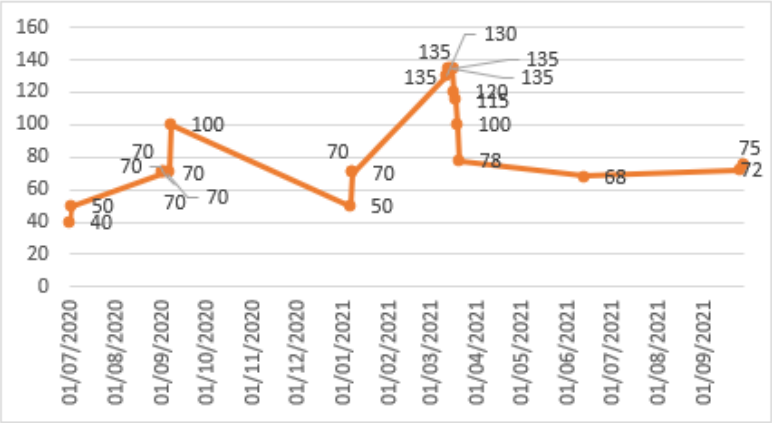
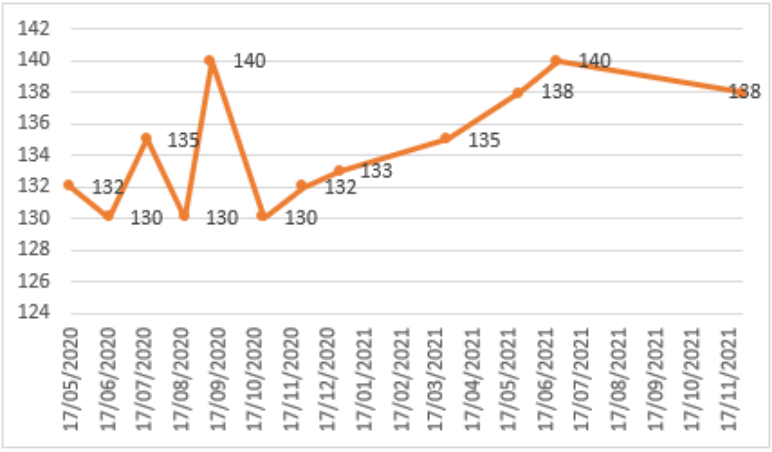


Figure 5.15: Example extracted from the questionnaire of the generated linguistic description from data of the in-hospital diastolic blood pressure.

Code	Item
Q1	The automatically generated linguistic description adequately explains the graph representing the data in the example
Q2	The linguistic description provided is similar to what you would do yourself to describe the evolution of the corresponding variable
Q3	Vocabulary is used correctly
Q4	The text contains the most relevant information related to the data
Q5	The text is correct in terms of spelling, punctuation and formatting (layout)

Table 5.23: Items of the expert evaluation questionnaire.

In Table 5.24 we present a summary of the experts' scores for each of the questions, the content (Q1, Q2) and layout (Q3, Q4, Q5) dimensions and the overall result. In general, the results show the experts agree with the presented linguistic descriptions since the scores average overall average is 3.81. Besides, mode shows the greatest used value is 4, and the



Blood glucose has been high, with a last record of **138 mg/dL**. In addition, an **increasing trend of approximately 5%** with an initial value of 132 until reaching 138 mg/dL is detected. In particular, it should be noted that particularly high values have been reached on 12/09/2020 and 27/06/2021 with **140 mg/dL**. In contrast, particularly low values have been reached on 18/06/2020, 20/08/2020 and 25/10/2020 with **130 mg/dL**.

Figure 5.16: Example extracted from the questionnaire of the generated linguistic description from data of the blood glucose.

	Average	SD	Mode	Median	IQR
Q1	3.40	1.23	2	4	2
Q2	3.26	1.23	2	3.5	2
Q3	4.26	0.70	4	4	1
Q4	3.64	1.23	4	4	3
Q5	4.48	0.51	4	4	1
Content	3.33	1.23	2	4	2
Layout	4.13	0.93	4	4	1
Overall	3.81	1.12	4	4	2

Table 5.24: Expert evaluation results from the questionnaire.

median is also 4 with an interquartile range of 2. However, the results show a low conformity of the experts with respect to the content of the linguistic descriptions (items Q1 and Q2) with

a mode of 2 whereas they agree with the layout of the generated linguistic descriptions (items Q3-Q5). The free-writing item of the questionnaire has allowed us to identify the following two weaknesses that explain this low score: *i)* indicate whether the values recorded for a variable are stable if applicable and *ii)* in the intensifications/exceptions there are cases where it could be confusing to talk about particularly high/low values as it could be understood as an anomaly. This suggests that it would be interesting to endow the system with some kind of personalization or adaption of the descriptions to each medical professional interest or information needs.

Overall, we can conclude that the generated linguistic descriptions are suitable both in content and form to describe medical temporal series, since the independent experts who assessed them indicated a high level of satisfaction with the automatically generated texts and their adequacy to this temporal series. Moreover, the experts' answer suggestions are very useful and will be taken into consideration to improve the quality of the descriptions generated by our system.

5.4 Discussion and final remarks

In the literature we found systems based on protoforms with simple structures that can be useful in some cases. Nevertheless, the information they provide is very limited and insufficient, especially in real use cases that have more demanding requirements. Thus, in this chapter, we presented a D2T model based on computational intelligence and natural language generation for the automatic linguistic description of numerical data series, including more complex description structures. Moreover, our model included a temporal ontology which covers a wide range of temporal references with the aim of addressing the problem of managing the inherent imprecision of the temporal knowledge, which is very relevant describing time series.

We provided two real use cases of application: *i)* in the meteorology field and *ii)* in the medical field. In the environmental information field, we developed the ICA2TEXT system, which provides linguistic descriptions about the Air Quality Index indicator (AQI), which measures the presence and density of several pollutants in the atmosphere. We designed a model to generate linguistic descriptions about real time series, which describes the AQI evolution and the prediction provided by the official Galician (NW Spain) Meteorology Agency, MeteoGalicia [154]. Results of the expert validation show the generated linguistic descriptions are very suitable since we obtained an average score of 4.37 for the content and 4.88

USC UNIVERSIDADE

for the linguistic description layout. Furthermore, ICA2TEXT has been integrated into MeteoGalicia's production website, so that after a period of testing, the deployment as a public service in the air quality section of MeteoGalicia's website is planned as a future work.

In the field of medicine, within the R+D project SMART CARDIA, we developed a system for the automatic generation of linguistic descriptions from the records of 19 different medical indicators for patients in the cardiac rehabilitation program developed in the cardiology area of the University Clinical Hospital of Santiago de Compostela (Galicia, NW Spain). The validation of this system was carried out by two expert cardiologists, with the aim of ensuring that the resulting linguistic descriptions complied with the requirements of the project in which it was developed. Results of the expert validation show the generated linguistic descriptions are suitable to describe this medical information since we obtained an overall average of 3.91.

Finally, we can conclude that the D2T model for the generation of linguistic descriptions of time series proposed in this chapter is suitable for the description of real data from different domains, which have very specific requirements, since it allows generating sufficiently representative linguistic descriptions of the information contained in the data.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER 6

CONCLUSIONS

The number of practical applications that can benefit from the extraction of the relevant information contained in their data increases as the amount of data generated is higher and, therefore, their manual analysis by humans is unachievable. This information is very important in many areas, such as decision making. In this context, in the literature arises the Natural Language Generation (NLG) field with the aim of generating comprehensible texts with relevant information from several sources of data. In particular, Data-To-Text systems (D2T), which can process large amounts of numerical or symbolical data, converting them into readable texts that contain relevant and understandable information for users.

Furthermore, in the literature, the application of fuzzy sets in the NLG field, and in particular D2T systems, has increased considerably in recent years, being of great use in these systems to tackle the problem of imprecision and uncertainty in natural language. These fields have been analyzed in the state-of-the-art, highlighting its strengths and weaknesses in Chapter 2 of this thesis. Many of the Linguistic Descriptions of Data (LDD) approaches lack the richness of NLG texts, as the fit into the unary (*"Q Y are S"*) and binary (*"Q KY are S"*) protoforms, which are not enough to fulfill most real application areas requeriments. Data-To-Text (D2T) systems based on Linguistic Descriptions of Data (LDD) have the problem of being insufficiently expressive to reach the level of detail that the data features and the related domain demand. The following three aspects of this problem have been addressed in this thesis:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- The content determination phase in NLG systems relies on extracting the set of the most representative set of fuzzy quantified sentences with respect to a data set as a search

space problem. Approaches in the literature are based both on heuristic and meta-heuristic algorithms However, this task has not been addressed yet in depth, especially with very large data sets.

- The generation of fuzzy quantified statements, also within the content determination stage, requires evaluating the candidate solutions to determine their representativeness with respect to the data with the aim of selecting the most suitable set of descriptions. Many fuzzy quantification methods have been proposed and theoretically studied in the literature, finding differences in their behavior from the point of view of their properties. Nevertheless, their behavior has not been studied yet from an experimental perspective.

- The proposed D2T systems lack in terms of the expressiveness of the resulting texts, since they are mainly based on unary and binary fuzzy quantified statements. Although these structures are useful in several scenarios, they provide general information that may lack in many areas, especially in real use cases, whose properties and requirements are more demanding.

Based on the weaknesses found in the literature, this thesis is based on the following hypothesis:
More expressive protoform structures and more efficient search methods are required in the design of Data-To-Text systems based on Linguistic Descriptions of Data to meet the needs identified in real use cases.

The work developed in this thesis has contributed to address these drawbacks of prior research in D2T systems based on LDD with the aim of meeting a set of objectives defined in Section 1.2. First of all, we have addressed the objective **O1** in Chapter 3 of this thesis, which is focused on the extraction of protoforms proposing meta-heuristic solutions from a data set in the content phase, approached as a search problem.

- In Section 3.1 we have presented the context of this problem, describing the two considered data sets (from observation and prediction meteorological data), and in Section 3.2 we have described the linguistic variables designed through fuzzy sets from the meteorological variables (sky, wind and temperature), the geographical descriptors and the quantifiers.

- Section 3.3 is composed of Section 3.3.1, were we have described deeply the design of the meta-heuristic approaches to address the problem of extracting LDD from the

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

data. Section 3.3.2 describes the metrics we have designed to assess the candidate fuzzy quantified descriptions, and in Section 3.3.3 we have described the measures to evaluate the resulting rankings from our meta-heuristic algorithms.

- Section 3.4 has introduced the results of the experiments. In Section 3.4.1, we have introduced the comparison between a random baseline and our meta-heuristic approaches to prove if the application of meta-heuristic solutions is justified. The results obtained both for unary descriptions, where the accuracy of the baseline was very low, and for binary descriptions, where the execution time was very high, justify the application of meta-heuristic solutions in these scenarios. Sections 3.4.2 and 3.4.3 have detailed the results of our meta-heuristic approaches with the considered data sets for the extraction of unary and binary candidates. Results of these experiments show the Genetic Algorithm performs better than the Simulated Annealing regardless of the data set features. In Section 3.4.5, we have shown the details of a validation performed by a meteorologist to assess the adequacy of the descriptions generated in the field of meteorology. Results of this assessment show no differences between our approaches in terms of quality in the unary case, whilst GA performs better than SA with regard to the quality in the binary case.

In Chapter 4, we have conducted an empirical study among the most widely used fuzzy quantification methods in the literature for the evaluation of unary and binary fuzzy quantified sentences to fulfill the objective **O2**. This chapter comprises an analysis of the properties of the methods considered and the two studies carried out.

- In Section 4.1, we have detailed the context of the experimental studies performed in this chapter, detailing the fuzzy quantification methods included in these studies. Moreover, we have also described the structure of the unary and fuzzy quantified statements considered. In Section 4.1.1, we have described the properties considered from the related work to theoretically analyze the behavior of the methods. In Sections 4.1.2 and 4.1.3, we have analyzed the fuzzy quantification methods in terms of these properties.

- Section 4.2 has described the details of the correlation comparison performed between the considered methods. In Section 4.2.1, we included the description of the involved data set and the design of the linguistic variables generated from the meteorological variables (sky, wind and temperature), the quantifiers and the geographical descriptions

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

to generate the fuzzy quantified statements. In Section 4.2.2, we have described the stages of the experimental execution with their inputs and outputs and the details of the correlation comparison. In Section 4.2.2 we have described the experimental results of this comparison study. Results of these experiments showed a nearly absolute correlation between all the pairs of methods (higher than 0.99 between all pairs of methods) for unary descriptions whilst for the binary case results are lower between the analyzed methods, finding three groups of methods with regard their correlation: *i)* Z_Q, GD_Q and ZS_Q, *ii)* Y_Q and *iii)* V_Q.

- In Section 4.3, we have presented the significance study between the selected fuzzy quantification methods. In Section 4.2.1, we have described the features of the considered 15 data sets and how the linguistic components have been designed. In Section 4.3.2, we described the stages of the experimental tests pipeline with their inputs and outputs. Moreover, we have detailed the process of the linguistic descriptions generation and the characteristics of the significance tests, defining the null hypothesis. In Section 4.3.3, we have discussed the results of the experiments. In the unary scenario, no significant differences were detected between the quantification methods, whilst for binary quantified descriptions, significant differences were detected in the behavior of pairs of quantification methods at least in a 50% of experiments, rejecting the null hypothesis.

Finally, the last objective **O3** has been developed in Chapter 5, which was focused on designing a D2T model that makes up for the shortcomings of existing models in terms of the simplicity of their descriptions. Furthermore, this objective also attempts to handle the temporal dimension found in many types of data.

- In Section 5.1, we have presented the proposed the Data-To-Text (D2T) model based on LDD. To manage the temporal dimension of the data, in Section 5.1.1 we have introduced the details of the temporal model defined within the D2T model, including its structure, consisting of a top-down hierarchy module and the definition of the temporary concepts included. In Section 5.1.2, we have described the architecture of the D2T model, detailing each of its stages: *i)* content determination, *ii)* document planning, *iii)* microplanning and *iv)* structure realization.
- In Section 5.2, we have described ICA2TEXT, a system developed as a real application of this model applied to the Air Quality Index (AQI) time series description.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Section 5.2.1 has detailed the definition of the necessary components to generate the linguistic descriptions, in particular, the calculation of the general labels and definition of the temporal references. In Section 5.2.5, we have described the validation of this system performed in collaboration with three expert meteorologists, where we have concluded the linguistic descriptions generated by our system were very suitable both in content and form to describe the data. ICA2TEXT is a system developed in collaboration with MeteoGalicia with the aim of provide useful information to the citizens. In Section 5.2.6, we described the details of the transference of the system to society, expecting to be deployed on the official MeteoGalicia website.

- Section 5.3 has included the specification of the reporting module developed within SMART CARDIA, an industrial R+D project in the medicine field, namely cardiac rehabilitation. In Section 5.3.1, we have detailed the tasks performed into each stage of the model pipeline. In Section 5.2.5, we have detailed the validation of this reporting module within this model with experts and examples of the resulting linguistic descriptions.

6.1 Future work

As a result of the work developed in this thesis, some directions of research that might be worthy to explore have been identified. These ideas focus on further improving the expresiveness of Data-To-Text (D2T) systems based on Linguistic Descriptions of Data (LDD) and expanding their applicability to other real-world problems. First of all, regarding the extraction of the most representative set of protoforms from a given data set within the content determination phase of a D2T system, this problem has not been addressed in depth in the literature, so that the existing proposals are very scarce. This is a critical phase within the pipeline that needs to be approached carefully to obtain the best set of descriptions, otherwise, the resulting text will not be as descriptive as it should. This remains an open problem in the literature.

In this regard, D2T systems are often highly conditioned by application case, so that most approaches are designed ad hoc. Our approaches have been developed so that they can be used with any type of numerical data as long as the input format matches that of the meta-heuristic solutions. Therefore, future work in this line should not only propose new approximations, but their characteristics should also allow researchers to compare solutions proposed in the literature easily with different data sources.

USC UNIVERSIDADE

Also within the content determination phase, as previously mentioned, the extraction of protoforms is a critical task, since obtaining a sub-optimal set will decrease its usefulness. Within this task, which is approached as a search problem, the evaluation of candidate protoforms is a very important task since their result is used to guide the search algorithm. The evaluation of fuzzy quantified sentences is performed using fuzzy quantification methods, which have been proposed and studied extensively theoretically in the literature. Within this thesis we have empirically studied the behavior of a set of widely used and well-known quantification methods, demonstrating that their selection in the evaluation of unary sentences can be guided by other criteria, since no significant differences have been found; notwithstanding, research should deepen the analysis of the behavior of quantification methods when evaluating binary quantified descriptions, which are widely used in the literature, whose behaviors have been shown to differ greatly in this task. Moreover, this experimentation has only focused on relative and coherent quantifiers. Since other types of quantifiers are used in the literature, researchers should extend the study to include them, as well as considering new fuzzy quantification methods, such as F^A [58], which allow evaluating different types of quantifiers as, for instance, semi-fuzzy; and more complex fuzzy quantified sentence structures.

Finally, regarding LDD-based D2T systems, most approaches are currently based on unary and binary quantified statements. In this thesis, we have proposed a model that allows the design of more complex structures that adapt to the needs of real use cases. Research should continue along these lines, proposing new models based on more complex structures that can be more descriptive in several scenarios. Furthermore, a fuzzy temporal model has been proposed in this model to handle the temporal dimension of the data. These models could be extended to add more complex references if necessary, moreover, other models could be included to address other critical dimensions in some data specifically.

Evaluating a D2T system is a difficult task, usually based on an expert evaluation by means of questionnaires. However, during the development of the ICA2TEXT and SMART CARDIA systems we have seen how difficult it is to perform this type of evaluation in real environments, since the availability of experts is usually very low as time availability is limited and it is not possible to count on the help of a large number of people. On the other hand, automatic metrics have proven not to be sufficient on their own, but as a complement to expert validations. Nowadays, there is an intense activity and discussion about the most appropriate methodologies in the NLG field in various contexts such as real applications or large data sets, where the number of possible descriptions is very large, and it is necessary to ensure that

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

the cases to be validated are representative. In addition, since this is an open issue, there are still problems within the evaluations that need to be addressed such as the presence of false positives.

6.2 Final remarks

The application of fuzzy sets to handle imprecision and uncertainty in NLG has proven to be a very promising line of research in the literature. These techniques are used showing very good results mainly in the content determination phase of an NLG system, however, other stages could also benefit from the use of fuzzy sets. The goal of using these techniques is to provide better systems that produce more human-friendly information from natural language texts by handling the imprecision of the language itself.

In the literature, we have found some shortcomings in the systems proposed. Therefore, in this thesis we have established a set of objectives, detailed in Chapter 1, focusing on the development of techniques with the aim of contributing to improve these problems. These objectives have been largely fulfilled in this thesis with very positive and promising results, however, this is still an open problem that needs a great attention in the literature.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Appendix: List of publications included in this thesis

This thesis reproduces the contents of the following publications:

- Chapter 3 reproduces the contents of A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín, *Meta-heuristics for generation of linguistic descriptions of weather data: experimental comparison of two approaches*. Fuzzy Sets and Systems, 2022. ISSN: 0165-0114. The publication is available at `https://doi.org/10.1016/j.fss.2022.02.016`. Journal Impact Factor: 3.343 (JCR 2020), 0.902 (SJR 2020). Journal ranked in JCR 2020, in Computer science, Theory & Methods (Q1, 23/110), Mathematics, Applied (D1, 18/265), Statistics & Probability (Q1, 20/125) categories.
  Author's contribution: research of the topic, conceptualization, design and implementation of the models, design and conduction of the experiments, and writing of the article.
  Reproduction authorization: This article has been published as open access. Elsevier allows its inclusion as part of doctoral theses without express permission (see `www.elsevier.com/about/policies/copyright#Author-rights`).

- Chapter 4 reproduces the contents of C. Heble-Lahera, A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín, *Empirical Study of Fuzzy Quantification Models for Linguistic Descriptions of Meteorological Data*. In proceedings of the 2020 IEEE International Conference of Fuzzy Systems (FUZZ-IEEE), pp. 1-7, Glasgow (United Kingdom), 2020. ISSN: 1558-4739. The publication is available at `https://doi.org/10.1109/FUZZ48607.2020.9177716`.
  Conference ranking (GRIN-SCIE): A-, Class 2.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Author's contributions: research, collaboration in the implementation of the models, design and supervision of the experiments, and writing of the article.
Reproduction authorization: IEEE allows the reuse of accepted papers in a thesis without express permission (see www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/permissions_faq.pdf)

- Chapter 4 reproduces the contents of A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín, *An experimental study on the use of fuzzy quantification models for linguistic descriptions of data*. In proceedings of the 24th European Conference of Artificial Intelligence (ECAI), pp. 267-274, Santiago de Compostela (Spain), 2020. ISBN: 978-1-64368-100-9 (print), 978-1-64368-101-6 (online). The publication is available at https://dx.doi.org/10.3233/FAIA325.
Conference Ranking (GRIN-SCIE): A, Class 2.
Author's contributions: research of the topic, conceptualization, design and implementation of the models, design and conduction of the experiments, and writing of the article.
Reproduction authorization: IOS Press allows to reuse articles under certain circumstances https://www.iospress.com/article-sharing-policy. In addition, we have received permission by mail from Carry Koolbergen, Contracts, Rights & Permissions Coordinator.

- Chapter 5 reproduces the contents of A. Cascallar-Fuentes, J. Gallego-Fernández, A. Ramos-Soto, A- Saunders, A. Bugarín, *Automatic generation of textual descriptions in data-to-text systems using a fuzzy temporal ontology: application in air quality index data series*. Applied Soft Computing, 119, 2022. ISSN: 1568-4946. The publication is available at https://doi.org/10.1016/j.asoc.2022.108612
Journal Impact Factor: 6.725 (JCR 2020) 1.290 (SJR 2020). Journal ranked in JCR 2020, in Computer Science, Artificial Intelligence (Q1, 24/139), and D1 Computer Science, Interdisciplinary applications (D1, 11/111) categories.
Author's contributions: research of the topic, conceptualization, design and implementation of the models, design and implementation of the systems, design and conduction of the experiments, and writing of the article.
Reproduction authorization: This article has been published as open access, therefore, Elsevier allows its inclusion as part of doctoral theses without express permission (see www.elsevier.com/about/policies/copyright#Author-rights).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Bibliography

[1] AccuWeather. AccuWeather website. `www.accuweather.com`, 2021. [Online; Accessed on September 2021].

[2] Esra Akgül, Yılmaz Delice, Emel Kızılkaya Aydoğan, and Fatih Emre Boran. An application of fuzzy linguistic summarization and fuzzy association rule mining to kansei engineering: a case study on cradle design. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–31, 2021.

[3] Ali M. Al-Khouri. Open data: A paradigm shift in the heart of government. *Journal of Public Administration and Governance*, 4:217–244, 2014.

[4] James F. Allen. Towards a General Theory of Action and Time. *Artificial Intelligence*, 23(2):123–154, 1984.

[5] Rui Jorge Almeida, Marie-Jeanne Lesot, Bernadette Bouchon-Meunier, Uzay Kaymak, and Gilles Moyse. Linguistic summaries of categorical time series for septic shock patient data. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, 2013.

[6] Tunahan Altintop, Ronald R. Yager, Diyar Akay, Fatih Emre Boran, and Muhammet Ünal. Fuzzy Linguistic Summarization with Genetic Algorithm: An Application with Operational and Financial Healthcare Data. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 25(04):599–620, 2017.

[7] Jacopo Amidei, Paul Piwek, and Alistair Willis. The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations. In *12th International Conference on Natural Language Generation (INLG)*, pages 397–402, 2019.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[8] Sena Aydogan, Diyar Akay, Fatih Emre Boran, and Ronald R. Yager. An Extension of Fuzzy Linguistic Summarization Considering Probabilistic Uncertainty. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 26(2):195–216, 2018.

[9] Silvana Badaloni and Massimiliano Giacomin. The algebra IAfuz: a framework for qualitative fuzzy temporal reasoning. *Artificial Intelligence*, 170(10):872–908, 2006.

[10] Senén Barro, Alberto Bugarín, Purificación Cariñena, and Félix Díaz-Hermida. A framework for fuzzy quantification models analysis. *IEEE Transactions on Fuzzy Systems*, 11(1):89–99, 2003.

[11] Senén Barro, Roque Marín, José Mira, and Alfonso R. Patón. A model and a language for the fuzzy representation and handling of time. *Fuzzy sets and Systems*, 61(2):153–175, 1994.

[12] Cristina Barros. *Proposal of a Hybrid Approach for Natural Language Generation and its Application to Human Language Technologies*. PhD thesis, Universidad de Alicante, 2019.

[13] Jon Barwise and Robin Cooper. Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence*, pages 241–301. 1981.

[14] Anja Belz. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4):431–455, 2008.

[15] Anja Belz and Ehud Reiter. Comparing Automatic and Human Evaluation of NLG Systems. In *11st Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006.

[16] Anja Belz, Mike White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. The First Surface Realisation Shared Task: Overview and Evaluation Results. In *13th European Workshop on Natural Language Generation (ENLG)*, pages 217–226, 2011.

[17] Kristin P. Bennett and Olvi L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization methods and software*, 1(1):23–34, 1992.

UDC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[18] Johan Benthem. Determiners and logic. *Linguistics and Philosophy*, 6(4):447–478, 1983.

[19] Ignacio Blanco, Miguel Delgado, María J. Martín-Bautista, Daniel. Sánchez, and María A. Vila. *Quantifier Guided Aggregation of Fuzzy Criteria with Associated Importances*, pages 272–287. 2002.

[20] Fatih E. Boran, Diyar Akay, and Ronald R. Yager. An overview of methods for linguistic summarization with fuzzy sets. *Expert Systems with Applications*, 61:356–377, 2016.

[21] Patrick Bosc. On the Comparison of the Sugeno and the Choquet Integrals for the Evaluation of Quantified Statements. In *3rd European Congress on Intelligent Techniques and Soft Computing (EUFIT)*, 1995.

[22] Patrick Bosc and Ludovic Lietard. Monotonic quantified statements and fuzzy integrals. In *Proceedings of the First International Joint Conference of The North American Fuzzy Information Processing Society Biannual Conference. The Industrial Fuzzy Control and Intelligence (NAFIPS/IFIS/NASA)*, pages 8–12, 1994.

[23] Patrick Bosc, Ludovic Lietard, and Olivier Pivert. Quantified Statements and Database Fuzzy Querying. In *Fuzziness in Database Management Systems*, pages 275–308, 1995.

[24] Khedidja Boulanouar, Allel Hadjali, and Mohand Lagha. A Hybrid Approach for Linguistic Summarization of Time Series. In *International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pages 1–5, 2020.

[25] Michel Boyer and Guy Lapalme. Generating paraphrases from meaning-text semantic networks. *Computational Intelligence*, 1(1):103–117, 1985.

[26] Alberto Bugarín, Nicolás Marín, Daniel Sánchez, and Gracián Triviño. Aspects of quality evaluation in linguistic descriptions of data. In *IEEE International Conference on Fuzzy Systems, FUZZ-IEEE*, pages 1–8, 2015.

[27] Stephan Busemann and Helmut Horacek. Generating air quality reports from environmental data. In *DFKI Workshop on Natural Language Generation*, pages 15–21, 1997.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[28] Erion Çano and Ondrej Bojar. Automating Text Naturalness Evaluation of NLG Systems. *CoRR*, abs/2006.13268, 2020.

[29] Giuseppe Carenini and Johanna D. Moore. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925–952, 2006.

[30] Purificación Cariñena, Alberto Bugarín, Manuel Mucientes, Félix Díaz-Hermida, and Senén Barro. *Fuzzy Temporal Rules: A Rule-based Approach for Fuzzy Temporal Knowledge Representation and Reasoning*, pages 237–250. 2002.

[31] Andrea Cascallar-Fuentes, Alejandro Ramos-Soto, and Alberto Bugarín. Adapting SimpleNLG to Galician language. In *11th International Conference on Natural Language Generation (INLG)*, pages 67–72, 2018.

[32] Andrea Cascallar-Fuentes, Alejandro Ramos-Soto, and Alberto Bugarín-Diz. An experimental study on the use of fuzzy quantification models for linguistic descriptions of data. In *24th European Conference on Artificial Intelligence (ECAI)*, pages 267–274. IOS Press, 2020.

[33] Andrea Cascallar-Fuentes, Alejandro Ramos-Soto, and Alberto J Bugarın-Diz. Descripciones lingüısticas de datos de observación meteorólogica usando temple simulado. In *XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018)*, pages 469–474. Asociación Española para la Inteligencia Artificial (AEPIA), 2018.

[34] Rita Castillo-Ortega, Nicolás Marín, Carmen Martínez-Cruz, and Daniel Sánchez. A proposal for the hierarchical segmentation of time series. Application to trend-based linguistic description. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 489–496, 2014.

[35] Rita Castillo-Ortega, Nicolás Marín, and Daniel Sánchez. Time Series Comparison Using Linguistic Fuzzy Techniques. In *13th International Conference on Information Processing and Management of Uncertainty (IPMU)*, volume 6178, pages 330–339, 2010.

[36] Rita Castillo-Ortega, Nicolás Marín, and Daniel Sánchez. A Fuzzy Approach to the Linguistic Summarization of Time Series. *Journal of Multiple-Valued Logic*, 17(2-3):157–182, 2011.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[37] Rita Castillo-Ortega, Nicolás Marín, and Daniel Sánchez. Linguistic local change comparison of time series. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 2909–2915. IEEE, 2011.

[38] Rita Castillo-Ortega, Nicolás Marín, Daniel Sánchez, and Andrea Tettamanzi. Linguistic summarization of time series data using genetic algorithms. In *EUSFLAT*, volume 1, pages 416–423, 2011.

[39] Rita Castillo-Ortega, Nicolás Marín, Daniel Sánchez, and Andrea Tettamanzi. Quality Assessment in Linguistic Summaries of Data. In *Advances on Computational Intelligence - 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, volume 298, pages 285–294, 2012.

[40] Arun Tejasvi Chaganty, Stephen Mussmann, and Percy Liang. The price of debiasing automatic metrics in natural language evalaution. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 643–653, 2018.

[41] Ernie Chang, Xiaoyu Shen, Dawei Zhu, Vera Demberg, and Hui Su. Neural Data-to-Text Generation with LM-based Text Augmentation. In *16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 758–768, 2021.

[42] Kai Chen, Fayuan Li, Baotian Hu, Weihua Peng, Qingcai Chen, Hong Yu, and Yang Xiang. Neural data-to-text generation with dynamic content planning. *Knowledge Based Systems*, 215:106610, 2021.

[43] José Coch. System Demonstration Interactive Generation And Knowledge Administration In Multimeteo. In *9th International Workshop on Natural Language Generation(INLG)*, 1998.

[44] José Coch. Multimeteo: multilingual production of weather forecasts. *ELRA Newsletter*, 3(2), 1998.

[45] Patricia Conde-Clemente, José Maria Alonso, Éldman de Oliveira Nunes, Ángel Sánchez, and Gracián Triviño. New types of computational perceptions: Linguistic descriptions in deforestation analysis. *Expert Systems With Applications*, 85:46–60, 2017.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[46] Licong Cui and Yongming Li. Linguistic quantifiers based on Choquet integrals. *International Journal of Approximate Reasoning*, 48(2):559–582, 2008.

[47] Jacek Czerniak and Hubert Zarzycki. Application of rough sets in the presumptive diagnosis of urinary system diseases. In *Artificial Intelligence and Security in Computing Systems*, pages 41–51. Springer, 2003.

[48] Robert Dale, Barbara Di Eugenio, and Donia Scott. Introduction to the special issue on natural language generation. *Computational Linguistics*, 24(3):345–353, 1998.

[49] Robert Dale and Ehud Reiter. Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, 19(2):233–263, 1995.

[50] Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750–757, 2008.

[51] Miguel Delgado, M. Dolores Ruiz, Daniel Sánchez, and M. Amparo Vila. Fuzzy quantification: a state of the art. *Fuzzy Sets and Systems*, 242:1–30, 2014.

[52] Miguel Delgado, Daniel Sánchez, Maria J. Martín-Bautista, and M. Amparo Vila. A probabilistic definition of a nonconvex fuzzy cardinality. *Fuzzy Sets and Systems*, 126(2):177–190, 2002.

[53] Miguel Delgado, Daniel Sánchez, and M. Amparo Vila. Un método para la evaluación de sentencias con cuantificadores lingüísticos. In *III Congreso Español sobre Tecnologías y Lógica Fuzzy*, pages 193–198, 1998.

[54] Miguel Delgado, Daniel Sánchez, and M. Amparo Vila. Fuzzy cardinality based evaluation of quantified sentences. *International Journal of Approximate Reasoning*, 23(1):23–66, 2000.

[55] Nina Dethlefs. Context-Sensitive Natural Language Generation: From Knowledge-Driven to Data-Driven Techniques. *Language & Linguistics Compass*, 8(3):99–115, 2014.

[56] Félix Díaz-Hermida. *Modelos de cuantificación borrosa basados en una interpretación probabilıstica y su aplicación en recuperación de información*. PhD thesis, Universidade de Santiago de Compostela, 2006.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[57] Félix Díaz-Hermida and Alberto Bugarín. Semi-fuzzy quantifiers as a tool for building linguistic summaries of data patterns. In *IEEE Symposium on Foundations of Computational Intelligence (FOCI)*, pages 45–52, 2011.

[58] Félix Díaz-Hermida, Alberto Bugarín, and David E. Losada. The probatilistic quantifier fuzzification mechanism FA: A theoretical analysis. *CoRR*, abs/1410.7223, 2014.

[59] Félix Díaz-Hermida, Purificación Cariñena, and Alberto Bugarín. Efficient Evaluation of Similarity Quantified Expressions in the Temporal Domain. In *Soft Methods for Integrated Uncertainty Modelling*, pages 191–198. 2006.

[60] Félix Díaz-Hermida, David E. Losada, Alberto Bugarín, and Senén Barro. A Probabilistic Quantifier Fuzzification Mechanism: The Model and Its Evaluation for Information Retrieval. *IEEE Transactions on Fuzzy Systems*, 13(5):688–700, 2005.

[61] Félix Díaz-Hermida, Marcos Matabuena, and Juan Carlos Vidal. The FA Quantifier Fuzzification Mechanism: analysis of convergence and efficient implementations. *CoRR*, abs/1902.02132, 2019.

[62] Félix Díaz-Hermida, Martin Pereira-Fariña, Juan Carlos Vidal, and Alejandro Ramos-Soto. Characterizing Quantifier Fuzzification Mechanisms: A behavioral guide for applications. *Fuzzy Sets and Systems*, 345:1–23, 2018.

[63] Félix Díaz-Hermida, Alejandro Ramos-Soto, and Alberto Bugarín. On the role of fuzzy quantified statements in linguistic summarization of data. In *11th International Conference on Intelligent Systems Design and Applications, (ISDA)*, pages 166–171. IEEE, 2011.

[64] Chrysanne DiMarco, H Dominic Covvey, Peter Bray, Donald Cowan, Vic DiCiccio, Eduard Hovy, Joan Lipa, and Doug Mulholland. The development of a natural language generation system for personalized e-health information. *Medinfo*, 2007, 2007.

[65] Carlos A Donis-Díaz, Rafael Bello, and Janusz Kacprzyk. Using ant colony optimization and genetic algorithms for the linguistic summarization of creep data. In *Intelligent Systems*, pages 81–92. 2015.

[66] Dheeru Dua and Casey Graff. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml`, 2017. [Online; Accessed September 2021].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[67] Didier Dubois and Henri Prade. *Possibility Theory - An Approach to Computerized Processing of Uncertainty*. Springer, 1988.

[68] Didier Dubois and Henri Prade. Processing fuzzy temporal knowledge. *IEEE Transactions on Systems, Man and Cybernetics*, 19(4):729–744, 1989.

[69] Pablo Ariel Duboue and Kathleen R. McKeown. Statistical Acquisition of Content Selection Rules for Natural Language Generation. In *Conference on Empirical Methods in Natural Language Processing*, pages 121–128, 2003.

[70] Sourmitra Dutta. Temporal reasoning in medical expert systems. In *First Annual IEEE Symposium on Computer-Based Medical Systems (CBMS)*, pages 118–122, 1988.

[71] Luka Eciolaza, Martin Pereira-Fariña, and Gracián Triviño. Automatic linguistic reporting in driving simulation environments. *Applied Soft Computing*, 13(9):3956–3967, 2013.

[72] Luka Eciolaza and Gracián Triviño. Linguistic reporting of driver behavior: Summary and event description. In *11th International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 148–153, 2011.

[73] Agoston E. Eiben, James E. Smith, et al. *Introduction to evolutionary computing*, volume 53. Springer, 2003.

[74] Dominic Espinosa, Rajakrishnan Rajkumar, Michael White, and Shoshana Berleant. Further Meta-Evaluation of Broad-Coverage Surface Realization. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 564–574, 2010.

[75] European Environment Agency. European Air Quality Index website. `www.eea.europa.eu`, 2021. [Online; Accessed September 2021].

[76] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we Need hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15(90):3133–3181, 2014.

[77] Manuel. Fernández-Delgado, Manisha Sirsat, Eva Cernadas, Sadi Alawadi, Senén Barro, and Manuel Febrero-Bande. An extensive experimental survey of regression methods. *Neural Networks*, 111(March):11–34, 2019.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[78] Richard Fikes and Nils J. Nilsson. STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. *Artificial Intelligence*, 2(3/4):189–208, 1971.

[79] Katja Filippova and Michael Strube. Generating Constituent Order in German Clauses. In *45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

[80] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

[81] Changzeng Fu, Chaoran Liu, Carlos Toshinori Ishi, Yuichiro Yoshikawa, Takamasa Iio, and Hiroshi Ishiguro. Using an Android Robot to Improve Social Connectedness by Sharing Recent Experiences of Group Members in Human-robot Conversations. *IEEE Robotics and Automation Letters*, 2021.

[82] Tian Gao. A Learning and Practicing System to Support Effective Poetry Generation Based on Neural Network. In *International Conference on Bioinformatics and Intelligent Computing*, pages 12–18, 2021.

[83] Juan Moreno García, Javier Abián-Vicén, Luis Jiménez Linares, and Luis Rodriguez-Benítez. Description of multivariate time series by means of trends characterization in the fuzzy domain. *Fuzzy Sets and Systems*, 285:118–139, 2016.

[84] Konstantina Garoufi. Planning-Based Models of Natural Language Generation. *Language and Linguistics Compass*, 8(1):1–10, 2014.

[85] Konstantina Garoufi and Alexander Koller. Generation of effective referring expressions in situated context. *Language, Cognition and Neuroscience*, 29(8):986–1001, 2014.

[86] Albert Gatt and Emiel Krahmer. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[87] Albert Gatt and François Portet. Multilingual generation of uncertain temporal expressions from data: A study of a possibilistic formalism and its consistency with human subjective evaluations. *Fuzzy Sets and Systems*, 285:73–93, 2016.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[88] Albert Gatt and Ehud Reiter. SimpleNLG: A Realisation Engine for Practical Applications. In *12th European Workshop on Natural Language Generation (ENLG)*, pages 90–93, 2009.

[89] Donald Gillies. *Artificial Intelligence and Scientific Method*. Oxford University Press, 1996.

[90] Ingo Glöckner. DFS–An Axiomatic Approach to Fuzzy Quantification. *TR97-06, Techn. Fakultät, Univ. Bielefeld*, 2(3):10, 1997.

[91] Eli Goldberg, Norbert Driedger, and Richard I. Kittredge. Using Natural-Language Processing to Produce Weather Forecasts. *IEEE Expert*, 9(2):45–53, 1994.

[92] Google. Google Earth website. `https://earth.google.com`, 2021. [Online; Accessed September 2021].

[93] Google. Google Maps website. `https://www.google.es/maps/`, 2021. [Online; Accessed June 2021].

[94] Herbert P. Grice. Logic and Conversation. In *Speech Acts*, pages 41–58. Brill, 1975.

[95] Michael Alexander Kirkwood Halliday, Christian MIM Matthiessen, Michael Halliday, and Christian Matthiessen. *An introduction to functional grammar*. Routledge, 2014.

[96] Javeria Hassan, Muhammad Ali Tahir, and Adnan Ali. Natural language understanding of map navigation queries in Roman Urdu by joint entity and intent determination. *PeerJ Computer Science*, 7:15, 2021.

[97] Hakan Hatipoglu, Fatih Emre Boran, Mustafa Avci, and Diyar Akay. Linguistic Summarization of Europe Brent Spot Price Time Series Along with the Interpretations from the Perspective of Turkey. *International Journal of Intelligent Systems*, 29(10):946–970, 2014.

[98] Balraj S Heran, Jenny M. H. Chen, Shah Ebrahim, Tiffany Moxham, Neil Oldridge, Karen Rees, David R Thompson, and Rod S Taylor. Exercise-based cardiac rehabilitation for coronary heart disease. *Cochrane database of systematic reviews*, (7), 2011.

[99] Jerry R. Hobbs and Feng Pan. An ontology of time for the semantic web. *ACM Transactions on Asian Language Information Processing*, 3(1):66–85, 2004.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[100] Saar Hommes, Chris van der Lee, Felix Clouth, Jeroen Vermunt, Xander Verbeek, and Emiel Krahmer. A Personalized Data-to-Text Support Tool for Cancer Patients. In *12th International Conference on Natural Language Generation (INLG)*, pages 443–452, 2019.

[101] Assaf Hoogi, Arjun Mishra, Francisco Gimenez, Jeffrey Dong, and Daniel L. Rubin. Natural Language Generation Model for Mammography Reports Simulation. *IEEE Journal on Biomedical and Health Informatics*, 24(9):2711–2717, 2020.

[102] David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In *13th International Conference on Natural Language Generation (INLG)*, pages 169–182, 2020.

[103] James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine*, 56(3):157–172, 2012.

[104] Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. Text Revision: A Model and Its Implementation. In *Aspects of Automated Natural Language Generation, 6th International Workshop on Natural Language Generation*, volume 587 of *Lecture Notes in Computer Science*, pages 215–230, 1992.

[105] Nan Jiang, Jing Chen, Ri-Gui Zhou, Changxing Wu, Honglong Chen, Jiaqi Zheng, and Tao Wan. PAN: Pipeline assisted neural networks model for data-to-text generation in social internet of things. *Information Sciences*, 530:167–179, 2020.

[106] Backus John. The syntax and semantics of the proposed international algebraic language of the Zurich ACMGAMM conference. *International Conference on Information Processing*, pages 125–132, 1960.

[107] Karen Sparck Jones and Julia R. Galliers. *Evaluating Natural Language Processing Systems*. Springer-Verlag Berlin Heidelberg, 1995.

[108] Aravind K. Joshi and Yves Schabes. Tree-Adjoining Grammars. In *Handbook of Formal Languages, Volume 3: Beyond Words*, pages 69–123. Springer, 1997.

UDC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[109] Janusz Kacprzyk and Anna Wilbik. Using Fuzzy Linguistic Summaries for the Comparison of Time Series: an application to the analysis of investment fund quotations. In *2009 International Fuzzy Systems Association and 2009 World Congress European Society of Fuzzy Logic and Technology Conference*, pages 1321–1326, 2009.

[110] Janusz Kacprzyk and Anna Wilbik. Temporal Linguistic Summaries of Time Series Using Fuzzy Logic. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods (IPMU)*, volume 80 of *Communications in Computer and Information Science*, pages 436–445, 2010.

[111] Janusz Kacprzyk, Anna Wilbik, and Sławomir Zadrożny. A Linguistic Quantifier Based Aggregation for a Human Consistent Summarization of Time Series. In *2006 International Workshop on Soft Methods in Probability and Statistics (SMPS)*, volume 37 of *Advances in Soft Computing*, pages 183–190, 2006.

[112] Janusz Kacprzyk, Anna Wilbik, and Sławomir Zadrożny. On some types of linguistic summaries of time series. In *3rd International IEEE Conference Intelligent Systems*, pages 373–378, 2006.

[113] Janusz Kacprzyk, Anna Wilbik, and Sławomir Zadrożny. Linguistic Summarization of Time Series by Using the Choquet Integral. In *12th International Fuzzy Systems Association World Congress (IFSA)*, volume 4529 of *Lecture Notes in Computer Science*, pages 284–294, 2007.

[114] Janusz Kacprzyk, Anna Wilbik, and Sławomir Zadrożny. An approach to the linguistic summarization of time series using a fuzzy quantifier driven aggregation. *International Journal of Intelligent Systems*, 25(5):411–439, 2010.

[115] Janusz Kacprzyk and Ronald R. Yager. Using Fuzzy Logic with Linguistic Quantifiers in Multiobjective Decision Making and Optimization: A Step Towards More Human-Consistent Models. In *Stochastic Versus Fuzzy Approaches to Multiobjective Mathematical Programming under Uncertainty*, pages 331–350. 1990.

[116] Janusz Kacprzyk and Ronald R Yager. Linguistic summaries of data using fuzzy logic. *International Journal of General System*, 30(2):133–154, 2001.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[117] Janusz Kacprzyk and Sławomir Zadrożny. Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural-language generation. *IEEE Transactions on Fuzzy Systems*, 18(3):461–472, 2010.

[118] Janusz Kacprzyk and Sławomir Zadrożny. Comprehensiveness of Linguistic Data Summaries: A Crucial Role of Protoforms. In *Computational Intelligence in Intelligent Data Analysis*, pages 207–221, 2013.

[119] Janusz Kacprzyk and Sławomir Zadrożny. Fuzzy logic-based linguistic summaries of time series: a powerful tool for discovering knowledge on time varying processes and systems under imprecision. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(1):37–46, 2016.

[120] Janusz Kacprzyk and Sławomir Zadrożny. Linguistic summarization of the contents of Web server logs via the Ordered Weighted Averaging (OWA) operators. *Fuzzy Sets and Systems*, 285:182–198, 2016.

[121] Janusz Kacprzyk, Sławomir Zadrożny, and Andrzej Ziołkowski. FQUERY III+: A "human-consistent" database querying system based on fuzzy logic with linguistic quantifiers. *Information Systems*, 14(6):443–453, 1989.

[122] Katarzyna Kaczmarek-Majer, Olgierd Hryniewicz, Monika Dominiak, and Łukasz Święcicki. Personalized linguistic summaries in smartphone-based monitoring of bipolar disorder patients. In *11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT)*, pages 400–407, 2019.

[123] Mihir Kale and Scott Roy. Machine Translation Pre-training for Data-to-Text Generation - A Case Study in Czech. In *13th International Conference on Natural Language Generation (INLG)*, pages 91–96, 2020.

[124] Edward L Keenan and Dag Westerståhl. Generalized quantifiers in linguistics and logic. In *Handbook of logic and language*, pages 837–893. 1997.

[125] Zhang Kerui, Hu Haichao, and Liu Yuxia. Automatic Text Summarization on Social Media. In *4th International Symposium on Computer Science and Intelligent Control*, pages 1–5, 2020.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[126] Hauke Kindler, Dirk Densow, and Theodor M. Fliedner. A pragmatic implementation of medical temporal reasoning for clinical medicine. *Computers in Biology and Medicine*, 28(2):105–120, 1998.

[127] Richard Kittredge, Alain Polguere, and Eli Goldberg. Synthesizing weather forecasts from formatted data. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*, 1986.

[128] Alexander Koller and Kristina Striegnitz. Generation as Dependency Parsing. In *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 17–24, 2002.

[129] Ioannis Konstas and Mirella Lapata. A Global Model for Concept-to-Text Generation. *Journal of Artificial Intelligence Research*, 48:305–346, 2013.

[130] Dirk Kraus. Suregen-2: A shell system for the generation of clinical documents. In *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 215–218, 2003.

[131] Werasak Kurutach. Modelling fuzzy interval-based temporal information: a temporal database perspective. In *IEEE International Conference on Fuzzy Systems*, volume 2, pages 741–748, 1995.

[132] Willem J.M. Levelt. *Speaking: From intention to articulation*, volume 1. MIT press, 1993.

[133] Willem J.M. Levelt. Producing spoken language: A blueprint of the speaker. The neurocognition of language, 1999.

[134] Percy Liang, Michael I. Jordan, and Dan Klein. Learning Semantic Correspondences with Less Supervision. In *47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99, 2009.

[135] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, 2004.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[136] Shih-Wei Lin, Jatinder Gupta, Kuo-Ching Ying, and Zne-Jung Lee. Using simulated annealing to schedule a flowshop manufacturing cell with sequence-dependent family setup times. *International Journal of Production Research*, 47:3205–3217, 2009.

[137] Hiroshi Maeda, Shinichiro Asaoka, and Shuta Murakami. Dynamical fuzzy reasoning and its application to system modeling. *Fuzzy Sets and Systems*, 80(1):101–109, 1996.

[138] Saad Mahamood and Ehud Reiter. Working with Clinicians to Improve a Patient-Information NLG System. In *7th International Natural Language Generation Conference (INLG)*, pages 100–104. The Association for Computer Linguistics, 2012.

[139] Y. M. W. Mak, W. K. Chan, and C. S. S. Yue. Barriers to participation in a phase II cardiac rehabilitation programme. *Hong Kong medical journal*, 11:472–475, 2005.

[140] Vida Maliene, Vytautas Grigonis, Vytautas Palevičius, and Sam Griffiths. Geographic information system: Old principles with new capabilities. *Urban Design International*, 16(1):1–6, 2011.

[141] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60, 1947.

[142] William C. Mann and James A. Moore. Computer Generation of Multiparagraph English Text. *American Journal of Computational Linguistics*, 7(1):17–29, 1981.

[143] Diego Marcheggiani and Laura Perez-Beltrachini. Deep Graph Convolutional Encoders for Structured Data to Text Generation. In *11th International Conference on Natural Language (INLG)*, pages 1–9, 2018.

[144] Nicolás Marín and Daniel Sánchez. Fuzzy sets and systems + natural language generation: A step forward in the linguistic description of time series. *Fuzzy Sets and Systems*, 285:1–5, 2016.

[145] Nicolás Marín and Daniel Sánchez. On generating linguistic descriptions of time series. *Fuzzy Sets and Systems*, 285:6–30, 2016.

[146] Carmen Martínez-Cruz, Antonio J. Rueda, Mihail Popescu, and James M. Keller. A New Linguistic Description Approach for Time Series and its Application to Bed Rest-

UGC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

lessness Monitoring for Eldercare. *IEEE Transactions on Fuzzy Systems*, pages 1–1, 2021.

[147] Sergio Martínez-Municio, Luis Rodríguez-Benítez, Ester Castillo-Herrera, Juan Giralt-Muiña, and Luis Jiménez-Linares. Modelado lingüístico y síntesis de series temporales heterogéneas de consumo energético. In *XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018) 23-26 de octubre de 2018 Granada, España*, pages 475–480, 2018.

[148] Carmen Martínez, Javier Medina Quero, Macarena Espinilla, and Sergio D. Gramajo. Analyzing daily behaviours from wearable trackers using linguistic protoforms and fuzzy clustering. In *Workshop on Intelligent Information Processing and Natural Language Generation*, pages 1–9, 2020.

[149] Daniel D. McCracken and Edwin D. Reilly. Backus-Naur Form (BNF). In *Encyclopedia of Computer Science*, page 129–131. 2003.

[150] Javier Medina, Macarena Espinilla, and Christopher Nugent. Real-time fuzzy linguistic analysis of anomalies from medical monitoring devices on data streams. In *10th EAI International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth*, pages 300–303, 2016.

[151] Chris Mellish, Donia Scott, Lynne J. Cahill, Daniel S. Paiva, Roger Evans, and Mike Reape. A Reference Architecture for Natural Language Generation Systems. *Natural Language Engineering*, 12(1):1–34, 2006.

[152] Carlos Menéndez-González and Gracián Triviño. OLAP navigation in the Granular Linguistic Model of a Phenomenon. In *IEEE Symposium on Computational Intelligence and Data Mining, (CIDM)*, pages 260–267, 2011.

[153] Carlos Menéndez-González and Gracián Triviño. Selection of the Best Suitable Sentences in Linguistic Descriptions of Data. In *Advances on Computational Intelligence - 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, volume 298 of *Communications in Computer and Information Science*, pages 295–304, 2012.

[154] MeteoGalicia. Galician Meteorological Agency website. `www.meteogalicia.gal`, 2021. [Online; Accessed June 2021].

UDC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[155] Meteosat. Meteosat website. `www.meteosat.com`, 2021. [Online; Accessed September 2021].

[156] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa C. Mensch, Alexander C. Berg, Tamara L. Berg, and Hal Daumé III. Midge: Generating Image Descriptions From Computer Vision Detections. In *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 747–756, 2012.

[157] Juan Moreno-García, Antonio Moreno-García, Luis Jiménez-Linares, and Luis Rodríguez-Benítez. Describing Time Series Using Fuzzy Piecewise Linear Segments. In *Computational Intelligence and Mathematics for Tackling Complex Problems*, pages 149–155. 2020.

[158] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[159] Gilles Moyse and Marie-Jeanne Lesot. Linguistic summaries of locally periodic time series. *Fuzzy Sets and Systems*, 285:94–117, 2016.

[160] Gilles Moyse, Marie-Jeanne Lesot, and Bernadette Bouchon-Meunier. Mathematical Morphology Tools to Evaluate Periodic Linguistic Summaries. In *Flexible Query Answering Systems - 10th International Conference (FQAS)*, volume 8132 of *Lecture Notes in Computer Science*, pages 257–268, 2013.

[161] Soichiro Murakami, Sora Tanaka, Masatsugu Hangyo, Hidetaka Kamigaito, Kotaro Funakoshi, Hiroya Takamura, and Manabu Okumura. Generating Weather Comments from Meteorological Simulations. In *16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1462–1473, 2021.

[162] Gábor Nagypál and Boris Motik. A Fuzzy Model for Representing Uncertain, Subjective, and Vague Temporal Knowledge in Ontologies. In *On The Move to Meaningful Internet Systems: CoopIS, DOA, and ODBASE*, volume 2888 of *Lecture Notes in Computer Science*, pages 906–923, 2003.

[163] Mahdi Namazifar, Alexandros Papangelis, Gökhan Tür, and Dilek Hakkani-Tür. Language Model is all You Need: Natural Language Understanding as Question Answer-

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7803–7807, 2021.

[164] Peter Nemenyi. *Distribution-free multiple comparisons*. PhD thesis, PhD thesis, Princeton University, 1963.

[165] Feng Nie, Jinpeng Wang, Jin-Ge Yao, Rong Pan, and Chin-Yew Lin. Operations Guided Neural Networks for High Fidelity Data-To-Text Generation. *CoRR*, abs/1809.02735, 2018.

[166] Sergei Nirenburg, Victor R. Lesser, and Eric Nyberg. Controlling a Language Generation Planner. In *11th International Joint Conference on Artificial Intelligence*, pages 1524–1530, 1989.

[167] Hans-Joachim Novak. Generating a Coherent Text Describing a Traffic Scene. In *11th Coference on Computational Linguistics*, page 570–575, 1986.

[168] Vilém Novák. Linguistic characterization of time series. *Fuzzy Sets and Systems*, 285:52–72, 2016.

[169] Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. Why We Need New Evaluation Metrics for NLG. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2241–2252, 2017.

[170] Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. RankME: Reliable Human Ratings for Natural Language Generation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 72–78, 2018.

[171] Hans Jürgen Ohlbach. Relations Between Fuzzy Time Intervals. In *11th TIME International Symposium on Temporal Representation and Reasoning*, pages 44–51, 2004.

[172] Feng Pan. *Representing complex temporal phenomena for the semantic web and natural language*. University of Southern California, 2007.

[173] Feng Pan and Jerry R Hobbs. Time in OWL-S. In *AAAI Spring Symposium on Semantic Web Services*, pages 29–36, 2004.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[174] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[175] María Dolores Peláez-Coca, Macarena Espinilla, María Rosa Fernández Olmo, and Javier Medina. Fuzzy Linguistic Protoforms to Summarize Heart Rate Streams of Patients with Ischemic Heart Disease. *Complexity*, 2019:2694126:1–2694126:11, 2019.

[176] François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816, 2009.

[177] Javier Medina Quero, Miguel López-Medina, Alberto Salguero Hidalgo, and Macarena Espinilla. Predicting the Urgency Demand of COPD Patients From Environmental Sensors Within Smart Cities With High-Environmental Sensitivity. *IEEE Access*, 6:25081–25089, 2018.

[178] Da qun Qian. Representation and use of imprecise temporal knowledge in dynamic systems. *Fuzzy Sets and Systems*, 50(1):59–77, 1992.

[179] Swapan Raha and Kumar Sankar Ray. Approximate reasoning with time. *Fuzzy Sets and Systems*, 107(1):59–79, 1999.

[180] Alejandro Ramos-Soto, Alberto Bugarín, and Senén Barro. On the role of linguistic descriptions of data in the building of natural language generation systems. *Fuzzy Sets and Systems*, 285:31–51, 2016.

[181] Alejandro Ramos-Soto, Alberto Bugarín, Senén Barro, Nuria Gallego, Catalina Rodríguez, Iván Fraga, and Anthony Saunders. Automatic Generation of Air Quality Index Textual Forecasts Using a Data-To-Text Approach. In *Advances in Artificial Intelligence - 16th Conference of the Spanish Association for Artificial Intelligence (CAEPIA)*, volume 9422 of *Lecture Notes in Computer Science*, pages 164–174, 2015.

[182] Alejandro Ramos-Soto, Alberto Bugarín, Senén Barro, and Juan Taboada. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems*, 23(1):44–57, 2015.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[183] Alejandro Ramos-Soto, Alberto Bugarín, Senén Barro, and Félix Díaz-Hermida. Automatic Linguistic Descriptions of Meteorological Data. A soft computing approach for converting Open Data to Open Information. In *8th Iberian Conference on Information Systems and Technologies*, pages 728–733, 2013.

[184] Alejandro Ramos-Soto, Alberto José Bugarín Diz, Senén Barro, and Juan Taboada. Linguistic Descriptions for Automatic Generation of Textual Short-Term Weather Forecasts on Real Prediction Data. *IEEE Transactions on Fuzzy Systems*, 23(1):44–57, 2015.

[185] Alejandro Ramos-Soto, Julio Janeiro Gallardo, and Alberto Bugarín. Adapting SimpleNLG to Spanish. In *10th International Conference on Natural Language Generation (INLG)*, pages 144–148, 2017.

[186] Raimundo Real and Juan M. Vargas. The Probabilistic Basis of Jaccard's Index of Similarity. *Systematic Biology*, 45(3):380–385, 1996.

[187] Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. Controlling Hallucinations at Word Level in Data-to-Text Generation. *CoRR*, abs/2102.02810, 2021.

[188] Ehud Reiter. Has a consensus NL Generation Architecture Appeared, and is it Psycholinguistically Plausible? In *7th International Workshop on Natural Language Generation (INLG)*, 1994.

[189] Ehud Reiter. An architecture for Data-to-Text systems. In *11th European Workshop on Natural Language Generation*, pages 97–104, 2007.

[190] Ehud Reiter. Natural language generation. In *The Handbook of Computational Linguistics and Natural Language Processing*, chapter 20, pages 574–598. 2010.

[191] Ehud Reiter and Anja Belz. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35(4):529–558, 2009.

[192] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[193] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, 2000.

[194] Ehud Reiter, Roma Robertson, and Liesl Osman. Types of Knowledge Required to Personalise Smoking Cessation Letters. In *Artificial Intelligence in Medicine. Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making (AIMDM)*, volume 1620 of *Lecture Notes in Computer Science*, pages 389–399, 1999.

[195] Ehud Reiter, Roma Robertson, and Liesl Osman. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2):41–58, 2003.

[196] Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2):137–169, 2005.

[197] Verena Rieser and Oliver Lemon. Natural Language Generation as Planning Under Uncertainty for Spoken Dialogue Systems. In *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 683–691, 2009.

[198] Verena Rieser and Oliver Lemon. *Reinforcement Learning for Adaptive Dialogue Systems - A Data-driven Methodology for Dialogue Management and Natural Language Generation*. Theory and Applications of Natural Language Processing. Springer, 2011.

[199] Lala Septem Riza, Brahma Putra, Yaya Wihardi, and Beta Paramita. Data to text for generating information of weather and air quality in the R programming language. *Journal of Engineering Science and Technology*, 14(1):498–508, 2019.

[200] Majid RobatJazi, Marek Reformat, and Witold Pedrycz. Ontology-based framework for reasoning with fuzzy temporal data. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2030–2035, 2012.

[201] María Ros, Manuel Pegalajar Cuéllar, Miguel Delgado, M. Amparo Vila, Derek T. Anderson, James M. Keller, and Mihail Popescu. Linguistic summarization of long-term trends for understanding change in human behavior. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 2080–2087, 2011.

[202] Enrique H. Ruspini. A new approach to clustering. *Information and control*, 15(1):22–32, 1969.

USC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

[203] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A Survey of Evaluation Metrics Used for NLG Systems. *CoRR*, abs/2008.12009, 2020.

[204] Daniel Sánchez-Valdes, Alberto Álvarez-Álvarez, and Gracián Triviño. Dynamic linguistic descriptions of time series applied to self-track the physical activity. *Fuzzy Sets and Systems*, 285:162–181, 2016.

[205] Xiaoyu Shen, Ernie Chang, Hui Su, Cheng Niu, and Dietrich Klakow. Neural Data-to-Text Generation via Jointly Learning the Segmentation and Correspondence. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7155–7165, 2020.

[206] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. CRC Press, 2003.

[207] Anastasia Shimorina and Claire Gardent. Handling Rare Items in Data-to-Text Generation. In *11th International Conference on Natural Language Generation (INLG)*, pages 360–370, 2018.

[208] Grégory Smits, Pierre Nerzic, Olivier Pivert, and Marie-Jeanne Lesot. Efficient Generation of Reliable Estimated Linguistic Summaries. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, 2018.

[209] Lukasz Sosnowski and Tomasz Penza. Generating Fuzzy Linguistic Summaries for Menstrual Cycles. In *Federated Conference on Computer Science and Information Systems (FedCSIS)*, volume 21, pages 119–128, 2020.

[210] Somayajulu Sripada, Ehud Reiter, and Ian Davy. SumTime-Mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10, 2003.

[211] Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. Evaluation of an NLG System using Post-Edit Data: Lessons Learnt. In *10th European Workshop on Natural Language Generation (ENLG)*, 2005.

[212] Somayajulu Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. Generating English summaries of time series data using the Gricean maxims. In *9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 187–196, 2003.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[213] Amanda Stent and Martín Molina. Evaluating Automatic Extraction of Rules for Sentence Plan Construction. In *10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 290–297, 2009.

[214] Robert Stewart and Sumithra Velupillai. Applied natural language processing in mental health big data. *Neuropsychopharmacology*, 46:252–253, 2021.

[215] Reza Tavakkoli-Moghaddam, Mir-Bahador Aryanezhad, Nima Safaei, and Amir Azaron. Solving a dynamic cell formation problem using metaheuristics. *Applied Mathematics and Computation*, 170(2):761–780, 2005.

[216] Craig Thomson and Ehud Reiter. A Gold Standard Methodology for Evaluating Accuracy in Data-To-Text Systems. In *13th International Conference on Natural Language Generation (INLG)*, pages 158–168, 2020.

[217] Gracián Triviño and Michio Sugeno. Towards linguistic descriptions of phenomena. *International Journal of Approximate Reasoning*, 54(1):22–34, 2013.

[218] Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.

[219] Ross Turner, Somayajulu Sripada, Ehud Reiter, and Ian Davy. Using Spatial Reference Frames to Generate Grounded Textual Summaries of Georeferenced Data. In *5th International Natural Language Generation Conference (INLG)*, 2008.

[220] Ross Turner, Somayajulu Sripada, Ehud Reiter, and Ian P. Davy. Selecting the Content of Textual Descriptions of Geographically Located Events in Spatio-Temporal Weather Data. In *Applications and Innovations in Intelligent Systems XV - 27th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 75–88, 2007.

[221] Albert van der Heide and Gracián Triviño. Automatically Generated Linguistic Summaries of Energy Consumption Data. In *9th International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 553–559, 2009.

[222] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In

UDC UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

12th International Conference on Natural Language Generation (INLG), pages 355–368, 2019.

[223] Chris van der Lee, Bart Verduijn, Emiel Krahmer, and Sander Wubben. Evaluating the text quality, human likeness and tailoring component of PASS: A Dutch data-to-text system for soccer. In *27th International Conference on Computational Linguistics (COLING)*, pages 962–972, 2018.

[224] Peter J. M. Van Laarhoven and Emile H. L. Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.

[225] Sebastian Varges, Heike Bieler, Manfred Stede, Lukas C. Faulstich, Kristin Irsig, and Malik Atalla. SemScribe: Natural Language Generation for Medical Reports. In *8th International Conference on Language Resources and Evaluation (LREC)*, pages 2674–2681, 2012.

[226] M. Amparo Vila, Juan-Carlos Cubero, Juan-Miguel Medina, and Olga Pons. Using OWA operator in flexible query processing. In *The ordered weighted averaging operators*, pages 258–274. Springer, 1997.

[227] M. Vítek. Fuzzy Information and Fuzzy Time. *IFAC Symposium on Fuzzy Information, Knowledge Representation and Decision Analysis*, 16(13):159–162, 1983.

[228] W3C. World Wide Web Consortium (W3C) website. `www.w3.org`, 2021. [Online; Accessed September 2021].

[229] Leo Wanner, Bernd Bohnet, Nadjet Bouayad-Agha, François Lareau, and Daniel Nicklaß. MARQUIS: Generation of User-Tailored Multilingual Air Quality Bulletins. *Applied Artificial Intelligence*, 24(10):914–952, 2010.

[230] Joseph Weizenbaum. ELIZA— A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM*, 9(1):36–45, 1966.

[231] Michael White and Ted Caldwell. EXEMPLARS: A practical, extensible framework for dynamic text generation. In *9th International Workshop on Natural Language Generation (INLG)*, 1998.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[232] Michael White and David M. Howcroft. Inducing Clause-Combining Rules: A Case Study with the SPaRKy Restaurant Corpus. In *15th European Workshop on Natural Language Generation (ENLG)*, pages 28–37, 2015.

[233] Anna Wilbik and Remco M. Dijkman. Linguistic summaries of process data. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, 2015.

[234] Anna Wilbik and Uzay Kaymak. Gradual Linguistic Summaries. In *15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, volume 443 of *Communications in Computer and Information Science*, pages 405–413, 2014.

[235] Anna Wilbik, Uzay Kaymak, and Remco M. Dijkman. A method for improving the generation of linguistic summaries. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2017.

[236] Anna Wilbik, Uzay Kaymak, James M. Keller, and Mihail Popescu. Evaluation of the Truth Value of Linguistic Summaries - Case with Non-monotonic Quantifiers. In *7th International Conference Intelligent Systems (IS)*, volume 322 of *Advances in Intelligent Systems and Computing*, pages 69–79, 2014.

[237] Anna Wilbik and James M. Keller. A distance metric for a space of linguistic summaries. *Fuzzy Sets and Systems*, 208:79–94, 2012.

[238] Anna Wilbik and James M. Keller. Anomaly detection from linguistic summaries. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, 2013.

[239] Anna Wilbik, James M. Keller, and Gregory L. Alexander. Linguistic summarization of sensor data for eldercare. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 2595–2599, 2011.

[240] Anna Wilbik, James M. Keller, and Gregory L. Alexander. Similarity evaluation of sets of linguistic summaries. *International Journal of Intelligent Systems*, 27(10):926–938, 2012.

[241] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

[242] Windguru. Windguru website. `www.windguru.cz`, 2021. [Oneline; Accessed September 2021].

[243] Ronald R. Yager. A new approach to the summarization of data. *Information Sciences*, 28(1):69–86, 1982.

[244] Ronald R. Yager. General Multiple-Objective Decision Functions and Linguistically Quantified Statements. *International Journal of Man-Machine Studies*, 21(5):389–400, 1984.

[245] Ronald R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):183–190, 1988.

[246] Ronald R. Yager. Quantifier guided aggregation using OWA operators. *International Journal of Intelligent Systems*, 11(1):49–73, 1996.

[247] Ronald R. Yager, Kenneth M. Ford, and Alberto J. Cañas. An Approach to the Linguistic Summarization of Data. In *3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, volume 521 of *Lecture Notes in Computer Science*, pages 456–468, 1990.

[248] Georgios N. Yannakakis and Héctor Perez Martínez. Ratings are Overrated! *Frontiers ICT*, 2:13, 2015.

[249] I.-Cheng Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808, 1998.

[250] I.-Cheng Yeh, King-Jang Yang, and Tao-Ming Ting. Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, 36(3):5866–5871, 2009.

[251] Jin Yu, Jim Hunter, Ehud Reiter, and Somayajulu Sripada. An approach to generating summaries of time series data in the gas turbine domain. In *2001 International Conferences on Info-Tech and Info-Net (ICII)*, volume 3, pages 44–51, 2001.

[252] Jin Yu, Ehud Reiter, Jim Hunter, and Somayajulu Sripada. SumTime-Turbine: A Knowledge-Based System to Communicate Gas Turbine Time-Series Data. In *16th*

International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE), volume 2718 of *Lecture Notes in Computer Science*, pages 379–384, 2003.

[253] Lofti A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3–28, 1978.

[254] Lotfi A. Zadeh. Fuzzy Sets. *Information and Control*, 8(3):338–353, 1965.

[255] Lotfi A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning - I. *Information Sciences*, 8(3):199–249, 1975.

[256] Lotfi A. Zadeh. PRUF - A Language for the Representation of Meaning in Natural Languages. In *5th International Joint Conference on Artificial Intelligence*, page 918, 1977.

[257] Lotfi A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. In *Computational linguistics*, pages 149–184. Elsevier, 1983.

[258] Lotfi A Zadeh. Fuzzy logic = computing with words. *IEEE Transactions on fuzzy systems*, 4(2):103–111, 1996.

[259] Lotfi A. Zadeh. From Computing with Numbers to Computing with Words - From Manipulation of Measurements to Manipulation of Perceptions. In *Intelligent Systems and Soft Computing: Prospects, Tools and Applications*, volume 1804 of *Lecture Notes in Computer Science*, pages 3–40, 2000.

[260] Lotfi A. Zadeh. A new direction in AI - Toward a Computational Theory of Perceptions. In *Computational Intelligence, Theory and Applications, International Conference, 7th Fuzzy Days*, volume 2206 of *Lecture Notes in Computer Science*, page 628, 2001.

[261] Lotfi A Zadeh. A prototype-centered approach to adding deduction capability to search engines-the concept of protoform. In *Intelligent Systems. 1st International IEEE Symposium*, volume 1, pages 2–3, 2002.

[262] Lotfi A. Zadeh. Generalized theory of uncertainty (GTU) - principal concepts and ideas. *Computational Statistics & Data Analysis*, 51(1):15–46, 2006.

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

[263] Lotfi A Zadeh and Janusz Kacprzyk. *Fuzzy logic for the management of uncertainty*. John Wiley & Sons, Inc., 1992.

[264] Lotfi A. Zadeh and Janusz Kacprzyk. Computing with Words in Information/Intelligent systems 1: Foundations. *Physica-Verlag, Heidelberg*, 1999.

[265] Sławomir Zadrożny and Janusz Kacprzyk. On the use of linguistic summaries for text categorization. *International Conference on the Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU)*, pages 1373–1380, 2004.

[266] Jiaming Zeng, Imon Banerjee, A. Solomon Henry, Douglas J. Wood, Ross D. Shachter, Michael F. Gensheimer, and Daniel L. Rubin. Natural Language Processing to Identify Cancer Treatments With Electronic Medical Records. *JCO Clinical Cancer Informatics*, (5):379–393, 2021.

[267] Li Zhou and George Hripcsak. Temporal reasoning with medical data - A review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40(2):183–202, 2007.

[268] Alberto Álvarez Álvarez and Gracián Triviño. Linguistic description of the human gait quality. *Engineering Applications of Artificial Intelligence*, 26(1):13–23, 2013.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

List of Figures

Fig. 1.1 Definition of fuzzy quantifiers as fuzzy sets. 5
Fig. 1.2 Linguistic variable composed of five labels defined as fuzzy sets. 5

Fig. 2.1 NLG pipeline defined in [188, 192, 193]. 18

Fig. 3.1 Example of maps with meteorological information showing the difficulty of interpreting its contents. 41
Fig. 3.2 Described locations for the weather observation service provided by Meteo-Galicia (one observation point per municipality). 42
Fig. 3.3 Described locations for the weather prediction service provided by Meteo-Galicia. Each point is the intersection of the grid lines (being each cell 4km. x 4km. size) 43
Fig. 3.4 Definition of the linguistic terms of temperature. 47
Fig. 3.5 Observation and prediction quantifiers definition for the whole territorial coverage (percentage of locations in the map). 48
Fig. 3.6 Reference coordinates from Google Earth [92]. 49
Fig. 3.7 Geographical descriptors in the observation scenario. 50
Fig. 3.8 Geographical descriptors in the prediction scenario. 50

Fig. 3.9 Description of the experimentation stages with their inputs and outputs with examples. . . . 53
Fig. 3.10 Structure of the unary chromosome in the Genetic Algorithm. . . . 56
Fig. 3.11 Structure of the binary chromosome in the Genetic Algorithm. . . . 57
Fig. 3.12 *fulfil_score* representation . . . 59
Fig. 3.13 *cov_score* representation . . . 61
Fig. 3.14 Distribution of the expert scores separated by algorithm and type of quantified statement. . . . 78

Fig. 4.1 Quantifiers' definition of the percentage of locations. . . . 91
Fig. 4.2 Description of the experimentation stages with their inputs and outputs. . . . 92
Fig. 4.3 Linguistic variable definition from numerical variable with 5 equidistant linguistic terms. . . . 99
Fig. 4.4 Representation of the equidistant definition of the quantifiers. . . . 100

Fig. 5.1 Graphical representation of Instant (1), Imprecise interval (2) and Fuzzy interval (3). . . . 110
Fig. 5.2 Top-down hierarchy module from "month" to "hour" and a bottom-up hierarchy module of period evaluation representation of the temporal model. . . . 112
Fig. 5.3 Pipeline of the proposed data-to-text model. Rectangles show the stages, while ellipses show the outputs. . . . 112
Fig. 5.4 Map of the meteorological stations network of the Galician Meteorological Agency. . . . 119
Fig. 5.5 31-hour AQI evolution time series where the station registered "good" values. 119
Fig. 5.6 General linguistic description label calculation. . . . 121
Fig. 5.7 Definition of fuzzy intervals (R30) to represent the beginning and ending of the trend description. . . . 122
Fig. 5.8 Fuzzy definition of a full day. . . . 123
Fig. 5.9 Fuzzy definition of the three moments of the day (morning, afternoon, night) (2) and of the {first, central, last} hours of a moment of the day. . . . 123
Fig. 5.10 Example from the AQI questionnaire designed for the expert validation. . . . 125
Fig. 5.11 Example from the AQI questionnaire designed for the expert validation where a bad situation is described. . . . 127

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Fig. 5.12 Definition of fuzzy intervals (R30) to represent the beginning and ending of the trend description. 130
Fig. 5.13 Calculation of the representative value in the general summary. 132
Fig. 5.14 Example of the generated linguistic description from data of the HDL cholesterol. 140
Fig. 5.15 Example extracted from the questionnaire of the generated linguistic description from data of the in-hospital diastolic blood pressure. 141
Fig. 5.16 Example extracted from the questionnaire of the generated linguistic description from data of the blood glucose. 142

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

List of Tables

Tab. 2.1 Types of protoforms structures in LDD 25

Tab. 3.2 List of wind codes with their respective icons defined in the MeteoGalicia style book. 46

Tab. 3.3 Values of the fuzzy linguistic variable “temperature”. 47

Tab. 3.1 Sky state codes with their respective icons defined in the MeteoGalicia’s style book. 51

Tab. 3.4 Template to describe each meteorological variable in the generated descriptions. 52

Tab. 3.5 Coverage values for each defined quantifier. 60

Tab. 3.6 Example of input data where the numerical data means the percentage of accuracy of our approaches when processing the available data sets. 64

Tab. 3.7 Results of skewness, kurtosis and Jarque-Bera test about the compared data. 66

Tab. 3.8 Results of the Breusch-Pagan test for heterocedasticity. 67

Tab. 3.9 Summary of the results of the unary random baseline. 69

Tab. 3.10 Summary of the results of the binary random baseline. 69

Tab. 3.11 Metrics results from the unary observation analysis. 70

Tab. 3.12 Example of the truth value obtained by the three quantification methods for a set of unary fuzzy quantified sentences. 71
Tab. 3.13 Metrics results from the binary observation analysis. 72
Tab. 3.14 Summary of the results of the statistical tests applied to the observation case. 73
Tab. 3.15 Metrics results from the unary prediction analysis. 73
Tab. 3.16 Metrics results from the binary prediction analysis. 75
Tab. 3.17 Summary of the results of the t-test applied to the prediction data. 75
Tab. 3.18 Summary of the results of the statistical tests between the baseline and the meta-heuristic approaches for the unary case. 76
Tab. 3.19 Summary of the results of the statistical tests between the baseline and the meta-heuristic approaches for the binary case. 76
Tab. 3.20 Summary of the significance difference detection between the GA and SA meta-heuristic where ✓ means "significant difference between GA and SA" and ✗ means "no significant differences detected between GA and SA". Differences are always in favor of GA. 81

Tab. 4.1 Recompilation of fulfillment of the unary properties from [51] by the involved fuzzy quantification methods. 89
Tab. 4.2 Recompilation of the fulfillmenting of the binary properties from [51] by the involved fuzzy quantification methods (blank spaces mean no information in the paper). 90
Tab. 4.3 Pearson correlation coefficient for quantification methods evaluating unary descriptions from 30 July 2019. 94
Tab. 4.4 Pearson correlation coefficient for quantification methods evaluating unary descriptions from the average data set describing temperature. 94
Tab. 4.5 Pearson correlation coefficient for quantification methods evaluating unary descriptions from the average data set including all meteorological variables. 95
Tab. 4.6 Pearson correlation coefficient for quantification methods evaluating binary descriptions from the 30th July 2019 data set describing temperature. 96
Tab. 4.7 Pearson correlation coefficient for quantification methods evaluating binary descriptions from the average data set describing temperature. 96
Tab. 4.8 Pearson correlation coefficient for quantification methods evaluating binary descriptions from the average data set including all meteorological variables. 96
Tab. 4.9 Some relevant features of the data sets used in the experiments 98

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Tab. 4.10 Number of experiments performed for each scenario. 101
Tab. 4.11 Percentages of cases where unary quantified models show similar behavior for all the variables (crisp and fuzzy). 103
Tab. 4.12 Percentages of cases where unary quantified models show similar behavior for the subset of fuzzy variables. 103
Tab. 4.13 Percentages of cases where binary quantified models show similar behavior for all the variables (crisp and fuzzy). 104
Tab. 4.14 Percentages of cases where binary quantified models show similar behavior for the subset of fuzzy variables. 104

Tab. 5.1 Description of the rules which compose the grammar of the designed linguistic summarization model. 115
Tab. 5.2 Air Quality Index (AQI) labels with their perception and numerical index. . 120
Tab. 5.3 Range of hours for the moments of the day {morning, afternoon, night} in summer or winter from the style guide of MeteoGalicia. 121
Tab. 5.4 Questions of the air quality index expert validation test. 126
Tab. 5.5 Expert validation results for AQI descriptions. 127
Tab. 5.6 Monitoring variables during the cardiac rehabilitation grouped by category. . 130
Tab. 5.7 Reference values of HDL cholesterol. 131
Tab. 5.8 Reference values of LDL cholesterol. 131
Tab. 5.9 Reference values of total cholesterol. 132
Tab. 5.10 Reference values of triglycerides. 132
Tab. 5.11 Reference values of alanine transaminase (ALT). 133
Tab. 5.12 Reference values of aspartate aminotransferase (AST). 133
Tab. 5.13 Reference values of alkaline phosphatase (ALP). 133
Tab. 5.14 Reference values of gamma-glutamyl transpeptidase (GGT). 134
Tab. 5.15 Reference values of diastolic blood pressure. 134
Tab. 5.16 Reference values of systolic blood pressure. 135
Tab. 5.17 Reference values of cooximetry. 135
Tab. 5.18 Reference values of blood sugar. 136
Tab. 5.19 Reference values of glycated hemoglobin. 136
Tab. 5.20 Reference values of body mass index (BMI). 137
Tab. 5.21 Reference values of weight. 137
Tab. 5.22 Reference values of body fat percentage. 137

Tab. 5.23 Items of the expert evaluation questionnaire. 141
Tab. 5.24 Expert evaluation results from the questionnaire. 142

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

List of Acronyms

GIS	Geographic Information Systems
AI	Artificial Intelligence
NLG	Natural Language Generation
D2T	Data-To-Text
LDD	Linguistic Descriptions of Data
NLP	Natural Language Processing
NLU	Natural Language Understanding
NICU	Neonatal Intensive Care Unit
CWW	Computing With Words
CTP	Computational Theory of Perceptions
GLMP	Granular Linguistic Model of a Phenomemon
PM	Perception Mappings
CP	Computational Perceptions
TGQ	Theory of Generalized Quantifiers
QFM	Quantifier Fuzzification Mechanisms
FTPs	Fuzzy Temporal Propositions

USC
UNIDADE
SANTIAGO
COMPOSTELA

GLiDTS	Generation of Linguistic Descriptions of Time Series
PLS	Piecewise Linear Segments
LTAG	Lexicalized Tree Adjoining Grammar
CFG	Context-Free Grammar
NN	Neural Network
SA	Simulated Annealing
GA	Genetic Algorithm
Z_Q	Zadeh's scalar quantification method
GD_Q	Delgado's GD method
ZS_Q	Zadeh's fuzzy cardinality based method
SD	Standard Deviation
DF	Degrees of Freedom
AQI	Air Quality Index
CRU	Cardiac Rehabilitation Units
CR	Cardiac Rehabilitation
ALT	Alanine Transaminase
AST	Aspartate Aminotransferase
ALP	Alkaline Phosphatase
GGT	Gamma-glutamyl Transpeptidase
Hba1C	Glycated Hemoglobin
BMI	Body Mass Index

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA





Hoxe en día, grandes cantidades de datos son consumidos e xerados, que teñen gran valor na medida que se pode extraer a información que conteñen para ser empregada nos procesos de toma de decisións. É neste contexto onde os sistemas data-to-text (D2T), que se centran na xeración automática de textos a partir de fontes de datos numéricos ou simbólicos, son de gran utilidade. Por outra banda, no campo da Lóxica Borrosa xurdiron propostas para describir datos empregando termos lingüísticos.
Os obxectivos desta tese son: estender e mellorar a fase de determinación de contido en sistemas D2T para representar o coñecemento impreciso e a busca intelixente, medir e comparar o impacto dos métodos de cuantificación na avaliación
de sentencias cuantificadas borrosas, e deseñar un modelo D2T para describir series temporais.